Discovery of Causal Time Intervals

Zhenhui Li* Guanjie Zheng* Amal Agarwal† Lingzhou Xue† Thomas Lauvaux‡

Abstract

Causality analysis, beyond "mere" correlations, has become increasingly important for scientific discoveries and policy decisions. Many of these real-world applications involve time series data. A key observation is that the causality between time series could vary significantly over time. For example, a rain could cause severe traffic jams during the rush hours, but has little impact on the traffic at midnight. However, previous studies mostly look at the *whole* time series when determining the causal relationship between them. Instead, we propose to detect the *partial* time intervals with causality. As it is time consuming to enumerate all time intervals and test causality for each interval, we further propose an efficient algorithm that can avoid unnecessary computations based on the bounds of F-test in the Granger causality test. We use both synthetic datasets and real datasets to demonstrate the efficiency of our pruning techniques and that our method can effectively discover interesting causal intervals in the time series data.

1 Introduction

In recent years, driven by a wide range of real-world applications and scientific discoveries, detecting causality from the data has gained increasing attention in the data mining community. Time series data is a type of data frequently seen in many applications such as stock, traffic, and climate. Various methods have been developed to determine the causality between two time series. Among all causality test methods, Granger causality test [10] is one of the most popular methods [25, 2, 20, 18, 17].

In short, the idea of Granger test is based on *prediction*: Given a pair of time series $X = x_1 x_2 \dots x_T$ and $Y = y_1 y_2 \dots y_T$, two different predictive models, called the full model and the reduced model, are fitted using the **whole** time series (i.e., all data points). The full model uses the past values of X and the past values of Y to predict Y, whereas the reduced model uses past values of Y only to predict Y. Then, we say that X causes Y by Granger test if the full model is significantly better than the reduced model (i.e., it is critical to use the values of X to predict Y).

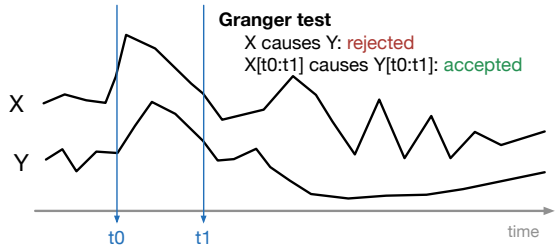


Figure 1: An example illustrating the causality in partial time series. The original Granger test fails to detect the causal relationship, since X causes Y only during the time interval $[t_0 : t_1]$. Our goal is to find such time intervals.

However, in real-world scenarios, causal relationships may exist only in **partial** time series. This is illustrated in Figure 1, where X causes Y only during the time interval $[t_0 : t_1]$. In such cases, if we apply the traditional Granger test (i.e., fitting the models using the whole time series), we may not detect any causality between X and Y. However, if we only test on the partial time series in interval $[t_0 : t_1]$, the causality can be discovered. Therefore, given two time series X and Y, our goal in this paper is to find the **causal time intervals** in which X causes Y. This is highly motivated by a number of real world applications. Consider the following two examples.

EXAMPLE 1.1. *Suppose that a local event (e.g., a football game) attracts many people, generating a large volume of traffic at the venue (e.g., a football stadium) during the event hours. In such scenario, time series X could be the frequency of geo-tagged tweets at the venue (as a proxy of the popularity of local events), and time series Y could be the taxi pick-up volume at the same location. Intuitively, we can say that the tweet frequency X (which represents the people attending the event) causes traffic volume Y* during the time of the event. *However, for the rest of the time, tweet frequency does not necessarily cause the pick-up volume. In fact, during the normal days, the pick-ups are mostly due to human routine behaviors (e.g., commuting between*

*College of Information Sciences and Technology, Pennsylvania State University. Email: jessieli@ist.psu.edu
†Department of Statistics, Pennsylvania State University.
‡Department of Meteorology, Pennsylvania State University.

home and work) and may not have any correlation with the frequency of geo-tagged tweets.

EXAMPLE 1.2. *Nowadays, environmental pollution is a big concern in people's daily life. For example, poor air quality in big cities in China [27, 28] and potential air/water contamination from shale gas development [15, 22] in the United States have drawn a lot of attention lately. Here, one biggest scientific question is* whether certain type of emissions (e.g., CO2) generated by industrial, vehicular, or other human activities causes the environmental change. *But the causal relationship between an emission source and the air quality may vary over time. For example, vehicle emissions react in the presence of sunlight and form ground level ozone, a primary ingredient of smog, whereas industrial emissions have a bigger impact on the air quality in cool and humid days.*

To detect all the casual time intervals, a naive way is to enumerate all the time intervals of the time series and conduct Granger test for each interval. But such an enumeration could be very time consuming in practice. To address this challenge, in this paper we propose an efficient algorithm for causal time interval discovery. Here, our key insight is that similar time intervals (i.e., intervals with significant overlaps) are likely to have similar Granger test results. In fact, given the fitted models of a time interval, we can derive useful upper and lower bounds for the statistical significance of Granger causality for similar time intervals. Such bounds can be used to quickly identify intervals that will definitely pass or fail the test without fitting the actual models.

We have conducted experiments to verify the effectiveness and efficiency of our method on two interesting real-world datasets. First, we study the causality relationship between local event (represented by the frequency of geo-tagged tweets at the venue) and the traffic volume (represented by the taxi drop-offs/pick-ups). We have observed that, when there is an event, taxi drop-offs cause geo-tagged tweets, and geo-tagged tweets cause pick-ups. Second, we analyze variables of climate and their causal relationship with temperature. We have observed that the causality varies over time and is different for different variables.

In summary, our key contributions are as follows:

- To the best of our knowledge, we are the first to study the problem of detecting partial time intervals in which a causal relationship exists between two time series based on the Granger test.
- We propose an efficient algorithm which utilizes the bounds of regression error to avoid unnecessary computations for the Granger test over similar time intervals.
- We conduct experiments on both synthetic data and real data to demonstrate the effectiveness and efficiency of the proposed method.

Before proceeding, we emphasize that our work is built upon the Granger test. We do not claim to improve the Granger test itself, which is known to have some weaknesses (more details in Section 2). Instead, we focus on an important yet unaddressed problem in practice, that is, discovering partial time intervals with Granger causality.

2 Related Work

Dependency/causality discovery on time series. Many methods have been proposed to discover the temporal dependency between time series, such as autocorrelation, cross-correlation [7], transfer entropy [6], randomization test [16], phase slope index [21]. However, temporal correlation or dependency does not necessarily indicate causality. Granger causality test [10] was first introduced in the area of econometrics for time series analysis and has since gained tremendous success across many domains due to its simplicity and robustness [12, 13, 3, 8].

Weakness of the Granger test. A well-known difficulty in causality reasoning (including the Granger test) is the confounding issue, that is, spurious causality may be detected if both processes X and Y are impacted by a third (possibly unknown) process (i.e., a confounder). In [14], variables which mediate the impact of an action on the outcome are identified to reduce the bias in assessing the causal effect for online advertising. [5] exploits confounder path delays in an attempt to cancel out the spurious confounder effect. However, a principled solution to this problem remains elusive.

Variations of the Granger test. While the original Granger causality test was designed for two time series, several methods [2, 18, 17] have been proposed to analyze time series data involving many features and to learn a causal graph structure. Following the work [2], [20] detects causality of spatial time series, [18] proposes to use hidden Markov Random Field method, [17] handles extreme values in time series, [4] detects Granger causality from irregular time series, and [5] presents Copula-Granger method to efficiently capture non-linearity in the data. Learning temporal causal graph has been applied to biology applications [25], climate analysis [9], microbiology [19], fMRI data analysis [24], anomaly detection [23], and longitudinal analysis [26]. However, none of these work address the issues that causality could change over time, and that

causality only exists in partial time intervals. Previous study [1] uses a fixed-length sliding window to test Granger causality, but they cannot detect all the causal time intervals in an efficient way.

3 Finding Causal Time Intervals

We use $X = x_1x_2 \dots x_T$ and $Y = y_1y_2 \dots y_T$ to represent the two synchronized time series of interest. As discussed before, our goal is to detect all time intervals in which a casual relationship exists between X and Y. We use $X[i:j]$ to denote a time interval in time series X, that is, $x_ix_{i+1} \dots x_j$. Similarly, $Y[i:j]$ denotes $y_iy_{i+1} \dots y_j$. In this paper, we use the *two-sided Granger test* to determine the causality, which is introduced next.

3.1 Granger Causality Test The Granger causality test [10] is a well-known hypothesis testing procedure for determining whether one time series is useful in forecasting another. It was introduced to find "predictive causality" in time series. The basic idea of Granger causality is that, if a time series X causes the changes in a time series Y, the past values of X should provide *unique information* in predicting the future values of Y, compared with using the past values of Y only for prediction.

Specifically, given two time series X and Y, Granger test is performed by running the following two regressions with a given lag parameter L. The full model M_f uses both lagged values of X and lagged values of Y:

$$y_t \sim \sum_{l=1}^{L} a_l \cdot y_{t-l} + \sum_{l=1}^{L} b_l \cdot x_{t-l} \qquad \text{(Full model } M_f\text{)}.$$

The reduced model M_r uses only the lagged values of Y:

$$y_t \sim \sum_{l=1}^{L} a_l \cdot y_{t-l} \qquad \text{(Reduced model } M_r\text{)}.$$

If X Granger–causes Y, we expect the full model provides a significantly more accurate prediction than the reduced model. To determine whether the full model would be statistically better than the reduced model or not, the F-test is commonly performed, and it calculates the F-statistic:

$$F = \frac{(SSE_r - SSE_f)/(d_f - d_r)}{SSE_f/(T - d_f - 1)}, \tag{3.1}$$

where SSE_M denotes the sum of squared errors of a regression model M. Specifically, let $\hat{y}_t$ denotes the value of y_t predicted by M, then $SSE_M = \sum_{t=1}^{T}(y_t - \hat{y}_t)^2$. Further, d_f and d_r denote the degrees of freedom (i.e., the number of independent variables) in the models M_f and M_r, respectively, and T is the length of the time series.

In our problem, we have $d_f = 2L$ and $d_r = L$. Therefore, the F-statistic can be simplified as follows:

$$F = \frac{(SSE_r - SSE_f)/L}{SSE_f/(T - 2L - 1)}. \tag{3.2}$$

Under the null hypothesis (i.e., the full model does not give a significantly better prediction), the F-statistic will have an F-distribution with parameters L and $T - 2L - 1$, or equivalently, $F \sim \mathcal{F}_{L,T-2L-1}$. The null hypothesis is rejected if the F-statistic calculated from the data is greater than the critical value of the F-distribution for some desired false-rejection probability (e.g., p-value $= 0.05$).

Based on the above discussion, we can now summarize the Granger causality test as follows.

Definition 3.1. (Granger Causality Test) *Give two time series X and Y, and a time lag parameter L, we say that X Granger–causes Y if the F-statistic calculated in Eq.* (3.2) *is greater than the critical value of the F-distribution for certain desired false-rejection probability.*

3.2 Two-Sided Granger Causality In [10], Granger pointed out the following feedback mechanism for two-sided causality: if X Granger–causes Y and Y also Granger–causes X, it is likely that there exists another variable which needs to be controlled or could be a better candidate for the Granger causation. Several follow-up articles [11, 12] also mentioned the necessity of considering of the direction of Granger causality. Following the discussion, we test for two-sided Granger causality and say that X causes Y only if X causes Y by Granger causality but Y does *not* cause X by Granger causality. Thus, we define the causality as follows.

Definition 3.2. (Causality) *Give two time series X and Y, and a time lag parameter L, we say that X causes Y if and only if:*

1. *X and Y are stationary; and*
2. *X causes Y by Granger causality; and*
3. *Y does* not *cause X by Granger causality.*

We further define the time interval with causality as follows.

Definition 3.3. (Causal Time Interval) *Give two time series X and Y, and a time lag parameter L, we*

call time interval $[i:j]$ *a causal time interval if* $X[i:j]$ *causes* $Y[i:j]$ *by Definition 3.2.*

Our objective is to find all the casual time intervals.

PROBLEM 3.1. *Give two time series* X *and* Y*, find all the time intervals* $[i:j]$ *such that* $X[i:j]$ *causes* $Y[i:j]$ *by Definition 3.2.*

4 Efficient Algorithm

The most straightforward approach to solve Problem 3.1 is to enumerate all the time intervals $[i:j]$ and check each criterion in Definition 3.2. Checking the causality of one time interval takes $O(T \cdot L^2)$ time (see Appendix A for details). Since there are $O(T^2)$ time segments, the overall time complexity is $O(T^3 \cdot L^2)$. Such time complexity could be too high for some real applications, especially when the time series is long. Next, we investigate efficient pruning rules to speed-up the computation.

Key insight to speed-up the computation. We observe that the major bottleneck of complexity lies in the fitting of the full model and the reduced model for each interval in order to compute the SSE values for F-test. To speed-up the computation, our key insight is that very often it is unnecessary to obtain the *exact* SSE values and F-statistic for an interval. Specially, suppose that for interval $[i:j]$ we have fitted the full model $M_f^{[i:j]}$ and the reduced model $M_r^{[i:j]}$. Then, consider the subsequent interval $[i:j+1]$. Since the models for $[i:j]$ and the models for $[i:j+1]$ share most of the data samples for fitting, we can actually estimate the *bounds* of the F-statistic for the interval $[i:j+1]$ using the models for $[i:j]$ and then do pruning based on the bounds. For example, if the upper bound of F-statistic for $[i:j+1]$ is smaller than or equal to the critical value of the corresponding F-distribution, then we can conclude that $[i:j+1]$ will fail the Granger test and there is no need to fit the actual models. Further, we can continue the same estimation of bounds for intervals $[i:j+2]$, $[i:j+3]$, ... (with less tight bounds), until we reach a time interval $[i:j+\delta]$ that requires an accurate computation of F-statistic. Below we formally describe the pruning technique based on our key insight, and prove its effectiveness.

4.1 Pruning Technique

To avoid unnecessary computation for time intervals that do not have causality, we first introduce the bounds for error in the regression model. Given a time interval $[i:j]$ and the corresponding data set $\{y_k, x_{k1}, \ldots, x_{kp}\}_{k=i}^{j}$, we use $M^{[i:j]}$ to denote the regression model obtained using the ordinary least squares fitting:

$$y_k = \beta_1 x_{k1} + \cdots + \beta_p x_{kp} + \varepsilon_k = \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_k, \quad k = i, \ldots, j,$$

where $\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\beta}$ is the inner product of vectors $\mathbf{x}_k$ and $\boldsymbol{\beta}$. The sum of squared errors for model $M^{[i:j]}$ is defined as:

$$SSE(M^{[i:j]}) = \sum_{k=i}^{j} \epsilon_k^2. \tag{4.3}$$

For a regression model M, we further denote $\epsilon_M(\mathbf{x}_k)$ as the prediction error of sample $\mathbf{x}_k$ using model M:

$$\epsilon_M(\mathbf{x}_k) = M(\mathbf{x}_k) - y_k, \tag{4.4}$$

where $M(\mathbf{x}_k)$ is the predicted value of y_k based on predictors $\mathbf{x}_k$ using model M. Then, for any $\delta \in \mathbb{Z}^+$, we have the following two lemmas regarding the error bounds of $M^{[i:j+\delta]}$:

LEMMA 4.1. *Given a regression model* $M^{[i:j]}$*, the upper bound of* $SSE(M^{[i:j+\delta]})$ *is:*

$$\lceil SSE(M^{[i:j+\delta]}) \rceil = SSE(M^{[i:j]}) + \sum_{k=j+1}^{j+\delta} \epsilon^2_{M^{[i:j]}}(\mathbf{x}_k). \tag{4.5}$$

Proof. By definition, we always have $SSE(M^{[i:j+\delta]}) = \min_{\boldsymbol{\beta}} \sum_{k=i}^{j+\delta} \epsilon_M^2(\mathbf{x}_k)$. It is obvious that $\min_{\boldsymbol{\beta}} \sum_{k=i}^{j+\delta} \epsilon_M^2(\mathbf{x}_k) \leq \sum_{k=i}^{j+\delta} \epsilon^2_{M^{[i:j]}}(\mathbf{x}_k)$. Now we also note that $\sum_{k=i}^{j+\delta} \epsilon^2_{M^{[i:j]}}(\mathbf{x}_k) = SSE(M^{[i:j]}) + \sum_{k=j+1}^{j+\delta} \epsilon^2_{M^{[i:j]}}(\mathbf{x}_k)$. Therefore, (4.5) gives an upper bound to $SSE(M^{[i:j+\delta]})$.

LEMMA 4.2. *Given a regression model* $M^{[i:j]}$*, the lower bound of* $SSE(M^{[i:j+\delta]})$ *is:*

$$\lfloor SSE(M^{[i:j+\delta]}) \rfloor = SSE(M^{[i:j]}). \tag{4.6}$$

Proof. Recall that $SSE(M^{[i:j]}) = \min_{\boldsymbol{\beta}} \sum_{k=i}^{j} \epsilon_M^2(\mathbf{x}_k)$. Note that $\min_{\boldsymbol{\beta}} \sum_{k=i}^{j} \epsilon_M^2(\mathbf{x}_k) \leq \sum_{k=i}^{j} \epsilon^2_{M^{[i:j+\delta]}}(\mathbf{x}_k)$ by definition. Then, $\sum_{k=i}^{j} \epsilon^2_{M^{[i:j+\delta]}}(\mathbf{x}_k) \leq \sum_{k=i}^{j+\delta} \epsilon^2_{M^{[i:j+\delta]}}(\mathbf{x}_k) = SSE(M^{[i:j+\delta]})$, since $\epsilon^2_{M^{[i:j+1]}}(\mathbf{x}_k), \cdots, \epsilon^2_{M^{[i:j+\delta]}}(\mathbf{x}_k)$ are all non-negative. Therefore, (4.6) gives a lower bound to $SSE(M^{[i:j+\delta]})$.

Now, let $M_f^{[i:j]}$ and $M_r^{[i:j]}$ denote the full model and reduced model obtained via least square fitting for time

segment $[i:j]$, respectively. We wish to know if time interval $[i:j+\delta]$ can pass the Granger test. Based on the F-test (Eq. (3.2)), the upper bound of F-statistic for $[i:j+\delta]$ is:

(4.7)
$$\lceil F^{[i:j+\delta]} \rceil = \frac{\left(\lceil SSE(M_r^{[i:j+\delta]}) \rceil - \lfloor SSE(M_f^{[i:j+\delta]}) \rfloor \right) / L}{\lfloor SSE(M_f^{[i:j+\delta]}) \rfloor / \left((j+\delta-i+1) - 2L - 1 \right)}.$$

For a fixed false-rejection probability (e.g., $p = 0.05$), let $c^*_{a,b}$ be the critical value of the F-distribution $\mathcal{F}_{a,b}$ with (a,b) degrees of freedom. Obviously, if the upper bound $\lceil F^{[i:j+\delta]} \rceil \leq c^*_{L,(j+\delta-i+1)-2L-1}$, there is no need to fit the regression model for time interval $[i:j+\delta]$. And we can safely move on to the next time interval $[i, j+\delta+1]$.

Similarly, we can also calculate the lower bound of F-statistic for $[i:j+\delta]$:

(4.8)
$$\lfloor F^{[i:j+\delta]} \rfloor = \frac{\left(\lfloor SSE(M_r^{[i:j+\delta]}) \rfloor - \lceil SSE(M_f^{[i:j+\delta]}) \rceil \right) / L}{\lceil SSE(M_f^{[i:j+\delta]}) \rceil / \left((j+\delta-i+1) - 2L - 1 \right)}.$$

Obviously, if the lower bound $\lfloor F^{[i:j+\delta]} \rfloor > c^*_{L,(j+\delta-i+1)-2L-1}$, the time interval $[i:j+\delta]$ will definitely pass the Granger causality test. In such cases, there is again no need to fit the actual models and calculate the actual F-statistic.

Note that the same pruning technique can be applied to the reverse Granger test. Specifically, if the upper bound of F-statistic is small or equal to the critical value, or the lower bound of F-statistic is greater than the critical value, then there is no need to fit the actual models or calculate the actual F-statistic. In other words, we only conduct the actual reverse Granger test if the upper bound is greater than the critical value but the lower bound is small or equal to the critical value. Finally, in the reverse direction, in which we test if Y causes X, we only keep those intervals that *fail* the F-test.

Finally, due to the space limit, we refer interested readers to Appendix B for a summarization of our algorithm and discussions on the implementation details.

5 Experiments

We conduct experiments on both synthetic datasets and real datasets to examine the efficiency and effectiveness of our proposed method. All the experiments are conducted on a 3.4GHz Intel Core i7 system with 16 GB memory.

5.1 Experiment on Synthetic Data To generate synthetic dataset, we first generate independent variable X and then generate time series Y based on X. Since both X and Y need to be stationary, we simulate X as an auto-regressive stationary time series of length T.

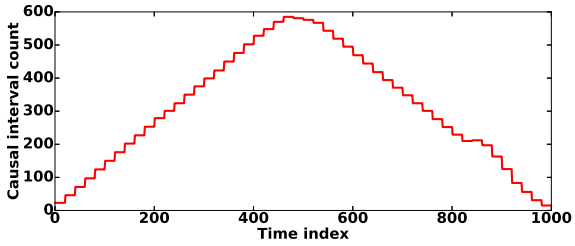


Figure 2: Count of causal time intervals with ground truth causal interval as $[450:550]$. We expect to see a Bell-shaped curve when overlaying all the causal time intervals.

To generate time series Y, we first choose an interval $[s^* : t^*]$ to be the causal time interval. Then, we simulate Y as

$$y_t = \begin{cases} \sum_{k=1}^{L} a_k y_{t-k} + \sum_{k=1}^{L} b_k x_{t-k} + \epsilon_t, & s^* \leq t \leq t^* \\ \sum_{k=1}^{L} a_k y_{t-k} + \epsilon_t, & \text{otherwise} \end{cases}$$

where a_k and b_k are parameters controlling the degree of auto-correlation and Granger causality respectively. The noise ϵ_t follows a Gaussian distribution and by default we set the standard deviation of this Gaussian distribution as 0 (i.e., no error).

By default, we set $T = 1000$, $[s^* : t^*] = [450:550]$, and $L = 2$. For the coefficients, we set $a_k = 0.2$ if k is odd, and $a_k = -0.2$ if k is even. We alternate the sign of a_k here to avoid the situations that the time series keeps increasing (or decreasing) and becomes extremely large (or small). In addition, we set $b_k = 0.2$.

Effectiveness Study. First, we note that although the ground truth time interval is $[450:550]$, there will be many intervals passing the causality test. Specifically, if an interval is causal, other intervals contained by this interval, as well as those which have significant overlaps with it, will also be causal. Thus, if we overlay all these intervals, we expect to see a Bell-shaped curve as shown in Fig. 2. As one can see, the middle time index (i.e., $t = 500$) is covered by the largest number of causal time intervals (i.e., about 600 in our setting). Note that timestamps that are not in $[450:550]$ may also be covered by a few causal intervals.

Now, suppose that the output contains p causal intervals $[s_1 : t_1]$, $[s_2 : t_2]$, ..., $[s_p : t_p]$. We denote c_t as the number of causal intervals covering timestamp t. Since we conduct causality test on the simulated data without any noise by default, we treat the count c_t as the ground truth count of causal intervals for timestamp

t. We then conduct the causality test on the noisy data and denote the count of detected causal intervals at t as c'_t. We define the detection error as the mean difference between c'_t and the ground truth c_t:

$$\frac{1}{T}\sum_{t=1}^{T}|c_t - c'_t|. \tag{5.9}$$

In Figure 3, we show the detection error as a function of the noise variance. As one can see, the detection error increases as the variance of noise increases. But the difference in the count of causal intervals remains relatively small when the variance is below 0.1.

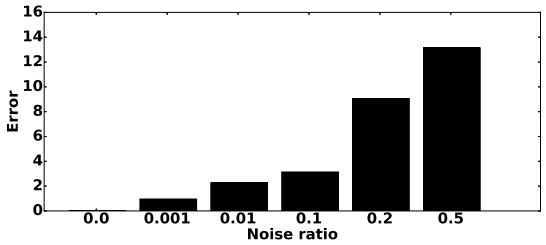


Figure 3: Detection error w.r.t. the variance of noise.

Efficiency Study. We conduct several sets of experiments to demonstrate the efficiency of our method under different settings.

First, in Table 1 we show the running time of the baseline method (without pruning) and our efficient algorithm (with pruning). In this experiment, we keep the causal interval length to be 100, and the standard deviation of noise to be 0. One can see that, as the length of the time series increases, it takes longer time to detect causal time intervals. However, our algorithm with pruning is at least 48% faster than the baseline method which does not do pruning. Note that, in the baseline method, we first test whether X Granger causes Y. Only when it passes the Granger test, we will run the reverse Granger test. And only if it fails the reverse test, we conduct stationarity test. So the baseline does not necessarily conduct all the model fittings and tests. But still, our pruning rule can save a significant amount of time on top of the baseline method.

Total length	Baseline	Pruning	Improvement
500	65.72	33.97	48%
1000	123.13	60.69	51%
1500	196.84	101.84	48%
2000	258.91	127.57	51%
2500	316.97	150.95	52%

Table 1: Running time (in seconds) and relative improvement w.r.t. total length of time series.

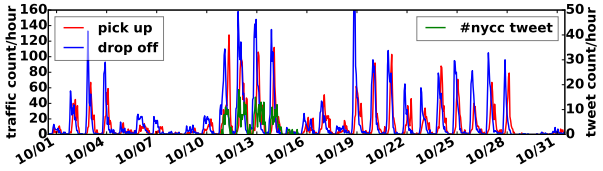
Next, we fix the total length of time series to be 1000, and examine the running time w.r.t the length of the causal time interval. As one can see in Table 2, the longer the causal interval is, the longer the overall running time is. However, our pruning method is robust to the change of causal interval length. We can always achieve at least 51% improvement in efficiency.

Causal length	Baseline	Pruning	Improvement
0	115.42	53.96	53%
50	114.45	53.52	53%
100	123.13	60.69	51%
150	132.46	61.90	53%
200	136.89	62.16	54%

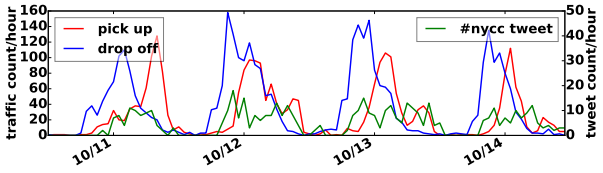
Table 2: Running time (in seconds) and relative improvement w.r.t. length of the causal interval.

5.2 Experiment on Traffic data In this section, we use real data to analyze the causal relationship between taxi pick-ups/drop-offs and geo-tagged tweets in New York City.

Data Description. The New York City generated about half million taxi trips every day. Such taxi trip data are public from `www.nyc.gov`. The dataset contains pick-up and drop-off of each trip. Each pick-up or drop-off record contains information of timestamp, latitude, and longitude. In order to study what causes taxi trips, we use geo-tagged tweets as a signal of local events. Each geo-tagged tweet includes information of timestamp, userid, latitude, longitude, and tweet content.



(a) Traffic and tweet data



(b) Traffic and tweet data (zoomed in)

Figure 4: Traffic and tweet data near Jacob K. Javits Convention Center in New York City in October, 2012. The zoomed in time series shows strong correlation between traffic and #nycc tweets.

Causality. Traffic and geo-tagged tweets could have a significant causal relationship when there is a big local event where many people tweet about the event and also many people take taxis before and after the event. We show such a case study in our dataset. We choose a 500 meter by 500 meter area near Jacob K. Javits Convention Center in New York City. In Figure 4, we show the frequency of taxi pick-ups and drop-offs from 10/01/2012 to 10/31/2012 in that area. We are interested in explaining the changes in traffic. We test the temporal frequency of different hashtags and find one hashtag "#nycc" has a strong causality with taxi volume from 10/11/2012 to 10/14/2012 in that area. Hashtag "#nycc" means New York Comic Conference. Figure 4 shows the frequency of this hashtag over time.

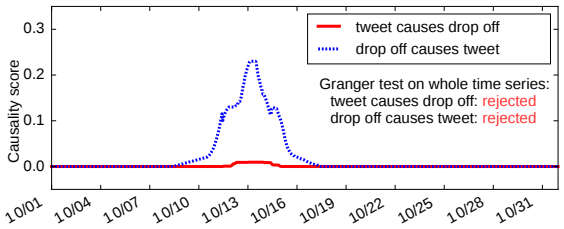
By testing the causality between the whole time series of traffic and tweet, we are unable to find any causality, as shown in Figure 5 where the causality score for the whole time interval is 0. Thus, we detect all the causal time intervals between traffic volume (pick-up frequency and drop-off frequency) around Jacob K. Javits Convention Center and the frequency of tweets with hashtag "#nycc" in October 2012. The time lag L is set to 5 hours. We only test time intervals less than 72 hours. To examine the strength of causal relationship at any given timestamp t, we further define the **causality score** of $CS(t)$ as:

$$CS(t) = \frac{\#\ causal\ time\ intervals\ covering\ timestamp\ t}{\#\ time\ intervals\ covering\ timestamp\ t}.$$

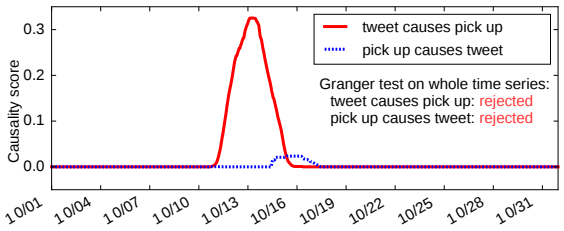
Figure 5(a) shows that from 10/11/2012 to 10/14/2012, there is a strong evidence that changes in taxi drop-off frequency cause changes in the frequency of tweets with hashtag "#nycc". This can be explained by the fact that many people took the taxi to Jacob K. Javits Convention Center, and posted tweets with hashtag "#nycc" during their stay in the comic event. Interestingly, it is much less evident that changes in the frequency of tweets cause changes in taxi drop-off frequency, because drop-offs typically happened before people tweeting about the event.

In addition, Figure 5(b) shows that from 10/11 to 10/14, there is a strong evidence that changes in the frequency of tweets with hashtag "#nycc" cause changes in the taxi pick-up frequency. This can be explained by the fact that, after attending the event, people took a taxi and left the venue. However, we do not see much causality in the reverse direction.

Running Time. In Table 3, we show the running time for different lengths of time series. Our pruning technique saves at least 36% of the time compared to the baseline method.



(a) Causality between tweet frequency and taxi drop-offs (no causality when testing on the whole time series of tweet and drop-off).



(b) Causality between tweet frequency and taxi pick-ups (no causality when testing on the whole time series of tweet and pick-up)

Figure 5: Causality between traffic and frequency of geo-tagged tweet containing hashtag "#nycc". During the nycc event period (10/11-10/14), we can see that drop-offs cause tweets but not vice versa; and tweets cause pick-ups but not vice versa. This suggests that people take taxi to the venue, tweet about the comic event, and then take taxi to leave.

5.3 Experiment on Climate Data

In recent years, global warming has been a major concern around the world. In this section, we use a real climate dataset [20] to study causal relationships between temperature and other meteorological factors.

Data description. The climate dataset [20] is composed of 125 climate monitoring locations in North America. These locations are located on 2.5 × 2.5 degree grids. Each sample location reported the monthly data from 1990 to 2002, with each sample containing temperature (TMP), CH4, CO2, CO, H2, Precipitation (PRE), Vapor (VAP), Cloud Cover (CLD), Wet Days (WET), and Frost Days (FRS). In this experiment, we examine the causal relationship between the temperature and the other nine factors for 125 locations separately and then aggregate the results of causality scores across all these locations.

Causality. We show the causality score obtained by our algorithm for each factor in Fig 6, with variance across locations represented by the grey area. The time lag length is set to be 3 months. In general, the causality

#Months	Baseline	Pruning	Improvement
1	53.46	30.64	43%
2	117.41	66.97	43%
3	179.82	101.98	43%
6	333.54	208.86	37%
12	711.72	458.61	36%

Table 3: Running time (in seconds) comparison on taxi data w.r.t. time series lengths.

of each feature on temperature varies over time. The causality of climate features (PRE, VAP, WET, CLD) varies a lot across different locations. However, the causality of gases (CH4, CO2, CO, H2) on temperature is relatively consistent across different locations. The well-known Greenhouses gases CH4 and CO2 contribute differently to the temperature change. The CH4 tends to have a stronger causality on temperature in recent years, but the causality of CO2 on temperature is becoming less significant.

Running time. Without pruning, it takes $11,680$ seconds to run over all stations of all variables w.r.t. temperature. With the pruning technique, it costs $7,162$ seconds. The pruning technique saves about one third of the computational time.

6 Conclusion

In this paper, we propose to study a new problem of detecting partial time intervals with Granger causality. This problem is motivated by our observation that causality could change over time. We further develop an efficient algorithm that utilizes the error bounds in the regression models to avoid repeatedly fitting regression models for similar time intervals. We have demonstrated the effectiveness and efficiency of our method on both synthetic and real datasets.

Acknowledgements

The work was funded from a gift to Penn State for the Pennsylvania State University General Electric Fund for the Center for Collaborative Research on Intelligent Natural Gas Supply Systems and was supported in part by NSF awards #1639150, #1618448, and #1544455. Lingzhou Xue's research is supported by the American Mathematical Society Simons Travel Grant and NSF award #1505256. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

[1] J. Aaltonen and R. Östermark. A rolling test of granger causality between the finnish and japanese security markets. *Omega*, 25(6):635–642, 1997.

[2] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical granger methods. In *KDD*, pages 66–75, 2007.

[3] I. Asimakopoulos, D. Ayling, and W. M. Mahmood. Non-linear granger causality in the currency futures returns. *Economics Letters*, 68(1):25–30, 2000.

[4] M. T. Bahadori and Y. Liu. Granger causality analysis in irregular time series. In *SDM*, pages 660–671, 2012.

[5] M. T. Bahadori and Y. Liu. An examination of practical granger causality inference. In *SDM*, 2013.

[6] L. Barnett, A. B. Barrett, and A. K. Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701, 2009.

[7] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[8] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler. Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by granger causality. *Proceedings of the National Academy of Sciences*, 101(26):9849–9854, 2004.

[9] J. B. Elsner. Granger causality and atlantic hurricanes. *Tellus A*, 59(4):476–485, 2007.

[10] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[11] C. W. Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.

[12] C. W. Granger. Some recent development in a concept of causality. *Journal of econometrics*, 39(1):199–211, 1988.

[13] C. Hiemstra and J. D. Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994.

[14] D. N. Hill, R. Moakler, A. E. Hubbard, V. Tsemekhman, F. Provost, and K. Tsemekhman. Measuring causal impact of online actions via natural experiments: Application to display advertising. In *KDD*, pages 1839–1847, 2015.

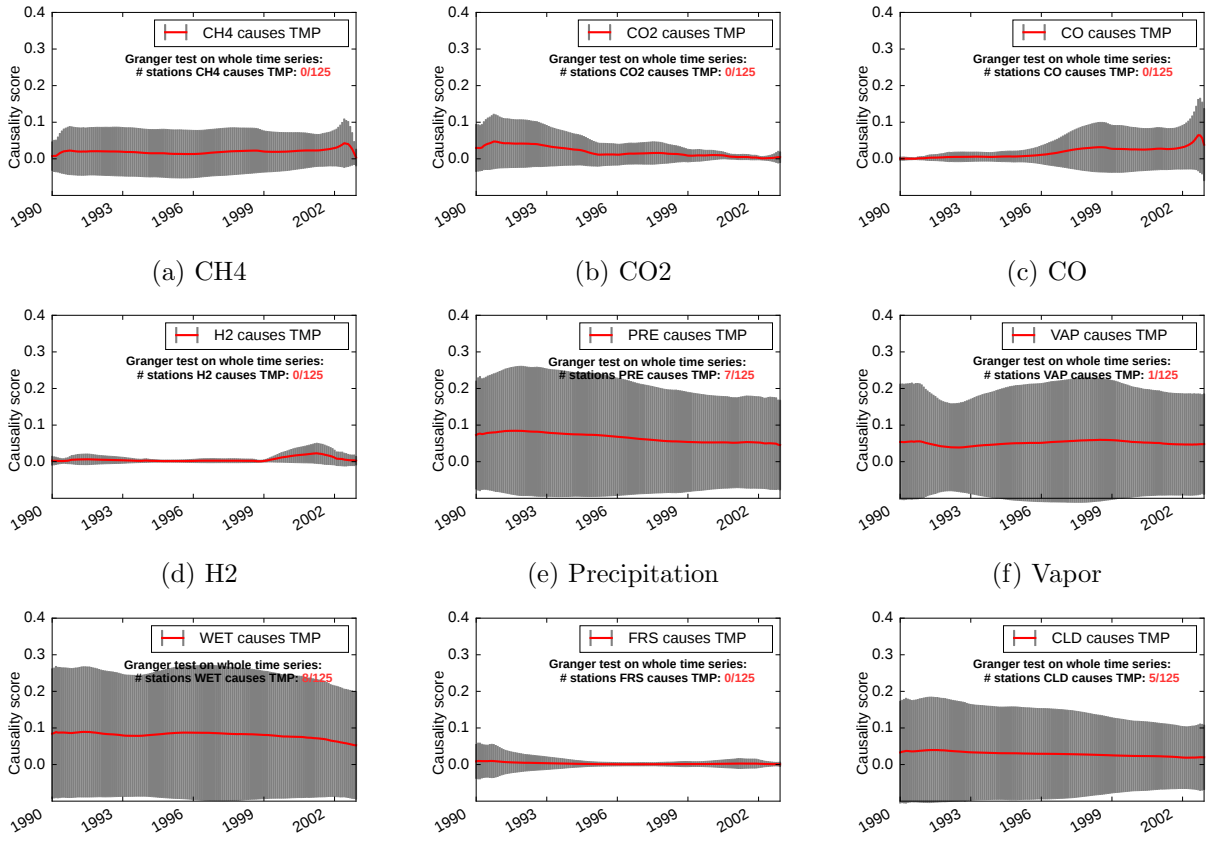
[15] R. W. Howarth, A. Ingraffea, and T. Engelder. Natural gas: Should fracking stop? *Nature*, 477(7364):271–275, 2011.

[16] T. La Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *WWW*, 2010.

[17] Y. Liu, T. Bahadori, and H. Li. Sparse-gev: sparse latent space model for multivariate extreme value time serie modeling. In *ICML*, 2012.

[18] Y. Liu, A. Niculescu-Mizil, A. C. Lozano, and Y. Lu. Learning temporal causal graphs for relational time-series analysis. In *ICML*, pages 687–694, 2010.

[19] A. C. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical granger modeling for gene expression regula-



(a) CH4 (b) CO2 (c) CO

(d) H2 (e) Precipitation (f) Vapor

(g) Wet Days (h) Frost Days (i) Cloud Cover

Figure 6: Causal relationship between temperature and other climate variables. There are 125 climate monitoring stations and we conduct causality test for each station. The red line shows the average causality score over 125 stations and the grey line indicates the variance. The larger the grey area is, the bigger the difference among different locations is. In each graph, we also report the number of stations that show causality when testing on the whole time series.

tory networks discovery. *Bioinformatics*, 25(12):i110–i118, 2009.

[20] A. C. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe. Spatial-temporal causal modeling for climate change attribution. In *KDD*, pages 587–596, 2009.

[21] G. Nolte, A. Ziehe, V. V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K.-R. Müller. Robustly estimating the flow direction of information in complex physical systems. *Physical review letters*, 100(23):234101, 2008.

[22] S. M. Olmstead, L. A. Muehlenbachs, J.-S. Shih, Z. Chu, and A. J. Krupnick. Shale gas development impacts on surface water quality in pennsylvania. *Proceedings of the National Academy of Sciences*, 110(13):4962–4967, 2013.

[23] H. Qiu, Y. Liu, N. A. Subrahmanya, and W. Li. Granger causality for time-series anomaly detection. In *ICDM*, pages 1074–1079, 2012.

[24] S. Ryali, K. Supekar, T. Chen, and V. Menon. Multivariate dynamical systems models for estimating causal interactions in fmri. *Neuroimage*, 54(2):807–823, 2011.

[25] P. A. Valdés-Sosa, J. M. Sánchez-Bornot, A. Lage-Castellanos, M. Vega-Hernández, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez. Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1457):969–981, 2005.

[26] T. Xu, J. Sun, and J. Bi. Longitudinal lasso: Jointly learning features and temporal contingency for outcome prediction. In *KDD*, pages 1345–1354, 2015.

[27] Y. Zheng, F. Liu, and H.-P. Hsieh. U-air: When urban air quality inference meets big data. In *KDD*, pages 1436–1444, 2013.

[28] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li. Forecasting fine-grained air quality based on big data. In *KDD*, 2015.

A Time Complexity of the Causality Test

According to Definition 3.2, we next examine the time complexity of each step in the procedure.

1. **Stationarity Test**. Typically, Augmented Dickey-Fuller (ADF) test [2] is applied to check the stationarity of a single time series. The test requires the fitting of an auto-regression model. The ordinary least squares fitting only involves matrix multiplication and inversion, and its time complexity is $O(np^2)$, where n is the sample size and p is the number of predictors. Hence, the time complexity of ADF test is $O(T \cdot q^2)$, where q is the lag parameter in the auto-regression model.

2. **Granger test on X causes Y**. This step requires fitting the full model $M_f^{[i:j]}$ and reduced model $M_r^{[i:j]}$ in order to obtain the sum of square errors (SSE) for the F-test shown in Eq. (3.2). Fitting the full model takes $O(T \cdot (2L)^2)$ time, and fitting the reduced model takes $O(T \cdot L^2)$ time. So the time complexity of this step is $O(T \cdot L^2)$.

3. **Reverse Granger test on Y causes X**. In this step, we need to fit the full model and the reduced model in the reverse direction, denoted by $H_f^{[i:j]}$ and $H_r^{[i:j]}$, respectively. As in Step 2, the time complexity of this step is $O(T \cdot L^2)$.

Based on the analysis above, conducting one causality test takes $O(T \cdot L^2)$ time.

B Algorithm

In Algorithm 1, we describe the pseudo code of our algorithm. Given two time series X and Y, we enumerate starting timestamp i and ending timestamp j (Line 3-5). For each interval, we first compute the upper bound of F-statistic using the latest regression model we have built $M^{[i:k_M]}$ (Line 7). If the upper bound of F-statistic is smaller than or equal to the critical value of the corresponding F-distribution, interval $[i:j]$ will definitely fail the Granger test and we continue to the next interval $[i:j+1]$ (Line 8-9). Otherwise, we further compute the lower bound of F-statistic (Line 10). If it is greater than the critical value, there is no need to fit the actual regression models because the interval $[i,j]$ will definitely pass the Granger test. If the lower bound is smaller than or equal to the critical value (Line 11), we need to fit the actual models (Line 12) and update the ending timestamp of the latest model (Line 13). Given the fitted models, we compute the actual F-statistic (Line 14). If time interval $[i:j]$ fails the Granger test (Line 15), we continue to the next interval; Otherwise, we move on to the reverse Granger test.

```
1:  INPUT: X = x_1 x_2 ... x_T and Y = y_1 y_2 ... y_T
2:  PARAMETER: lag L
3:  for i = 1 to T do
4:      k_M ← i, k_H ← i
5:      for j = i to T do
6:          {Granger test}
7:          Compute ⌈F_M^{[i:j]}⌉ using M^{[i:k_M]} according to Eq. (4.7)
8:          if ⌈F_M^{(i,j)}⌉ ≤ c*_{L,j-i+1-2L-1} then
9:              Continue
10:         Compute ⌊F_M^{[i:j]}⌋ using M^{[i:k_M]} according to Eq. (4.8)
11:         if ⌊F_M^{(i,j)}⌋ > c*_{L,j-i+1-2L-1} then
12:             Fit full model M_f^{[i:j]} and reduced model M_r^{[i:j]}
13:             k_M ← j
14:             F_M^{[i:j]} ← ((SSE(M_r^{[i:j]}) - SSE(M_f^{[i:j]}))/L) / (SSE(M_f^{[i:j]})/(j-i+1-2L-1))
15:             if F_M^{[i:j]} ≤ c*_{L,j-i+1-2L-1} then
16:                 Continue
17:         {Reverse Granger test}
18:         Compute ⌊F_H^{[i:j]}⌋ using H^{[i:k_H]} according to Eq. (4.8)
19:         if ⌊F_H^{[i:j]}⌋ > c*_{L,j-i+1-2L-1} then
20:             Continue
21:         Compute ⌈F_H^{[i:j]}⌉ using H^{[i:k_H]} according to Eq. (4.7)
22:         if ⌈F_H^{[i:j]}⌉ > c*_{L,j-i+1-2L-1} then
23:             Fit full model H_f^{[i:j]} and reduced model H_r^{[i:j]}
24:             k_H ← j
25:             F_H^{[i:j]} ← ((SSE(H_r^{[i:j]}) - SSE(H_f^{[i:j]}))/L) / (SSE(H_f^{[i:j]})/(j-i+1-2L-1))
26:             if F_H^{[i:j]} > c*_{L,j-i+1-2L-1} then
27:                 Continue
28:         {Stationarity Test}
29:         if X[i : j] is stationary AND Y[i : j] is stationary then
30:             Output [i : j] as a causal time interval
```

Algorithm 1: Finding Causal Time Intervals

For the reverse Granger test, we also reverse the pruning criteria as we wish to only output intervals that fail the test. This time, we first compute the lower bound of F-statistic using the latest regression model we have built $H^{[i:k_H]}$ (Line 18). If the lower bound is greater than the critical value of the corresponding F-distribution, interval $[i:j]$ will definitely pass the test and therefore should not be output (Line 19-20). Otherwise, we further compute the upper bound (Line 21). If the upper bound is smaller than or equal to

the critical value, $[i : j]$ will definitely fail the test; if the upper bound is greater than the critical value (Line 22), we need to fit the actual models (Line 23) and update the ending timestamp of the latest model (Line 24). Given the fitted models, we compute the actual F-statistic (Line 25). If time interval $[i : j]$ passes the reverse Granger test (Line 26), we continue to the next interval; Otherwise, we move on to the Stationarity test.

Finally, we test the stationarity of both time series (Line 29-30). Since stationarity test also involves fitting regression models, We only do the stationarity test for intervals that are not filtered by the Granger tests.

B.1 Implementation Details In this section, we provide further technical details in our implementation.

Speed up the fitting of regression models. As we move from time interval $[i : j]$ to $[i : j + \delta]$, it is not necessary to fit the regression model for $[i : j + \delta]$ from scratch. Instead, we can incrementally update the model built for time interval $[i : j]$ with the new data samples as described in [1].

Adjustment for best efficiency. In Algorithm 1, we present Granger test first and then the reverse Granger test for clarity of presentation. Actually, it is faster to calculate and check the bounds for the reverse Granger test (Line 18-22) before fitting the models for the Granger test (Line 12), because bound checking only takes $O(1)$ time. If the interval $[i : j]$ passes the reverse Granger test (which violates the third condition in Definition (3.2)), there is no need to fit the models in Line 12.

Lag selection. Granger causality test can be sensitive to the choice of the lag parameter L [5]. One way to choose an appropriate lag is to measure the quality of the auto-regression model [4] for different lags and choose the lag that produces the best model. Akaike information criterion (AIC) or Bayesian information criterion (BIC) can be used to measure the quality of a model.

Aggregating and visualizing the results. We note that if time interval $[i : j]$ is causal, it is very likely that intervals with significant overlaps with $[i : j]$ are also causal (e.g., $[i : j + 1]$). As a result, the output of our algorithm typically consists of many overlapping time intervals. An effective way to visualize the result is to overlay all these time intervals and generate a time series indicating the strength of causality at each timestamp. In particular, if there is a real causal relationship at interval $[s : t]$, we will observe a Bell shape time series with the peak in the middle of $[s : t]$ (see Figure 2 for an example). Alternatively, one can modify our problem definition to only output the maximal time intervals or closed time intervals (similar to the definition of maximal/closed frequent patterns [3]). As in the case of frequent patterns, such a new definition may leads to more efficient algorithms, but a through discussion on this topic is outside the scope of this paper.

References

[1] J. M. Chambers. Regression updating. *Journal of the American Statistical Association*, 66(336):744–748, 1971.

[2] W. A. Fuller. *Introduction to statistical time series*, volume 428. John Wiley & Sons, 2009.

[3] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques (3rd ed.)*. Morgan Kaufmann, 2011.

[4] V. Ivanov and L. Kilian. A practitioner's guide to lag order selection for var impulse response analysis. *Studies in Nonlinear Dynamics & Econometrics*, 9(1), 2005.

[5] D. L. Thornton and D. S. Batten. Lag-length selection and tests of granger causality between money and income. *Journal of Money, credit and Banking*, 17(2):164–178, 1985.