A close-up, low-angle shot of a pinball machine's playfield. The machine is illuminated with warm, golden light, highlighting the metallic surfaces and the vibrant, colorful artwork. The artwork features a character in a blue and yellow suit, possibly a superhero, with the word "Skill" written in a stylized font. The background is dark, with some distant lights visible, suggesting a dimly lit arcade or casino setting.

# Causality, Probability, and Time

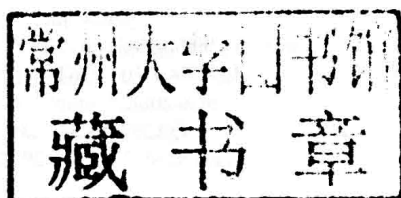
SAMANTHA KLEINBERG

CAMBRIDGE

# CAUSALITY, PROBABILITY, AND TIME

SAMANTHA KLEINBERG

*Stevens Institute of Technology, Hoboken, New Jersey*



CAMBRIDGE  
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS  
Cambridge, New York, Melbourne, Madrid, Cape Town,  
Singapore, São Paulo, Delhi, Mexico City  
Cambridge University Press  
32 Avenue of the Americas, New York, NY 10013-2473, USA  
www.cambridge.org  
Information on this title: www.cambridge.org/9781107026483

© Samantha Kleinberg 2013

This publication is in copyright. Subject to statutory exception  
and to the provisions of relevant collective licensing agreements,  
no reproduction of any part may take place without the written  
permission of Cambridge University Press.

First published 2013

Printed in the United States of America

*A catalog record for this publication is available from the British Library.*

*Library of Congress Cataloging in Publication Data*

Kleinberg, Samantha, 1983– author.  
Causality, probability, and time / Samantha Kleinberg, Stevens Institute of Technology,  
Hoboken, New Jersey.  
pages cm.

Includes bibliographical references and index.

ISBN 978-1-107-02648-3 (hardback)

1. Computational complexity. I. Title.

QA267.7.K54 2012

511.3'52–dc23 2012021047

ISBN 978-1-107-02648-3 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for  
external or third-party Internet Web sites referred to in this publication and does not guarantee  
that any content on such Web sites is, or will remain, accurate or appropriate.

## CAUSALITY, PROBABILITY, AND TIME

Causality is a key part of many fields and facets of life, from finding the relationship between diet and disease to discovering the reason for a particular stock market crash. Despite centuries of work in philosophy and decades of computational research, automated inference and explanation remain an open problem. In particular, the timing and complexity of relationships have been largely ignored even though this information is critically important for prediction, explanation, and intervention. However, given the growing availability of large observational datasets, including those from electronic health records and social networks, it is a practical necessity. This book presents a new approach to inference (finding relationships from a set of data) and explanation (assessing why a particular event occurred), addressing both the timing and complexity of relationships. The practical use of the method developed is illustrated through theoretical and experimental case studies, demonstrating its feasibility and success.

Samantha Kleinberg is Assistant Professor of Computer Science at Stevens Institute of Technology. She received a PhD in Computer Science and a BA in Computer Science and Physics from New York University.



# Acknowledgments

From the beginning, this work has been profoundly interdisciplinary. I am deeply grateful to the collaborators and colleagues who have enabled me to explore new fields and who have enthusiastically shared their expertise with me. Immersing myself in philosophy, bioinformatics, finance, and neurology among other areas has been challenging, exciting, and necessary to this work. I also thank the audiences and anonymous referees from the conferences, workshops, and seminars where I have presented earlier versions of this work for their feedback.

This book began when I was a graduate student at NYU and was completed during my post-doc at Columbia. The support of my colleagues and their generosity with their time and data have significantly shaped and improved the work, bringing it closer to practice. In particular, collaboration with medical doctors has given me a new appreciation for the importance of automating explanation and for the practical challenges this work faces.

This material is based on work supported by the NSF under Grant #1019343 to the Computing Research Association for the CIFellows project. That fellowship provided salary and research support for the last two years and was instrumental to the completion of this work. This work has also been funded in part with federal funds from the NLM, NIH, DHHS, under Contract No. HHSN276201000024C.

Prior versions of this work have appeared in conference proceedings and journals. In particular, material in Chapters 4 and 7 are partly based on work that appeared in Kleinberg and Mishra (2009); Chapter 6 is partly based on work that appeared in Kleinberg and Mishra (2010); and portions of Chapter 2 previously appeared in Kleinberg and Hripcsak (2011).

Finally I thank my editor, Lauren Cowles, at Cambridge University Press for her support and enthusiasm for this book.



# Contents

*Acknowledgments* *page vii*

1	Introduction	1
1.1	Structure of the Book	8
2	A Brief History of Causality	11
2.1	Philosophical Foundations of Causality	11
2.2	Modern Philosophical Approaches to Causality	13
2.3	Probabilistic Causality	18
2.4	Causal Inference Algorithms	32
3	Probability, Logic, and Probabilistic Temporal Logic	43
3.1	Probability	43
3.2	Logic	49
3.3	Probabilistic Temporal Logic	58
4	Defining Causality	65
4.1	Preliminaries	65
4.2	Types of Causes and Their Representation	76
4.3	Difficult Cases	96
5	Inferring Causality	111
5.1	Testing Prima Facie Causality	111
5.2	Testing for Causal Significance	120
5.3	Inference with Unknown Times	127
5.4	Correctness and Complexity	133
6	Token Causality	142
6.1	Introduction to Token Causality	142
6.2	From Types to Tokens	150



6.3	Whodunit? (Examples of Token Causality)	161
6.4	Difficult Cases	170
7	Case Studies	183
7.1	Simulated Neural Spike Trains	183
7.2	Finance	195
8	Conclusion	206
8.1	Broader Connections	211
8.2	Looking Forward	213
A	A Little Bit of Statistics	217
A.1	Preliminaries	217
A.2	Multiple Hypothesis Testing	218
B	Proofs	224
B.1	Probability Raising	224
B.2	Equivalence to Probabilistic Theory of Causality	224
B.3	Leads-to with Both Lower and Upper Time Bounds	230
	<i>Glossary</i>	237
	<i>Bibliography</i>	241
	<i>Index</i>	251

# 1

## Introduction

Whether we want to know the cause of a stock's price movements (in order to trade on this information), the key phrases that can alter public opinion of a candidate (in order to optimize a politician's speeches), or which genes work together to regulate a disease causing process (in order to intervene and disrupt it), many goals center on finding and using causes. Causes tell us not only that two phenomena are related, but how they are related. They allow us to make robust predictions about the future, explain the relationship between and occurrence of events, and develop effective policies for intervention.

While predictions are often made successfully on the basis of associations alone, these relationships can be unstable. If we do not know why the resulting models work, we cannot foresee when they will stop working. Lung cancer rates in an area may be correlated with match sales if many smokers use matches to light their cigarettes, but match sales may also be influenced by blackouts and seasonal trends (with many purchases around holidays or in winter). A spike in match sales due to a blackout will not result in the predicted spike in lung cancer rates, but without knowledge of the underlying causes we would not be able to anticipate that failure. Models based on associations can also lead to redundancies, since multiple effects of the true cause may be included as they are correlated with its occurrence. In applications to the biomedical domain, this can result in unnecessary diagnostic tests that may be invasive and expensive.

In addition to making forecasts, we want to gain new knowledge of how things work. Causes enable us to explain both the occurrence of events and the connection between types of events. We do not want to know only that a particular drug is associated with renal failure, but rather we want to distinguish between whether this association is due to an adverse drug reaction or the disease being treated causing both renal failure and prescription of the drug. Associations do not have this type of explanatory power, nor can they help us with a second type of explanation, that of why a particular

event occurred. When attempting to explain why a patient developed a secondary brain injury after a stroke, the goal is to determine which factors are responsible so that these can be treated to potentially prevent further brain damage. Knowing only that a particular event is correlated with secondary injury is insufficient to determine which factors made a difference to its occurrence in a particular case.

Finally, knowledge of the underlying causes of a phenomenon is what allows us to intervene successfully to prevent or produce particular outcomes. Causal relationships (actual or hypothesized) prompt us to make decisions such as taking vitamin supplements to reduce our risk of disease or enacting policies decreasing sodium levels in food to prevent hypertension. If we did not at least believe that there is a causal connection between these factors, we would have no basis for these interventions. Intervening on a side effect of the underlying cause would be like banning the sale of matches to reduce lung cancer rates. This is clearly ineffective, since smokers can also use lighters, but banning smoking or reducing smoking rates does have the ability to lower lung cancer rates. In general, to bring about desired outcomes we must know that the factor being acted upon is capable of preventing or producing the effect of interest.

However, causality alone is not enough. To use causes to effectively predict, explain, or alter behavior, we must also know the time over which a relationship takes place, the probability with which it will occur, and how other factors interact to alter its efficacy.

When finding factors that affect stock prices, we need to know when the effect starts and how long it persists to be able to trade on this information. Individual phrases may positively influence voter perception of a politician, but candidates must combine these into coherent speeches, and two phrases that are positive individually may have a negative impact in combination. With multiple targets for drug development, the likelihood of each being effective must be weighed against its potential risks to determine which candidates to pursue.

Few relationships are deterministic, so even if we know the details of a cause that can produce the desired effect and how long it takes to do so, we cannot be certain that this outcome will occur in all instances. In many cases, this is due to the limits of our knowledge (as it is rare that all factors relevant to the success of the cause can be enumerated) while in others the relationship itself may be probabilistic. Knowing both the timing of relationships and their probabilities is important for making decisions and assessing risk, as there are often multiple effects of a cause and multiple

causes of a particular effect. Thus, we can rarely influence a cause in isolation, and must also choose between potential candidates. For many medical conditions, doctors have a choice of treatments where some may be extremely effective, yet come with the potential for severe side effects, while other less effective drugs may be desirable because of their limited side effects. When choosing a target for interventions, one must evaluate the strength of the relationship (likelihood of the cause producing the effect, or the magnitude of influence) against potentially undesirable side effects. This has been partly addressed by artificial intelligence work on planning, which finds both direct and indirect effects (ramifications) of actions to determine whether a strategy will achieve a goal. These methods assume that we already have a model of how the system works, but in many cases the first step of research is finding this model or creating it with the input of domain experts. By starting with a set of causal facts (essentially, ways of changing the truth value of formulas), these methods free themselves from answering the most difficult question: what exactly is causality?

This question has plagued researchers in many areas, but it has been a fundamental practical problem in medicine where doctors must always act with incomplete information. Causality is at the center of every facet of medicine, including diagnosis of patients (Rizzi, 1994), identification of adverse drug events (Agbabiaka et al., 2008), comparative effectiveness research (Johnson et al., 2009), and epidemiological studies linking environmental factors and disease (Parascandola and Weed, 2001). Yet as central as causality is to biomedical research, work on understanding what it is and how to find it has primarily taken a pragmatic approach, disconnected from the philosophical literature in this area. As a result, randomized controlled trials (RCTs) have come to be treated as the gold standard for causal inference, even though these can answer only a subset of the many causal questions researchers and clinicians aim to answer and sidestep the question of what causality actually is. The basic idea of an RCT is that a subset of a population has been randomly assigned to a particular treatment while the control group does not receive the treatment. Both are measured the same way for the same time, and when there is a difference in outcomes between the groups it is said that the therapy is responsible for it (as it is meant to be the only difference between them). These methods have many well-known limitations, in particular that the ideal of randomization to eliminate confounding may not always occur in practice (Schulz et al., 1995), and that the internal validity of these studies (that they can answer the questions being asked) often comes at the expense of external validity (generalizability to

other populations) (Dekkers et al., 2010; Rothwell, 2006). Similarly, due to the difficulty and expense of enrolling patients, these studies follow fairly small populations over fairly short time periods.

Instead, new large-scale observational datasets from electronic health records (EHRs) may address some of these limitations (by studying the same population being treated, following patients over a long timescale, and using a large population). Columbia University Medical Center, for example, has a database of 3 million patients over twenty years. In other systems with less in and out-migration, these records can capture a patient's health over nearly their entire lifespan. Further, while many RCTs involve homogeneous sets of patients with few comorbidities, EHRs contain a more realistic set of patients (though they exclude those who have not sought or do not have access to medical care). Despite the potential benefits of using EHRs for research, they have been underused, as these observational data are outside the traditional paradigm of RCTs (here we have no control over the data gathered and patients may have many gaps in their records) and have been difficult to analyze using prior computational methods for causal inference (as few of their assumptions hold in these types of real-world datasets).

To address the challenge of causal inference from observational data, though, we first need to understand what causality is in a domain-independent way. Attempts have been made to create guidelines for evaluating causality in specific scenarios, such as Hill's viewpoints on causality (Hill, 1965), but these are simply heuristics. Over time though they have come to be treated as checklists, leading to a conflation of what causality might be with the evidence needed to establish it and tools we can use to recognize it. While I aim to develop practical inference methods, we must be clear about what is being inferred and this requires us to engage with the philosophical literature.

There is no single accepted theory of what it means for something to be a cause, but understanding this distinction between the underlying fact of causality and how inference algorithms identify causes (and which causes they identify) is critical for successful inference and interpretation of results. As will become clear in the later chapters, causality is far from a solved problem, but philosophical theories have succeeded in capturing many more aspects of it than are addressed in the computational literature. There is a small set of cases on which all theories agree, with only partial overlaps in others. Since there are generally no corresponding algorithms that can be applied to test datasets, the primary method for evaluating and comparing philosophical theories of causality has been by posing counterexamples to each, following a battery of tests that have evolved over the years. As no one

theory addresses all potential challenges, this provides some idea of which theories apply to which scenarios, but has also indicated that the search for a unified theory may be unlikely to succeed.

In this book, I will not attempt to provide a unifying theory of causality, but rather aim to make clear where there are areas of disagreement and controversy and where certain assumptions are generally accepted. The book begins with a review of philosophical approaches to causality, because these works give us a vocabulary for talking about it and they provide the foundation for the computational literature. In particular, philosophy is one of the few fields that has extensively studied both type-level causality (general relationships such as that between an environmental factor and a disease) and token causality (specific relationships instantiated at particular times and places, such as the cause of a particular patient's hypertension), as well as the link between these levels. While philosophical approaches have attempted to find one theory that accounts for all instances of causality (arguing against any approach that does not act as expected in at least one case), this has so far not succeeded but has yielded a rich set of competing theories. Given the lack of a unified solution after centuries of effort, some philosophers have recently argued for causal pluralism (with a plurality of things one might be plural about, including methodologies, causality itself, and so on). On the other hand, computational work has honed in on a few inference methods, primarily based on graphical models (where edges between nodes indicate causal dependence), but these may not be appropriate for all cases. Instead, we may once again take inspiration from the philosophical literature to guide development of a set of complementary methods for causal inference.

One of the most critical pieces of information about causality, though – the time it takes for the cause to produce its effect – has been largely ignored by both philosophical theories and computational methods. If we do not know when the effect will occur, we have little hope of being able to act successfully using the causal relationship. We need to know the timing of biological processes to disrupt them to prevent disease. We need to know how long it takes for conditions to trigger political instability if we want to react quickly to it. We need to know a patient's sequence of symptoms and medical history to determine her diagnosis. Further, personal and policy decisions may vary considerably with the length of time between cause and effect (and how this relates to the relationship's probability). The warning that "smoking causes lung cancer" tells us nothing about how long it will take for lung cancer to develop nor how likely this is to occur. We often see people who smoke and do not develop lung cancer, so we immediately know

that either this must occur on such a long timescale that other causes of death occur first, or that the relationship must be probabilistic. Without these details though, an individual cannot adequately assess their risk to make a decision about whether or not to smoke. While a deterministic relationship that takes 80 years may not affect a person's behavior, a relationship with a significantly lower probability at a timescale of only 10–15 years might be significantly more alarming.

To successfully make and use causal inferences we need to understand not only what causality is, but how to represent and infer it in all of its complexity.

I argue that it is futile to insist on a single theory that can handle all possible counterexamples and applications, and instead focus on developing an approach that is best equipped for inferring complex causal relationships (and their timing) from temporal data. While this method builds on philosophical work, the goal is not to develop a theory of causality itself, but rather a method for causal inference and explanation that aims to be philosophically sound, computationally feasible, and statistically rigorous. Since the goal is to use these methods in many areas – such as biology, politics, and finance – the definitions must be domain independent and should be compatible with the types of data that are realistically encountered in practice. This method needs to capture the probabilistic nature of the relationships being inferred, and be able to reason about potentially complex relationships as well as the time between cause and effect. I will discuss why previous methods for causal inference (those that result in the creation of networks or graphs, and those allowing simple lags between cause and effect but not windows of time) do not achieve these goals. Instead, I present an alternative approach based on the idea of causal relationships as logical statements, building on philosophical theories of probabilistic causality and extending probabilistic temporal logics to meet the representation needs of the complex domains discussed.

In this approach, cause, effect, and the conditions for causality are described in terms of logical formulas. This allows the method to capture relationships such as: “smoking and asbestos exposure until a particular genetic mutation occurs causes lung cancer with probability 0.6 in between 1 and 3 years.” While I focus on the case of temporal data, the working definitions developed 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 efficiently test any relationship that can be described in the logic. The approach is based on probabilistic theories of causality, but probability raising alone



is insufficient for identifying causal relationships since many non-causes may precede and seem to raise the probability of other events. Instead, to determine which relationships are significant, I introduce a new measure for the significance of a cause for its effect that assesses the average impact a cause makes to an effect's probability. Using the properties of this measure we are also able to determine the timing of relationships with minimal prior knowledge. Similarly, the distribution of this measure allows standard statistical methods to be applied to find which causal significance values should be considered statistically significant. The inference methods here build on philosophical theories of probabilistic causality, but introduce new computationally feasible methods for representing and inferring relationships.

In addition to inferring general relationships such as that smothering someone causes their death, we also aim to find causes for specific events, such as that Othello smothering Desdemona caused her death. These singular, token-level, relationships need not correspond exactly to type-level relationships. For example, seatbelts may prevent death in the majority of accidents, but can cause it in others by preventing escape from vehicles submerged under water. However, methods that make use of type-level relationships without being constrained by them can enable us to automate this type of reasoning. Finding the causes of particular events is a significant practical problem in biomedicine, where clinicians aim to diagnose patients based on their symptoms and understand their individual disease etiology. Algorithms that can do this without human input can have a particularly large impact in critical care medicine, where doctors face an enormous volume of streaming data that is too complex for humans to analyze, yet knowing not only what is happening but why is essential to treatment. Since treatments can come with potential risks, doctors must be sure they are treating the underlying cause of a patient's illness and not simply symptoms that indicate their level of health. Timing is critical for automating this type of explanation, since it allows objective determination of whether an observed sequence can be considered an instance of the known general relationship and provides information on when a cause is capable of producing its effect. This must also be done with incomplete data (as we may not observe all variables and may have gaps in their recording), and must allow for deviations in timing (as we do not usually have continuous data streams at an arbitrarily fine level of granularity). There are many reasons inferred timings may differ from particular timings even though the particular events are still instances of the general relationship. Inferring, for instance, that a factor causes decreased potassium levels in 60–120 minutes



does not necessarily mean that it is not possible for this to occur in 59 to 121 minutes. The need for this type of reasoning is not limited to biomedicine, but may also apply to finding causes of stock market crashes and software failures. In this book, I aim to close the loop from data to inference to explanation, developing methods for assessing potential token causes for an effect while allowing for incomplete and uncertain information.

### 1.1. Structure of the Book

This book is written primarily for computer scientists and philosophers of science, but it is intended to be accessible to biomedical scientists and researchers in finance among other areas. For that reason, the book is mostly self-contained, and assumes very minimal background in statistics, logic, or philosophy. Chapters 2 and 3 contain all needed background on causality, probability, and logic. Before discussing methods for inferring causes, one needs to understand what is being inferred. Thus, chapter 2 begins with a short introduction to philosophical theories of causality, beginning with historical foundations and then continuing with a critical discussion of probabilistic and counterfactual theories. This discussion covers the problem of defining and recognizing causal relationships, which is necessary before we can discuss how to find these in an automated way. The goal of this section is to make readers from all backgrounds familiar with potential problems in defining causality, providing a framework for evaluating other methods. Finally, I review recent approaches to inference, including graphical models and Granger causality. Chapter 3 is a gentle introduction to probability (covering what is needed for the later examples and algorithms) and temporal logic, concluding with a discussion of the probabilistic temporal logic that the approach builds on.

In the remaining chapters, we turn our attention to a new approach to causal inference. In chapter 4, I begin by defining the types of causes we will aim to identify. Rather than partitioning relationships into causal and non-causal, I focus on calculating the significance of relationships, introducing a new measure for this purpose that is computationally feasible, but based on the philosophical theories discussed in chapter 2. I relate the definitions to probabilistic temporal logic formulas and discuss how they deal with common counterexamples posed to theories of causality. By representing causal relationships as temporal logic formulas (and later extending this logic for use with data), this approach can address the previously ignored problem of representing and inferring complex, temporal, causal relationships. This will allow us to find relationships and their timing (how long it takes for a