Using Imprecise Complexes To Computationally Recognize Causal Relations In Very Large Data Sets

Lawrence J. Mazlack
Applied Artificial Intelligence Laboratory
University of Cincinnati
Cincinnati, Ohio 45221
mazlack@uc.edu

Abstract

Computationally recognizing causal relationships in data is fundamentally important to good decision-making. There are vast amounts of computer stored, multi-faceted data. Understanding how stored data items affect each other is crucial in making good decisions. The most important decisional information is an understanding of causal relationships. A method to discover causality from large, observational data sets would be transformative. Broadly effective computational methods are computationally untested.

An abundance of digital data riches promise a profound impact in both the quality and rate of discovery and innovation in science and engineering, as well as in other societal contexts. Worldwide, researchers are producing, accessing, analyzing, integrating and storing massive amounts of digital data daily, through observation, experimentation and simulation, as well as through the creation of collections of digital representations of tangible artifacts and specimens. Modern experimental and observational instruments generate and collect large sets of data of varying types (numerical, video, audio, textual, multi-modal, multi-level, multi-resolution) at increasing speeds. Often, the data users are not the data producers, and they thus face challenges in harnessing data in unforeseen and unplanned ways. In many science or engineering applications, for example, in mesoscale weather prediction or critical infrastructure protection applications, the ability to gather, organize, analyze, model, and visualize large, multi-scale, heterogeneous data sets in rapid fashion is often crucial.

Keywords: Causality, Imprecision, Complexes, Granularity, Graphs

1 Introduction

The ability to recognize and develop causal relationships is essential for reasoning; it forms the basis for learning to act intelligentially in the world. Knowledge of causal relationships provides a deep understanding of a system; and, potential control over the system coming from the ability to predict action's consequences that have yet to be performed. Starting with the ancient Greeks, philosophers, mathematicians, computer scientists, cognitive scientists, psychologists, and others have formally explored questions of causation.

The recognition of causal relationships from observational data, while sought for generations, is still very unsatisfactory. Most investigators no longer believe in the efficacy of astrology or in reading animal entrails; but an alternative, broadly effective methodology has yet to be developed. Causality has been widely considered by the earliest recorded investigators such as Zeno [28] and Plato [14]. However, little has been done to *computationally recognize causality* in large amounts of observational data.

Causal relationships exist in the commonsense world. When a glass is pushed off a table and breaks on the floor, we might say that being pushed from the table caused the glass to break.

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 698–704
Torremolinos (Málaga), June 22–27, 2008

(Although, being pushed from a table is not a *certain* cause of breakage; sometimes the glass bounces and no break occurs; or, someone catches the glass before it hits the floor.) More weakly, counterfactually, *not* falling to the floor *may* prevent breakage. (Sometimes, a glass breaks when it is tipped over on the table.) So, knowledge of at least some causal effects is imprecise. Perhaps, complete knowledge of all possible factors might lead to a crisp description of whether an effect will occur. However, in our commonsense world, it is unlikely that all possible factors can be known with certainty.

This lack of complete, precise knowledge should not be discouraging. We do things in the world by exploiting our commonsense *perceptions* [27] of cause and effect. When trying to precisely reason about causality, we need complete knowledge of all of the relevant events and circumstances. In commonsense, every day reasoning, we use approaches that do not require complete knowledge. Often, approaches follow what is essentially a *satisficing* [23] paradigm.

2 Complexes

When events happen, there are usually other related events. The entire collection of events can be called a complex. The events can be called the elements of the complex.

A mechanism [24] or a "causal complex" [3] [4] is a collection of events whose occurrence or non-occurrence results in a consequent event happening. A causal complex is the *complete* set of events and conditions necessary for the causal effect (consequent) to occur. Hobbs [3] suggests that using a causal complex does not require precise, complete knowledge of the complex.

Each complex can be considered to be a granule. Larger complexes can be decomposed into smaller complexes. Thus, going from larger-grained to smaller-grained. For example, in Figure 1, the largest-grained event is the sole causal element: *turn on the ignition switch*. The complex of other elements represents the finer-grains. These elements in turn could be broken down into still finer-grains; for example, *available fuel* can be broken down into: *fuel in tank, working fuel pump, intact fuel lines.*

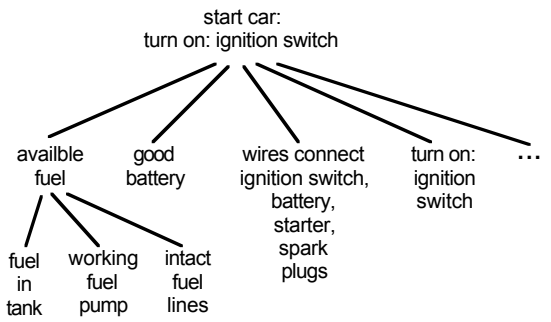

Figure 1. Nested causal complex.

3 Relationship Between Granularity and Imprecision

Causality is often granular. This is true both for commonsense reasoning as well as with more formal mathematical and scientific theories. At a very fine-grained level, the physical world itself (including time) may be granular (at least if modern string theory is correct).

Commonsense causal perceptions are often granular. Some causal events may be larger-grained while some supporting causal structures may be described as relatively fine-grained. In turn, there may be still finer identifiable causal elements. Each successively finer level may be thought of as a nested granular causal structure. Larger-grained causal objects are necessarily more imprecise as some of their constituent components. Some components of a larger-grained causal object may be precisely known, while others maybe be imprecise, and others unknown. The larger the grain, the greater is the likelihood that there might be underlying missing or unknown supporting components.

How to algorithmically evaluate the impreciseness of a larger-grained causal object when some of the underlying components are imprecise is not clear. Perhaps, some form of type-II fuzzy logic [11] manipulation might be helpful. A more detailed discussion of granularity and causality may be found in Mazlack [10].

4 Coincidence Is Not Causation

4.1 Statistics

The use of statistics has enabled investigators to have an idea of what items have a better chance of co-occurring. However, coincidence is not causality. For example, when someone shops in a food store, there is a relatively good chance that both bread and milk will be purchased during the same store visit; but, the inter-item causal relationship is minimal. Alternatively, when strawberries and whipped-cream are both purchased during the same store visit, the inter-item causal relationship has some strength.

The standard method in the experimental sciences of recognizing causality is to perform randomized, controlled experiments. Depending on their design, randomized experiments may remove reasons for uncertainty whether or not a relationship is casual. However, the very large, non-experimental data that need to be examined are not the product of controlled experiments. Some work has been done using statistical testing to reduce the search space in non-experimental data [1] [21].

4.2 Association Rules

There are several different data mining products that are sometimes naïvely considered to be causal [9]. The most common are *association rules*.

Customers who
buy beer and sausage
also tend to buy hamburger
with {confidence = 0.7}
in {support = 0.2}

Figure 2. Example of an association rule

At first glance, association rules, seem to imply a cause-effect relationship; that is:

> A customer's purchase of both sausage and beer causes the customer to also buy hamburger.

But, all that is discovered is the *existence* of a statistical relationship between the items. They have a degree of joint occurrence. The *nature* of the relationship is not specified. It is not known whether the presence of an item or sets of items causes the presence of another item or set of items; or the converse, or some other phenomenon causes them to occur together.

Purely accidental relationships are not nearly of the same level of interest as causal relationships. For example, if it is true that buying both *beer and sausage* somehow causes someone to *buy hamburger*, then a merchant might profitably put *beer* (or *sausage*) on sale and then increase the price of *hamburger* to compensate for the sale price. On the other hand, knowing that *bread* and *milk* are often purchased together may not be useful information as both products are commonly purchased on every store visit.

One of the reasons why association rules are developed is to aid in making retail decisions. However, simple association rules may lead to errors. Errors might occur; either if causality is recognized where there is no causality; or if the direction of the causal relationship is wrong [21] [7].

5 Existing Causal Recognition Methods

5.1 Positive Causation

There are different approaches to causality. The idea of "positive" causation ($\alpha \rightarrow \beta$) is at the core of common sense causal reasoning. Often a positive causal relationship is represented as a network of nodes and branches [6].

5.2 Counterfactuals

Negation or counterfactuals ($\neg\alpha \rightarrow \neg\beta$) have a place; although it may result in errors in reasoning. For example:

> If a person drinks *wine*, they may become inebriated.

cannot be simply negated to

> If a person does not drink *wine*, they will not become inebriated.

Generally, counterfactual reasoning can play a useful role in causal reasoning [Ortiz, 1999]. Counterfactuals can represent useful tools for identifying the role that an event plays in a collection of events. Unfortunately, because of computational needs, their use in non-experimental data is limited [26].

Effects can be *over determined* [16]*;* that is: more than one item can cause an effect. In the previous case,

If a person drinks *wine*, they may become inebriated.

people may, at the same time, also may be

Drinking *beer* and become *inebriated* whether or not they do *not drink wine*.

The issue of overdetermination, and the degree of overdetermination makes causal discovery more complex.

5.3 Uncorrelatedness, Unresponsiveness

Another idea that can be involved in causal reasoning is *causal uncorrelatedness* [18]; where if two variables have no common cause they are causally uncorrelated. This occurs if there are no single events that cause them to both change.

Similarly, Dawid [2] speaks in terms of *unresponsiveness* and *insensitivity* when β is unresponsive to α whatever the value of a might be set to, β will be unchanged. Along the same vein, Shoham [19] [20] distinguishes between *causing, enabling,* and *preventing*.

5.4 Graph Based Methods

Various graph-based methods have been suggested to recognize causality. Probably the best known is the class of Bayesian based methods based on Directed Acyclic Graphs (DAGs). The most fully developed approach is Pearl [13]. Silverstein [21] [22] followed a similar approach. (Other graph-based methods include [Khoo, 2000] [5] [Pogliano, 1995].) The constraints on DAGs are:

- The causal relationship is acyclic
- The Markoff condition holds
- Granularity is fixed
- Significant random events do not occur
- Data and events can be precisely and unambiguously described

Pearl [12] and Spirtes [25] make the claim that it is possible to infer causal relationships between two variables from associations found in observational (non-experimental) data without substantial domain knowledge. Spirtes [25] claims that directed acyclic graphs can be used if (a) the sample size is large and (b) the distribution of random values is faithful to the causal graph. Robins [15] argues that this is incorrect. Lastly, Scheines [17] claims that only in some situations will it be possible to determine causality. Besides the constraints on the DAGs, the directed graph methods all have similar liabilities, specifically:

- *Discrete or continuous data must be reduced to Boolean values.*

  Objection: This is an early technique that was and is used in data mining when analyzing market basket data. However, it is essentially flawed. Quantities do matter; some data co-occurrences are conditioned on there being a sufficiency of a co-occurring attribute. Also, some relationships may be non-linear based on quantity. For more extensive examples, see Mazlack [9]

- *Completeness: There is no missing data.*

  Objection: There is almost always missing data of some sort. Data collection is rarely fully representative and complete. Incremental data is often acquired that is at variance with previously acquired data. What is needed is a methodology that is not brittle in the face of incompleteness.

5.5 DAG Liability: Cyclic Graphs Exist

Directed Acyclic Graphs (DAGs) only work if causal relationships are not cyclic, either directly or indirectly (through another attribute).

This is at variance with our commonsense understanding of the world. While some causal situations may be acyclic, clearly others are cyclic.

General cycles can occur [6] [8]. Non-cyclic elements can influence cyclic elements. Depending on the conditioning of the cyclic nodes, the causal path might remain within the cycle, or it might branch out. There may or may not be a cumulative effect (feedback).

Figure 3 illustrates a cumulative effect cycle that has external input. The depression cycle could also be linked to a cycle increasing the *significant other's lack of interest* as *depression* increased. (This would also be another factor that would increase the *depression* intensity.)

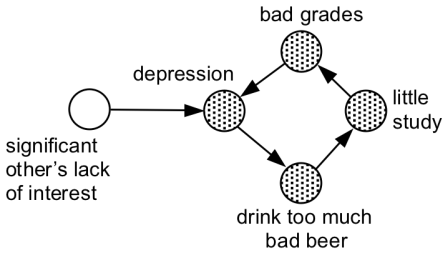

Figure 3. Cyclic causal dependency with cumulative effect.

Possibly the cycle might be collapsed to an imprecise larger grained single node labeled ***depression***. The down side of doing this is that the cyclic nature of ***depression*** would be obscured. As indicated earlier, the price of greater grain size, is greater imprecision.

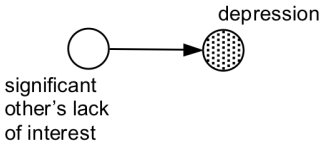

Figure 4. Previous figure's cycle nested into greater-grained representation.

5.6 Problems Meeting The Markov Conditions

There are situations where the Markoff time independent and memoryless conditions are not met, and consequently DAGs cannot be used.

- *Markov Stationary Condition* holds (Probabilities are time independent).

  Objection: This does not correspond to our commonsense understanding of the world. If one event is dependent on two other causal events, if one causal event happens much earlier (or later) than the other causal event, we may well have a different result.

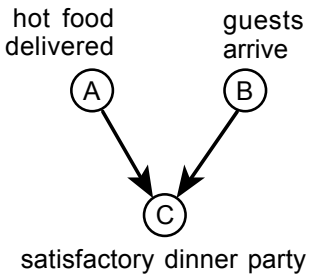

Figure 5. Case where differing times in causal events affects probability of causal result.

- The *Markoff Condition (Memoryless)* holds: Let *A* be a node in a causal Bayesian network, and let *B* be any node that is not a descendant of *A* in the network. Then the Markoff (Markov) condition holds if *A* and *B* are independent, conditioned on the parents of *A*. The intuition of this condition is: If *A* and *B* are dependent, then *B* must either be (a possibly indirect) cause of *A* or (possibly indirectly) caused by *A*. In the second case, *B* is a descendant of *A*, while in the first *B* is an ancestor of *A* and has no effect on *A* once *A*'s immediate parents are fixed. This makes sense in the example shown in Figure 6.

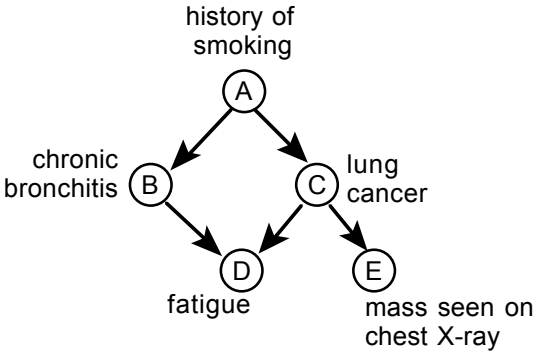

Figure 6. "Memoryless" Markoff condition holds

However, not all commonsense perceptions of causality work this way. Often, we believe that history matters as in the presumptive stereotypical example shown in Figure 7.

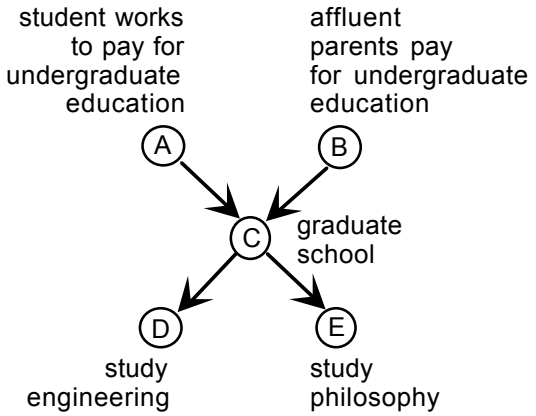

Figure 7. Causality where memory plays a part.

6. LONG TERM GOALS

The *long term goal* of the work is to discover (data mine) causal relationships in stored, observational data.

There are several *specific aims:*

- *Determine whether cyclic causal relationships can be recognized.*

  It has proven to be difficult to recognize any causal relationships, acyclic or cyclic, so far there are *no successful efforts to identify* cyclic causal relationships. It is clear that they exist. However, it is not clear whether a causal cycle can be recognized from data, even if it exists.

- *Determine whether optimum grain size can be discovered:*

  Discover whether it is possible to (a) discover whether the grain size of causal complexes can be recognized and subsequently (b) whether an optimum grain size can be determined.

- *Determine whether undefinability can be managed:*

  Can issues of undefinability and imprecision be recognized and defined? Can it be recognized when specific tools such as fuzzy sets or rough sets can be used to manage imprecision?

These issues are distinct in the sense that they involve at least potentially distinguishable questions. Resolving one does not require resolving another; however, resolving at least one could lead to a worthwhile follow-on investigation.

7. Epilogue

New methods are required that create knowledge and understanding from an abundance of digital data across the science and engineering frontier, and that accelerate the transformation of knowledge into new products and services that stimulate economic growth as well as other societal benefits.

Recognizing causality in very large digital data sets would place us on the threshold of a transformation in our understanding of the world around us. This promises a profound impact on the ability to generate and apply new knowledge. In addition, this will stimulate further advances in computational thinking.

The research outcomes will produce paradigm shifts in our understanding of a wide range of science and engineering phenomena and socio-technical innovations.

References

[1] G. Cooper [1997] "A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationships," *Data Mining and Knowledge Discovery,* v1, n 2, 203 - 224

[2] A. Dawid [1999] "Who Needs Counterfactuals" in **Causal Models and Intelligent Data Management** (ed) A. Gammerman) Springer-Verlag, Berlin

[3] J. Hobbs [2001] "Causality," *Proceedings, Common Sense 2001, Fifth Symposium on Logical Formalizations of Commonsense Reasoning*, New York University, New York, May, 145-155

[4] J. Hobbs [2003] "Causality And Modality: The Case Of 'Would'," to appear in *Journal of Semantics*

[5] S. Lauritzen [2001] "Causal inference from graphical models", D. C. O.E. Barndor - Nielsen, C. Kluppelberg (eds.), **Complex Stochastic Systems**, Chapman and Hall/CRC, London/Baton Rouge, 2001, 63-107

[6] L. Mazlack [2003a] "Commonsense Causal Modeling In The Data Mining Context," IEEE ICDM Proceedings, Melbourne, Florida, November 19 - 22, 2003

[7] L. Mazlack [2003b] "Causality Recognition For Data Mining In An Inherently Ill Defined World," 2003 BISC FLINT-CIBI International Joint Workshop On Soft Computing For Internet And Bioinformatics, December, 2003

[8] L. Mazlack [2004a] "Causal Satisficing And Markoff Models In The Context Of Data Mining," NAFIPS 2004 Proceedings, Banff

[9] L. Mazlack [2004b] "Naïve Decisions Using Rules That Do Not Take Into Account The Underlying Causality," IEEE International Conference on Data Mining (ICDM'04), Workshop: Foundations of Data Mining

[10] L. J. Mazlack [2005] "Granular Nested Causal Complexes," book chapter in: Intelligent Data Mining Techniques and Applications: Studies in Computational Intelligence v 5, D. Ruan, C. Chen, E. Kerre, G. Wets (eds), July, 2005, 23-49

[11] J. Mendel [2000] **Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions**, Prentice Hall, New Jersey

[12] J. Pearl, T. Verma [1991] "A Theory Of Inferred Causation," *Principles Of Knowledge Representation And Reasoning: Proceedings Of The Second International Conference,* Morgan Kaufmann, 441-452

[13] J. Pearl [2000] **Causality**, Cambridge University Press, New York, NY

[14] Plato [1981] **Five Dialogues, Euthyphro, Apology, Crito, Meno, Phaedo,** (note: Zeno lived about 428 B.C.) Hackett Publishing Company, 6-22

[15] R. Robins, L. Wasserman [1999], "On The Impossibility Of Inferring Causation From Association Without Background Knowledge," in (eds) C. Glymour, G. Cooper, **Computation, Causation, and Discovery** AAAI Press/MIT Press, Menlo Park, 305-321

[16] Schaffer, J. [2003] "Overdetermining Causes," *Philosophical Studies,* v 114, 23-45

[17] R. Scheines, P. Spirtes, C. Glymour, C. Meek [1994] **Tetrad II: Tools For Causal Modeling**, Lawrence Erlbaum, Hillsdale, NJ C.

[18] G. Shafer [1999] "Causal Conjecture," in **Causal Models and Intelligent Data Management** (ed) A. Gammerman) Springer-Verlag, Berlin

[19] Y. Shoham [1990] "Nonmonotonic Reasoning And Causation," *Cognitive Science,* v14, 213-252

[20] Y. Shoham [1991] "Remarks On Simon's Comments," *Cognitive Science*, v15, 301-303

[21] C. Silverstein, S. Brin, R. Motwani [1998a] "Beyond Market Baskets: Generalizing Association Rules to Dependence Rules," *Data Mining and Knowledge Discovery,* v 2, n 1, 39-68

[22] C. Silverstein, S. Brin, R. Motwani, J. Ullman [1998*b*] "Scalable techniques for mining causal structures," *Proceedings 1998 VLDB Conference,* New York, NY, August 1998, 594-605

[23] H. Simon [1955] "A Behavior Model Of Rational Choice," *Quarterly Journal of Economics,* v 59, 99-118

[24] H. Simon [1991] "Nonmonotonic Reasoning And Causation: Comment," *Cognitive Science,* v 15, 293-300

[25] P. Spirtes, C. Glymour, R. Scheines [1993] **Causation, Prediction, and Search,** Springer-Verlag, New York

[26] C. Winship, S. Morgan [1999] "The estimation of Causal Effects From Observational Data," Annual Review of Sociology, v 25, 659-706

[27] L. Zadeh [2002] "A New Direction In AI: Toward A Computational Theory Of Perceptions," **Technologies For Constructing Intelligent Systems: Tasks**, Physica-Verlag GmbH, Heidelberg, Germany, 3-20

[28] Zeno [2001] **Zeno's Paradoxes**, (note: Zeno lived about 545 B.C.), Hackett Publishing Company