Causal Inference with Rare Events in Large-Scale Time-Series Data

Samantha Kleinberg
Stevens Institute of Technology
Hoboken, NJ
samantha.kleinberg@stevens.edu

Abstract

Large-scale observational datasets are prevalent in many areas of research, including biomedical informatics, computational social science, and finance. However, our ability to use these data for decision-making lags behind our ability to collect and mine them. One reason for this is the lack of methods for inferring the causal impact of rare events. In cases such as the monitoring of continuous data streams from intensive care patients, social media, or finance, though, rare events may in fact be the most important ones – signaling critical changes in a patient's status or trading volume. While prior data mining approaches can identify or predict rare events, they cannot determine their impact, and probabilistic causal inference methods fail to handle inference with infrequent events. Instead, we develop a new approach to finding the causal impact of rare events that leverages the large amount of data available to infer a model of a system's functioning and evaluates how rare events explain deviations from usual behavior. Using simulated data, we evaluate the approach and compare it against others, demonstrating that it can accurately infer the effects of rare events.

1 Introduction

With the increasing availability of big data from the social sciences, finance, health, and climatology among other areas, rare events are increasingly common. Given the amount and granularity of observations (ICU patients are monitored every 5 seconds during their stays, Twitter users generate 12TB of data per day and tick by tick stock data is at the scale of milliseconds), it is often the case that we want to know not only how a system works, but to explain changes from usual functioning. This is particularly critical given the streaming nature of many of these datasets, where the ultimate goal is to monitor them in real-time in order to alert users when something unusual has occurred that they may be able to act on, whether this is an adverse event for a patient, the potential for political riots, or an arbitrage opportunity.

Analyzing large amounts of data to uncover rare events has traditionally been a focus of data mining and machine learning efforts aimed at problems such as detecting credit card fraud and network intrusions [Chandola *et al.*, 2009]. To act on this information, though, we need to know not just that a rare event has occurred, but whether it will have an impact on the functioning of a system. For example, doctors and nurses in intensive care units (ICUs) are currently overloaded with information from a large array of monitoring systems, and want to know not only when something out of the ordinary has occurred (when a patient is not continuing to function as expected) but that this is actionable knowledge that will influence how they treat the patient. The fact that a patient is having a seizure is not, for example, useful information unless doctors know whether or not the seizure will cause further harm (since treatment can have severe side effects in critically ill patients). Thus we need to determine which rare events are causal (rather than say, outliers or measurement artifacts) and what their effects are. However, causal inference methodologies [Pearl, 2000; Granger, 1980] have primarily taken a probabilistic approach, which is ineffective when dealing with rare events whose probabilities cannot be calculated in a statistically significant way.

This paper introduces a new method, ARC (assessment of rare causes), that evaluates rare events by determining how well they account for deviations from normal behavior (predicted by an inferred model). We demonstrate that this approach can find rare causes occurring as few as two times in thousands of observations when they have a substantial impact, and can in some cases identify even weak rare causes.

2 Background

The approach described here first requires inference of a system's usual behavior before quantifying deviations from this and how they may be explained by a rare event. To do this we build on efforts to infer complex causal relationships in time series data [Kleinberg and Mishra, 2009; Kleinberg, 2012]. The idea is that instead of inferring a model that explains all independencies in a set of data, we can accept or reject each causal relationship individually with a measure of its strength that is based on how much of an effect's value each cause accounts for. In this work we focus on continuous-valued effects (which are prevalent in medicine), and use a method that evaluates causal relationships using conditional expectation rather than conditional probability [Kleinberg,

2011].

In this approach, cause and effect are represented as logical formulas of the form:

$$c \leadsto^{\geqslant r,\leqslant s} (e > E[e]), \tag{2.1}$$

which means that after c occurs, the value of e will be greater than expected in between r and s time units. Note that c and e may themselves be complex logical formulas (such as conjunctions or sequences of factors). Methods for testing these formulas (based on model checking and verification [Chan *et al.*, 2005; Clarke *et al.*, 1999]) in data are linear in the size of the formula, allowing for efficient evaluation of potentially complex relationships.

The basic idea of the approach is to generate a set of potential causes and then determine the causal significance of each using the average difference in conditional expected value after holding fixed other potential causes of the same effect.

Definition 2.1. c is a *potential cause* of e if, with c being earlier than e: $E[e|c] \neq E[e]$.

To determine the significance of a potential cause, c, for an effect, e, we calculate:

$$\varepsilon_{avg}(c, e) = \frac{\sum_{x \in X} E[e|c \wedge x] - E[e|\neg c \wedge x]}{|X \backslash c|}, \tag{2.2}$$

where the set X is composed of potential causes of e. Note that there are time windows associated with each relationship between x and e, and c and e (as shown in equation (2.1)), and that $c \wedge x \wedge e$ in this context means that the windows for c and x's occurrences at particular times overlap such that they have a non-zero intersection and e occurs in that intersection.

We can now define significant and insignificant causes. Note that the terminology used [Kleinberg, 2011; 2012] is not genuine/spurious but rather insignificant/significant. An insignificant cause may be either a spurious cause or simply a very weak genuine one. Similarly, a significant cause can only be guaranteed to be genuine in the case where (akin to the assumptions made for BNs) all common causes of c and e are included in the set X (and one's belief in whether it is genuine should be proportional to their belief that this assumption holds)[1]. In general, inferences from observational data are best viewed as a targeted set of hypotheses to later be validated experimentally.

Definition 2.2. A potential cause c of an effect e is an ε-insignificant cause of e if $|\varepsilon_{avg}(c, e)| < \varepsilon$.

Definition 2.3. A potential cause c of an effect e that is not an ε-insignificant cause of e is an ε-significant or *just-so* cause of e.

Intuitively the significance measure tells us, relative to other possible pieces of information, how valuable c is for determining the value of e. The larger ε_{avg}, the more of e's value is explained by c. For instance, if c and e are uncorrelated, this value will be near zero. On the other hand, if c and e have a common cause, d, the value will be nonzero but much lower than that for d. In this work the word "causes" will be used as a shorthand, but should be interpreted in the sense of ε-significant.

After calculating the ε_{avg} values it remains to determine a threshold for ε (to determine which values are statistically significant). In the absence of genuine causal relationships, these values follow a normal distribution (whose mean and standard deviation can be inferred empirically [Efron, 2004]), and methods for controlling the false discovery rate (FDR) under multiple hypothesis testing can be used to find a threshold at the desired FDR level [Efron, 2010].

3 Evaluating rare events

We now turn our attention to developing a new method for causal inference with rare events in large-scale data. The general approach is to use the large volume of data to infer how a system normally functions and then determine whether events not explained by this model can be explained by the occurrence of rare causes. There are three components of this process: 1) inferring a model of usual functioning (as described in the previous section); 2) determining for each variable how much of its value at each measured instance is explained by the model; and 3) calculating how explanatory the rare event is when it occurs. We begin with a discussion of the rationale behind the approach before discussing components 2 and 3 in section 3.2.

3.1 Linking type and token

The basis for the approach developed here is the link between type (also called general) and token (also called singular, or actual) causality. Broadly, type-level relationships are ones that describe the behavior of a system, such as side effects of medications or a gene regulatory network. Token-level relationships on the other hand relate to causes of particular events at specific points in time and space, such as the cause of an individual's cancer at age 42 or the U.S. recession that began in December 2007. The relationship between these "levels" of causality has been studied primarily in philosophy (focusing on the meaning of each and how they are conceptually related), with much less attention in computer science (exceptions include the work of Pearl and collaborators [Halpern and Pearl, 2005; Hopkins and Pearl, 2007]).

It has been argued alternately that type-level relationships are generalizations of token-level ones [Hausman, 2005], that token-level cases are specific instances of type-level ones [Woodward, 2005] and that these levels are distinct [Eells, 1991]. Regardless of what the true underlying relationship is between type and token (or if there is one at all), we can make use of the idea that a type-level relationship is inferred by computational means specifically because it seems to be observed many times [Kleinberg, 2012]. Thus, these type-level relationships provide us with an expectation of what should occur in individual instances (though these expectations may not necessarily be fulfilled).

We exploit this link to aid in distinguishing between expected and unusual events, where instead of detecting sta-

[1] Note that in practice we avoid many spurious inferences due to the inclusion of temporal information. In order to incorrectly find a causal relationship between two effects of a hidden common cause, one would have to regularly precede the other in a stable window of time. Latent variables in the form of factors are included in the simulated data developed here and in [Kleinberg, 2011].

tistical anomalies, we determine whether events are causally explained given our prior inferences. While an event that is fully explained by inferred type-level causes is not likely to be an effect of a rare event, one that deviates significantly from what is expected after its occurrences may be.

3.2 Calculating the impact of a rare event

The average difference in conditional expected value, shown in equation (2.2), indicates what value of an effect should be observed when a cause actually occurs. That is, if we have inferred all of the causes of an effect, and assume their influence is additive, we should be able to sum the ε_{avg} values of actually occurring causes at any given time in order to determine what value of an effect will be observed. Holding fixed the other potential influences (other possible causes) when calculating the average difference isolates each cause's impact on the effect. For example, if $c \rightsquigarrow^{\geqslant 1,1} e$ with $\varepsilon_{avg}(c, e) = 5$, and where c is the only known cause of e, then the expected value of e after each time c actually occurs is 5. Using this approach, we can calculate how much of a variable's value at each observed time is not accounted for by the inferred relationships. When the result is non-zero, this means that there are influences other than the known causes, such as rare events (that cannot be assessed using the probabilistic methods discussed) and unmeasured (latent) variables.

More formally, we define the unexplained value of a variable as follows.

Definition 3.1. The *average unexplained value* of a continuous-valued variable, e, is the average difference between its actual and expected values relative to a time series T and set of type-level causes, R (inferred by the approach in section 2). It is given by:

$$u(e) = \frac{\sum_t e_t - E[e_t]}{\#e}, \tag{3.1}$$

where $\#e$ is the number of measurements of e in T, e_t is the value of e measured at time t and we sum over all measured values ($t \in T$).

The expected value of e at time t, $E[e_t]$, is defined as:

$$E[e_t] = \sum_{a \in A_t} \varepsilon_{avg}(a, e) \tag{3.2}$$

where each a is a type-level cause of e in R and

$$A_t = \{a : a \rightsquigarrow^{\geqslant r, \leqslant s} (e > E[e]) \wedge \exists a_{t'} : t' \in [t-s, t-r]\}. \tag{3.3}$$

The set A_t is composed of causes of e that actually occur before the instance of it at time t, consistent with their known time windows (so that if c causes e in one time unit, e_t is preceded by c_{t-1}).

The impact of a rare event cannot be well approximated with ε_{avg} (equation (2.2)) as this measure relies on there being a sufficient number of occurrences of a particular cause, c, and its absence (or negation), $\neg c$, along with each x. Further the calculation of ε_{avg} uses probabilities of more complex events, namely the conjunction of multiple events. These events will be at least as infrequent as c and some may not be observed at all.

Instead, to evaluate infrequent events we can compare the average unexplained value of an effect after such an event's occurrence to the overall average. This approach enables calculation of the impact of a rare event and determination of its statistical significance without the need for frequency-based probabilities. We define the average unexplained value of one variable conditioned on the occurrence of another as follows. As with the inferred relationships, there is a hypothesized time window[2] when the rare event may cause the effect, where $1 \leqslant r'' \leqslant s'' \leqslant \infty$, and $r'' \neq \infty$.

Definition 3.2. The *average conditional unexplained value* of a variable e given another variable v is the average difference between e's actual and expected values after each instance of v. It is given by:

$$u(e|v) = \frac{\sum_t e_t - E[e_t|v]}{\#e \wedge v}, \tag{3.4}$$

where the time window associated with v potentially causing e is $[r'', s'']$ and the times summed over are those where v has occurred prior to t (that is, $t : e_t \wedge v_{t''}, t'' \in [t-r'', t-s'']$).

The conditional expected value of e at time t given v is defined as:

$$E[e_t|v] = \sum_{a \in A'_t} \varepsilon_{avg}(a, e) \tag{3.5}$$

where again each a is a type-level cause of e and:

$$\begin{aligned} A'_t = \{a : a \rightsquigarrow^{\geqslant r, \leqslant s} (e > E[e]) \\ \wedge \exists a_{t'} : t' \in [t-s, t-r] \\ \wedge \exists v_{t''} : t'' \in [t-r'', t-s'']\}. \end{aligned} \tag{3.6}$$

This is similar to the unconditional unexplained value, but we now sum the difference in expected versus actual value over only the instances of e where the rare event v has occurred in the window of time $[r'', s'']$ before the effect. The number of instances of $e \wedge v$ in the denominator is the number of times e occurs after v.

The value $u(e|v)$ in equation (3.4) allows us to compare the overall amount of an effect that is unexplained to that that is unexplained after the rare event. A rare event with no genuine impact will lead to the same value as the overall average, ensuring this approach does not incorrectly find a rare event to be significant due to unmeasured causes or background conditions, though statistically significant differences should still be interpreted in the sense of being ε-significant (rather than genuine) for the same reasons discussed earlier. Note that this approach can handle a non-rare latent common cause, due to the comparison against the overall unexplained value. There could potentially be cases with rare latent common causes, but as discussed in section 2 to result in incorrect inferences this requires stable ordering and timing of the common effects. Algorithms for inference follow directly from equations (3.1) and (3.4).

3.3 Computational complexity

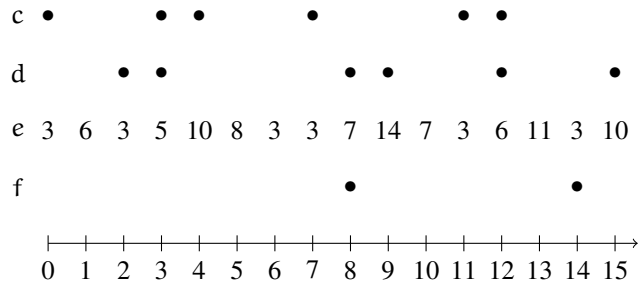
Determining how much of a variable's value is unexplained (calculation of eq. (3.1)) is linear in the number of times it

[2]Other work has shown how to infer this time window without any prior knowledge of it [Kleinberg, 2012].

occurs and bounded by the length of the observation series, T. Evaluation of a rare event as a potential cause of a particular effect is linear in the number of times the rare event occurs and by definition a rare event is infrequent. Testing whether V rare events are causes of N variables is then $O((V+T)N)$ though the major component of this is the term TN. While other approaches for inference of the normal model may be used, the one discussed here is $O(N^3T)$, where N is the number of variables in the system and T is the length of the time series (N^2 of these computations are independent and can proceed in parallel).

3.4 Example

To illustrate how this approach works, take the following simplified example and sequence of observations:



Here c, d, and f are discrete variables and e is continuous. The rare event, f, is shown occurring more frequently than such an event ordinarily would in order to better illustrate the calculations. The sequence shown is assumed to be part of a much longer time series, from which the relationships $c \rightsquigarrow^{\geqslant 1,\leqslant 1} e$ and $d \rightsquigarrow^{\geqslant 1,\leqslant 1} e$ were inferred. Assume it was found that $\varepsilon_{avg}(c,e) = 4$ and $\varepsilon_{avg}(d,e) = 3$.

The expected value of e at a particular timepoint given that both c and d have occurred at the previous timepoint is the sum of their influence: 7. If e instead has the value 10 (as at time 4 above):

$$e_4 - E[e_4] = 10 - 7 = 3. \tag{3.7}$$

This means that 3 is how much of e's value is "unexplained" by what is known. It may be that e's value is a function of its causes plus a constant (i.e. 3), or that there are unmeasured causes of it. At time 12, e is preceded by only c, making its expected value 4. Since it takes the value 6, the difference is 2. Doing this for all instances of e yields the average of equation (3.1). For e in this sequence, removing the timepoints[3] immediately after f, $u(e)$ is $42/14 = 3$.

To determine whether f is significant, we find $u(e|f)$, the average unexplained value of e immediately after f and compare this to the overall average. This value for instances following one time unit after f is $(11+10)/2 = 10.5$ since d occurs before e at time 9 and no other causes occur right before e at time 15. Statistical techniques for hypothesis testing can then be applied to determine whether such a difference is significant. In this case the p-value using an unpaired t-test is significantly less than 0.01, so f does indeed make a significant difference to e when it occurs. When a factor's results are hypothesized to be permanent, one may instead test how well the event explains all future values rather than only those immediately following it.

[3] As the length of the time series increases, this step is less important, as the few instances of the rare event will have less of an impact on the overall average. Failing to do this, though, will only lead to an underestimate of the rare event's influence.

Probability	Exp. 4K	Act. 4K	Exp. 10K	Act. 10K
0.005	20	20.44	50	51.03
0.0025	10	9.88	25	24.41
0.0005	2	2.35	5	4.91

Table 1: Number of expected and mean actual occurrences of each rare event in the 4,000 and 10,000 day time series.

4 Experimental results

4.1 Simulated data

To validate the approach developed here and compare it to other methods, it was applied to simulated financial time series data where two primary types of rare causes were embedded. These are 1) causes that lead to a fixed increase in the price changes of their effects; and 2) causes such that the effect's value is a function of the cause's value. In both cases, the data consist of 25 variables with five different causal structures (two with 10 causal relationships in the system and three with 20) that additionally include 1 or 3 rare causes. Data was generated for two time periods with each of two observation lengths (4,000 and 10,000 timepoints) while varying the probabilities of occurrence of rare events (P= 0.005, 0.0025, or 0.0005). The expected number of occurrences and mean number of occurrences (averaged over all datasets with the same probability and observation length) are reported in table 1. For each parameter setting (structure, type of causality, probability, time period) four runs of the system were created. This resulted in 480 datasets (5 structures $\times$ 2 types of causes $\times$ 2 time periods $\times$ 2 observation lengths $\times$ 3 probabilities $\times$ 4 runs).

The time series were generated using the approach of [Kleinberg, 2011] following a Fama-French [1993] factor model, where the return for stock portfolio i at time t is given by:

$$r_{i,t} = \sum_j \beta_{i,j} f_{j,t} + \varepsilon_{i,t}. \tag{4.1}$$

Here $f_{j,t}$ is the value of factor j at time t and $\beta_{i,j}$ is the weighting of factor j for portfolio i. The ε terms are portfolio specific error terms. Causality is embedded in two ways. First, if portfolio i influences portfolio j at a lag of 1 day, then $r_{j,t} = r_{j,t} + \varepsilon_{i,t-1}$. This is how the non-rare model (with 10 or 20 such relationships depending on the dataset) is embedded. In the case of rare causes where the effect is a function of the cause (i.e. it has a non-constant influence) then when the rare event occurs, it influences the effect's price movement in the same way. In the second case, when the rare cause has a constant (and substantial) influence then where $c_{i,j}$ is the influence of portfolio i on portfolio j, when event i occurs in the time window before t then $r_{j,t} = r_{j,t} + c_{i,j}$. Note that the factor time series are not included in the data used for inference.

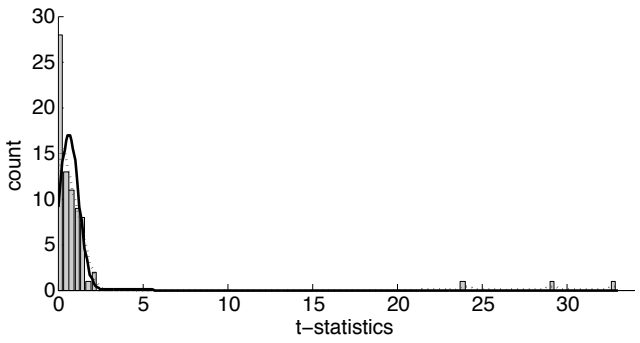


Figure 1: Histogram of t-statistics for one 4K timepoint constant influence dataset. Solid line depicts empirically inferred null distribution, dashed line shows fit to graph. The statistically significant results (all genuinely causal) have $t > 20$.

4.2 Methods

The approach described in this paper, referred to as ARC (assessment of rare causes), was compared against two others for inferring relationships in time series data: bivariate Granger causality [Granger, 1980] and dynamic Bayesian networks (DBNs) [Murphy, 2002]. We are not aware of any methods that deal specifically with evaluating rare causes, and did not compare against data mining methods as these would only identify the rare causes, not the relationship between them and the other variables.

ARC used the approach described here to evaluate the rare events along with the method developed in [Kleinberg, 2011] for inferring the normal model. After calculating the unexplained and conditional unexplained values, an unpaired t-test was used to determine the statistical significance of each relationship. The locfdr package [Efron *et al.*, 2011] was used to infer the null distribution empirically (the theoretical null $N(0,1)$ was used when there were too few results to infer it). The FDR was controlled at 0.01 using the local fdr approach [Efron, 2010]. Figure 1 shows a histogram of significance scores from one dataset along with a fit to them and the inferred null distribution.

Bivariate Granger causality was tested using the granger.test function in the MSBVAR R package [Brandt, 2012] along with fdrtool (using a significance level of 0.01) [Strimmer, 2008] to determine which of the inferred relationships are statistically significant.

The Banjo implementation of DBN structure learning was used with the default settings and a run time of 1 hour with 6 threads [Hartemink, 2008]. This required the data to be discretized so returns, which are relative to the previous day's value, were transformed to be either up or down.

The FDR (false discovery rate) and FNR (false negative rate) are compared using only the embedded rare events, and all methods used a lag between cause and effect of exactly one day. The FDR is the proportion of incorrect effects of the rare events that are identified out of the total number of effects identified. The FNR is the proportion of effects that are missed out of all effects of the rare events.

Method	FDR_c	FNR_c	FDR_f	FNR_f
ARC	0.0000	0.1553	0.0000	0.9671
DBN	0.4875	0.379	0.9130	0.9394
bivariate Granger	0.1198	0.1932	1.000	1.000

(a) 4,000 timepoints

Method	FDR_c	FNR_c	FDR_f	FNR_f
ARC	0.0076	0.0151	0.0833	0.9583
DBN	0.1830	0.2727	0.7903	0.9508
bivariate Granger	0.0121	0.0720	1.000	1.000

(b) 10,000 timepoints

Table 2: Results are broken into two cases: 1) rare events with a constant impact (FDR_c and FNR_c) and 2) where the effect is a function of the value of the rare cause (FDR_f and FNR_f).

4.3 Results

Empirical results are shown in tables 2a and 2b.

Constant influence datasets

First, when the rare cause has a constant, significant, influence on its effect, then regardless of the length of the time series, ARC makes very few false discoveries (FDR 0, and 0.0076 in the 4K and 10K timepoint datasets respectively), below the 0.01 level at which the FDR was controlled. The primary difference is that with more data points, there are fewer false negatives. Note that the FNR is calculated based on whether a relationship with a rare cause embedded in a dataset is identified. There was no correction for cases where the rare cause occurred only once (making it such that a t-statistic could not be calculated). This measure is thus extremely strict. This situation did not happen in the 10K long time series, but happened 19 times in the 4K time series with constant influence (and 24 times in the 4K time series with functional relationships). The effective false negative rate (adjusting the denominator to only count events that occur at least twice and can thus possibly be found by this approach) for 4K timepoints with constant influence is approximately 9%. On the other hand, both bivariate Granger causality and DBNs have significant FDR and FNR rates (particularly in the shorter time series). Bivariate Granger causality outperforms DBNs in both cases, though both improve with more data.

Functional influence datasets

While the FDR for ARC remains low in both cases (0 and 0.0833), one difference in results between the datasets with constant impact and a functional relationship is the performance of Granger causality. It fares better than DBNs on the constant influence datasets, but this is reversed on the functional datasets, where its FDR and FNR are 1 for both data lengths. The FDR achieved by DBNs improves somewhat with more data, but is still over 79%. This scenario is extremely challenging, since here a rare event may occur only 2-5 times and may only have a small influence (as it simply adds its value to that of the effect). The FNR is in fact increased in such cases even with non-rare events [Kleinberg, 2012]. Thus while extremely infrequent rare events (that occur only 1-2 times) with a small influence may be missed, we

can be confident that incorrect effects will not be falsely identified and that in some cases we can even find these rare and less significant causes.

5 Related work

5.1 Data mining

Much of the prior work on rare events[4] has come from data mining and information theory [Chandola *et al.*, 2009; Szathmary *et al.*, 2007; Weiss, 2004]. However, the focus has been on accurately identifying rare events (or classes of rare events [Joshi *et al.*, 2001]) such as network intrusions [Lee and Xiang, 2001], disease outbreaks [Wong *et al.*, 2003], or credit card fraud [Aggarwal and Yu, 2001]. This is also referred to as anomaly detection. These approaches do not find the causal relationship between the anomaly and the rest of the variables and thus cannot determine a) why the rare event occurred and b) what its implications are. If we do not aim solely to identify these outlying events but rather to determine if and how to act based on them (such as in order to improve medical treatment or create public health policies), we need additional causal information. For example, in order to have a beneficial impact on a patient's disease process, we must target interventions at underlying causes and not mere downstream effects, as targeting these is not only ineffective but may be harmful (many therapeutic interventions also carry non-negligible risks). However, for datasets with vast numbers of rare events, this prior work may potentially be used to reduce the computational complexity by selecting which events should be examined further using the algorithms developed here.

5.2 Causal inference

Thus far no causal inference methods have specifically addressed the challenge of evaluating rare events. The most similar work to that proposed here is on general methods for causal inference, including Bayesian [Pearl, 2000; Spirtes *et al.*, 2000] and dynamic Bayesian [Murphy, 2002] networks. Causal inference methods based on Bayesian networks (BNs) use graphical models (along with other assumptions about the structures and data from which they are inferred) to represent causal relationships using conditional independencies in a graph. Edges are directed from cause to effect and a node in the graph (variable) is independent of all of its non-descendants given its parents. Methods for inferring these structures take two main approaches: adding nodes one at a time using repeated conditional independence tests [Spirtes *et al.*, 2000], or searching the space of possible graphs for a set of variables and scoring how well each accounts for the independencies in the data [Cooper and Herskovits, 1992]. However, the conditional probabilities associated with the rare events cannot be accurately estimated, and rare events will have little impact on a graph's score (particularly when a rare event's effect has other, non-rare, causes), making it difficult to accurately infer these relationships.

[4]There is no single probability threshold for what makes an event rare. In some cases 5% is considered rare, though values closer to .05-.5% are also used.

Granger [1980] causality, an econometric method that determines whether one time series is predictive of another at various time lags after accounting for other possible information, has been applied to continuous time series data outside of finance more generally but its more accurate multivariate form is computationally complex while the bivariate form may erroneously find relationships between effects of a common cause [Eichler, 2006]. Granger causality is often evaluated using vector autoregressive (VAR) models and testing whether inclusion of a variable with non-zero coefficients in the regression leads to less variance in the error term. However, unless the rare event has a consistently large influence on the value of the effect and occurs often enough that failure to account for these instances will affect the VAR model, the rare event will not be identified as a cause. Extensions to Granger causality, such as those linking Granger causality and graphical models [Eichler and Didelez, 2007], face these same challenges.

6 Conclusions

Being able to automatically detect the impact of rare events is critical for using big data for decision-making, and particularly for enabling this to be done in real-time. Prior methods for causal inference based on graphical models, Granger causality, and logical formulas have taken a probabilistic approach, making them unable to reliably infer causal relationships involving these infrequent occurrences. On the other hand, data mining approaches do not solve the problem of causal inference, and cannot identify the impact of a rare event. While prior approaches have aimed to predict and identify rare events, we must know what their effects will be in order to determine what action should be taken. Many rare events (such as extreme laboratory values, or seizures) occur in medical settings, yet clinicians must weight the relative risks and benefits of treatment and thus need to know whether they are targeting a symptom or cause.

To address this problem we have developed a new approach, called ARC, for assessing the impact of rare events. Using the idea of rare events as explanations for deviations from usual behavior, we first infer a causal model of a system, then compare how much of the value of variables is unexplained by the model to how much is unexplained conditioned on the occurrence of a rare event. The approach was evaluated on simulated data with two types of rare events (those having a consistent and large impact and those with a more subtle influence), showing that this method leads to significantly lower FDR and FNR rates than approaches not specifically developed for rare events. Future work will involve extending this approach to non-linear relationships (where causes can interact and have non-additive results), and potentially to discrete effects. It may also be possible to build on the idea of how much of an effect's value a cause explains to develop new approaches for detecting latent variables.

References

[Aggarwal and Yu, 2001] C. C. Aggarwal and P. S. Yu. Outlier Detection for High Dimensional Data. *ACM Sigmod Record*, 30(2):37–46, 2001.

[Brandt, 2012] P. Brandt. MSBVAR R package 0.7-2. R package, 2012.

[Chan *et al.*, 2005] K. Chan, I. Poernomo, H. Schmidt, and J. Jayaputera. A Model-Oriented Framework for Runtime Monitoring of Nonfunctional Properties. *Lecture Notes in Computer Science*, 3712:38, 2005.

[Chandola *et al.*, 2009] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15, 2009.

[Clarke *et al.*, 1999] E. M. Clarke, O. Grumberg, and D. A. Peled. *Model Checking*. MIT Press, 1999.

[Cooper and Herskovits, 1992] G. F. Cooper and E. Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9(4):309–347, 1992.

[Eells, 1991] E. Eells. *Probabilistic Causality*. Cambridge University Press, 1991.

[Efron *et al.*, 2011] B. Efron, B. Turnbull, and B. Narasimhan. locfdr: Computes local false discovery rates. R package, 2011.

[Efron, 2004] B. Efron. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association*, 99(465):96–105, 2004.

[Efron, 2010] B. Efron. *Large-scale Inference : Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2010.

[Eichler and Didelez, 2007] M. Eichler and V. Didelez. Causal Reasoning in Graphical Time Series Models. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

[Eichler, 2006] M. Eichler. Graphical Modeling of Dynamic Relationships in Multivariate Time Series. In *Handbook of Time Series Analysis*, pages 335–372. Wiley-VCH, 2006.

[Fama and French, 1993] E. F. Fama and K. R. French. Common Risk Factors in the Returns on Stocks And Bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

[Granger, 1980] C. W. J. Granger. Testing for Causality: A Personal Viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

[Halpern and Pearl, 2005] J. Y. Halpern and J. Pearl. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, 2005.

[Hartemink, 2008] A. J. Hartemink. Banjo: Bayesian Network Inference with Java Objects 2.2.0, 2008.

[Hausman, 2005] D. M. Hausman. Causal Relata: Tokens, types, or variables? *Erkenntnis*, 63(1):33–54, 2005.

[Hopkins and Pearl, 2007] M. Hopkins and J. Pearl. Causality and Counterfactuals in the Situation Calculus. *Journal of Logic and Computation*, 17(5):939, 2007.

[Joshi *et al.*, 2001] M. V. Joshi, R. C. Agarwal, and V. Kumar. Mining Needles in a Haystack: Classifying Rare Classes via Two-Phase Rule Induction. In *ACM SIGMOD Record*, volume 30, 2001.

[Kleinberg and Mishra, 2009] S. Kleinberg and B. Mishra. The Temporal Logic of Causal Structures. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

[Kleinberg, 2011] S. Kleinberg. A Logic for Causal Inference in Time Series with Discrete and Continuous Variables. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, 2011.

[Kleinberg, 2012] S. Kleinberg. *Causality, Probability, and Time*. Cambridge University Press, 2012.

[Lee and Xiang, 2001] W. Lee and D. Xiang. Information-Theoretic Measures for Anomaly Detection. In *IEEE Symposium on Security and Privacy*, 2001.

[Murphy, 2002] K. Murphy. *Dynamic Bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkley, 2002.

[Pearl, 2000] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[Spirtes *et al.*, 2000] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

[Strimmer, 2008] K. Strimmer. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics*, 24(12):1461, 2008.

[Szathmary *et al.*, 2007] L. Szathmary, A. Napoli, and P. Valtchev. Towards Rare Itemset Mining. In *International Conference on Tools with Artificial Intelligence*, 2007.

[Weiss, 2004] G. M. Weiss. Mining with Rarity: A Unifying Framework. *ACM SIGKDD Explorations Newsletter*, 6(1):7–19, 2004.

[Wong *et al.*, 2003] W. Wong, A. Moore, G. Cooper, and M. Wagner. Bayesian Network Anomaly Pattern Detection for Disease Outbreaks. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[Woodward, 2005] J. Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, 2005.