

# Extracting Precursor Rules from Time Series - A Classical Statistical Viewpoint\*

*João B. D. Cabrerá and Raman K. Mehra* <sup>†</sup>

**Abstract** In many applications of interest one is faced with the problem of identifying precursor events for extraordinary phenomena. We investigate this problem within the framework of Temporal Data Mining. The concept of Precursor Rule is defined in terms of events and sequences of events. Precursor Rules relate Precursor Events extracted from input time series with Phenomenon events extracted from output time series. A methodology is proposed for extracting Precursor Rules from databases containing time series related to different regimes of a system. Given a fixed output time series containing one or more Phenomenon events, a key contribution of this paper is to show that the Granger Causality Test (GCT) can be used for ranking candidate time series according to the likelihood that Precursor Rules exist. Time Series Quantization is performed for extracting Phenomenon events and Precursor events, but GCT is applied to the raw time series, before Quantization and the definition of Event Types. The paper presents an analytic investigation of the utilization of the GCT for time series pairs containing impulsive time-localized structure. Following a number of approximations, the Granger Causality Index is related with the confidence of the Precursor Rules extracted from these time series pairs. An example from Network Security illustrates the effectiveness of the methodology. Using MIB (Management Information Base) datasets collected from real experiments involving Distributed Denial of Service Attacks, it is shown that Precursor Rules relating activities at Attacking Machines with Traffic Floods at Target Machines can be extracted by the method.

## 1 Introduction

In many applications of interest one is faced with the problem of identifying precursor events for extraordinary phenomena. Well known examples are earthquakes (eg. [20]) and financial market crashes (eg. [23]), where the benefits of determining reliable precursors cannot be overemphasized. Network Security is another field in which the determination of reliable precursors is of great importance. One would be interested in identifying abnormal activities in certain nodes of the network that precede a computer attack, such as a Denial-of-Service traffic flood (eg. [8], [9]). If precursor events could be identified in any of these three situations, and appropriate alarming mechanisms are in place, one is given the possibility of preventing, or at least minimizing the deleterious effects of the phenomenon. The nature of the alarming mechanisms and preventive actions in each case will depend on the confidence of the precursor event as an indicator of the occurrence of the disrupting phenomena in the near future, and the estimated time lag between the precursor event and the phenomenon. The specific nature of the phenomenon - domain information - also dictates the response in each case. As an example, precursors for computer attacks determined on a training set could

---

\*This work was supported by the Air Force Research Laboratory (Rome, NY - USA) under contract F30602-01-C-057 to Scientific Systems Company.

<sup>†</sup>The authors are with Scientific Systems Company, 500 West Cummings Park, Suite 3000 Woburn MA 01801 USA email: cabrerá@ssci.com, rkm@ssci.com .

be used to trigger the disconnection of offending nodes, or to trigger mechanisms to protect potential targets against disruptive traffic. Such rules could also be used on a forensics mode, to support human personnel in the detection of the source and nature of the security violation (eg. [8]). In these examples, the data records are numerical sequences (or time series), and the objective is to extract rules relating certain events (the precursors) in a certain variable, with events (the phenomenon) on another variable. Given an ensemble of time series, three interrelated questions can be posed: (1) Determining the variables where the phenomenon is most evident - the outputs, and characterize the phenomenon as an event in these variables (a sudden drop, an oscillation, etc.); (2) Given the outputs, determine the variables where the precursors can be found - inputs; and (3) Given the inputs, determine the precursors, again as events. Problems of this nature have been addressed implicitly or explicitly by the Statistics and Control Theory research communities for many decades now, under the headings of Time Series Analysis (eg. [15]) and System Identification (eg. [18]). The main emphasis in these disciplines has been in first extracting global models describing the evolution of the ensemble of the time series, and then to infer relationships among the variables using these models. The models themselves are very important for these disciplines, as they can be used for designing controllers or compensators in control engineering applications (eg. [17]), or for forecasting in many fields, such as Economics and Geophysics (eg. [7]). For lack of a better name, we call these efforts and their attending techniques as the Classical Statistical approach.

The Data Mining approach on the other hand is concerned with the direct extraction of the rules themselves. In [1] the authors introduced an unifying framework, allowing the study of problems related to classification, associations and sequences in an integrated fashion. More recently, several schemes directed to the extraction of rules from sequential data have been proposed - [2], [3], [5],[12], [16] and [19]. In special, the concepts of Events and Episodes formalized in [19], and the extraction of local patterns in time series discussed in [12] bear special significance for the subject of this paper. As discussed in the sequel, Phenomenon Characterization in the output time series and Precursor Characterization in the input time series can be understood as special cases of Time Series Quantization and resulting Event Alphabet Construction, in the spirit of [12]. Furthermore, the **Precursor**  $\Rightarrow$  **Phenomenon** rules can be understood as a special type of Serial Episode Rule, in the spirit of [19].

Given a fixed output time series containing one or more **Phenomenon** events, a key contribution of this paper is to show that a Classical Statistical Method - the Granger Causality Test (GCT, [14]) - can be used for ranking candidate time series according to the likelihood that rules of the form **Precursor**  $\Rightarrow$  **Phenomenon** will be present. GCT is applied to the raw time series, i.e. before Quantization and Event Construction. Hence, the time granularity ([5]) needed for optimal Phenomenon Characterization and Precursor Characterization is not an explicit parameter in the proposed method. Generally speaking, the candidate input time series are ranked according to the likelihood that good rules will exist. It is also expected that the higher the rank (Granger Causality Index - GCI), the simpler will be the resulting Quantization Problem. The GCT has been originally formulated for inferring causality among stationary processes. We show in this paper that valuable inferences can be drawn when applying the test to pairs of time series containing time-localized information. Under certain idealized conditions and suitable approximations, it is shown that the GCI corresponding to the input-output pair  $(u, y)$  is proportional to the confidence of the **Precursor**  $\Rightarrow$  **Phenomenon** rules to be extracted from this input-output pair. It suggests the GCT as an *Exploratory Tool* to extract pairs of time series more likely to contain good rules. Roughly speaking, if the GCI is high for the pair  $(u, y)$ , and time-localized structure is present in both time series, it signifies that **Precursor**  $\Rightarrow$  **Phenomenon** rules having high confidence are likely to exist for the pair.

The remaining of the paper is organized as follows: In section 2 we define Precursor Rules in terms of events and sequences of events, and propose a methodology for extracting such rules from databases containing time series related to different regimes of a system. Section 3 discusses the Granger Causality Test, and suggests its role as an Exploratory Tool for extracting Precursor Rules. Section 4 presents an analytic investigation of the utilization GCT for time series pairs containing impulsive time-localized structure. Following a number of approximations, the GCI is related with the confidence of the Precursor Rules extracted from time series pairs. Section 5 illustrates the applicability of the methodology discussed in 2 for the case of Network Security. Using MIB (Management Information Base) datasets collected from real experiments involving Distributed Denial of Service Attacks (eg. [11]), it is shown that Precursor Rules relating activities at Attacking Machines with Traffic Floods at Target Machines can be extracted by the method. Section 6

closes the paper, with our Conclusions.

## 2 A Methodology for Extracting Precursors - Assumptions, Objectives and Procedure

### 2.1 Notation and Definitions

#### Time Series, Multivariate Time Series and Collections

A time series is an ordered finite set of numerical values recorded from a variable of interest. It is assumed that the time elapsed between the recording of two consecutive elements is constant. The  $k$ th element of the time series  $\{z(k)\}$  is denoted as  $z(k)$ , where  $k = 0, 1, \dots, N - 1$  and  $N$  denotes the number of elements in  $\{z(k)\}$ . A multivariate time series of dimension  $m$  is an ordered finite set of  $m \times 1$  numerical vectors collected from  $m$  variables of interest. The  $k$ th element of the multivariate time series  $\{Z(k)\}$  is denoted as  $Z(k) = [z_1(k) \ z_2(k) \ \dots \ z_m(k)]^T$ , where  $z_i(k)$ ,  $i = 1, 2, \dots, m$  and  $k = 0, 1, \dots, N - 1$  are the  $k$ th elements of the individual time series  $z_i$  that form  $Z$ . It is assumed that all  $z_i(k)$  are recorded at the same instant of time, which allows the common index  $k$  to be used for their specification. The multivariate time series  $\{Z(k)\}$  is represented as an  $m \times N$  matrix of numerical variables. We call  $m$  the dimension of  $\{Z(k)\}$ , and  $N$  its size. A collection of multivariate time series is a finite set of multivariate time series corresponding to the same variables, but not necessarily having the same size. The  $j$ th element of the collection  $\mathcal{Z}$  is denoted as  $Z^j$ ,  $j = 1, 2, \dots, C$ . Collections of multivariate time series will be associated with the regimes of operation of the system of interest. In the earthquake prediction example, one can consider a Normal Collection  $\mathcal{N}$ , corresponding to periods of time when no earthquake is detected, and an Abnormal Collection  $\mathcal{A}$ , corresponding to periods of time during which earthquakes are known to be present. In the case of Network Security,  $\mathcal{N}$  corresponds to normal network activity, while  $\mathcal{A}$  corresponds to periods of time during which an attack is detected.  $C_{\mathcal{N}}$  denotes the number of elements in the collection  $\mathcal{N}$ , while  $C_{\mathcal{A}}$  denotes the number of elements in collection  $\mathcal{A}$ . Finally, we call the dataset  $\mathbf{ID}$  as the union of the two collections.  $\mathbf{ID}$  represents the Training Set, from where knowledge is to be extracted.

#### Events, Event Sequences and Precursor Rules

Events are defined in [19] as an ordered pair  $(A, \kappa)$  where  $\kappa = 0, 1, 2, \dots, K - 1$  is a time index representing the occurrence time of the event and  $A \in \mathcal{E}$  is an Event Type.  $\mathcal{E}$  is a finite set in [19], which we call the Event Alphabet. Event types provide a framework for transforming the raw time series data into more meaningful descriptions. As an example, consider the following procedure for transforming a time series  $\{z(k)\}$  having an even size  $N$ , into an event sequence  $\{\epsilon(\kappa)\}$  with size  $K = \frac{N}{2}$  and Event Alphabet  $\mathcal{E} = \{E_1, E_2\}$ :

- If  $[z(k + 1) + z(k)] \leq 200$ , Then  $\epsilon(\kappa) = E_1$ , for  $\kappa = \frac{k}{2}$  and  $k = 0, 2, 4, \dots, N - 2$ .
- Otherwise,  $\epsilon(\kappa) = E_2$ , for  $\kappa = \frac{k}{2}$  and  $k = 0, 2, 4, \dots, N - 2$ .

$z(k)$  can denote the number of alarms issued by a network monitoring device during a single day. A system administrator may only be interested in monitoring a higher level alarm, defined by  $\epsilon(\kappa)$ , i.e. every couple of days<sup>1</sup> check if more than 200 alarms were issued. If yes, a message is sent. Otherwise, nothing happens. The transformation from the time series space into the event space is the process of Time Series Quantization. The selection of the “right” parameters for performing the Quantization depends on the problem at hand. If  $m$  time series  $\{z_i(k)\}$ ,  $k = 1, 2, \dots, m$  are quantized according to the same procedure along the time index, producing  $m$  event sequences  $\{\epsilon_i(\kappa)\}$ ,  $i = 1, 2, \dots, m$ , we can define Multivariate Event Sequences and Collections of Multivariate Event Sequences the same way we defined their Time Series counterparts in section 2.1. It is understood that the events  $\epsilon_1(\kappa) \in \mathcal{E}_1$ ,  $\epsilon_2(\kappa) \in \mathcal{E}_2$ ,  $\dots$   $\epsilon_m(\kappa) \in \mathcal{E}_m$  are all recorded at the same instant  $\kappa$ , although the individual Event Alphabets  $\mathcal{E}_i$ ,  $i = 1, 2, \dots, m$  are not necessarily the same.

**Definition 1 (Causal Rule - [12]).** If  $A$  and  $B$  are two events, define  $A \xrightarrow{\tau} B$  as the rule: If  $A$  occurs, then  $B$  occurs within time  $\tau$ . We say that  $A \xrightarrow{\tau} B$  is a Causal Rule.  $\square$

<sup>1</sup>A sliding window could also have been used in defining  $\epsilon(\kappa)$ . It would not change our discussion and developments.

**Definition 2 (Precursor Rule).** If  $A$  and  $B$  are two events, define  $A \stackrel{\tau}{\leftarrow} B$  as the rule: If  $B$  occurs, then  $A$  occurred not earlier than  $\tau$  time units before  $B$ . We say that  $A \stackrel{\tau}{\leftarrow} B$  is a Precursor Rule.  $\square$

Causal Rules and Precursor Rules are special cases of Temporal Rules, discussed in [1]. Clearly, the rules  $A \stackrel{\tau}{\leftarrow} B$  and  $A \stackrel{\tau}{\Rightarrow} B$  are not the same. Notice that  $B$  is the antecedent of the Precursor Rule, while  $A$  is the antecedent of the Causal Rule. Hence, the confidence of the Precursor Rule -  $c(A \stackrel{\tau}{\leftarrow} B)$  - is the fraction of occurrences of  $B$  that were preceded by  $A$  within  $\tau$  units. In the problem at hand,  $A$  and  $B$  are events recorded at two different event sequences. If  $c(A \stackrel{\tau}{\leftarrow} B) = 1$ , it means that if  $B$  occurs, then  $A$  always occurred not earlier than  $\tau$  units before  $B$ , and is therefore a precursor of  $B$ , in the usual sense of the word. It *does not* mean however that all occurrences of  $A$  are followed by  $B$ .

The proposed methodology discovers Precursor Rules of the type  $A \stackrel{\tau}{\leftarrow} B$  in  $\mathbb{ID}$ , but utilizes the associated Causal Rule  $A \stackrel{\tau}{\Rightarrow} B$  for detection. The reason for this procedure is clear: we first characterize the phenomenon (item  $B$ ) and then search for the precursors (item  $A$ ). In summary, we mine Precursor Rules, but check for Causal Rules.

## 2.2 Assumptions, Problem Set-Up, Objectives and Procedure

### Assumptions

1. The variables are recorded as two collections of multivariable time series of dimension  $m$ . Collection  $\mathcal{N}$  corresponds to normal operation, while collection  $\mathcal{A}$  corresponds to abnormal operation. Typically  $C_{\mathcal{A}} \ll C_{\mathcal{N}}$ .
2. The  $m$  variables can be split into two subsets: Output variables  $y_i, i = 1, 2, \dots, m_1$  and **candidate** input variables  $u_i, i = 1, 2, \dots, m_2$ , with  $m_1 + m_2 = m$ .
3. The output variables are the ones in which the phenomenon of interest manifests itself. In the Financial Markets example, typical outputs are the various indices of market performance, or the performance of an individual stock of interest. The phenomenon is typically a sudden increase or decrease of the index, although more subtle statistical variations could be studied within this framework. It is assumed however the Phenomenon is time-localized.
4. The Phenomenon is only observed at collection  $\mathcal{A}$ .
5. The candidate input variables correspond to variables that may be or may not be related to the occurrence of the phenomenon observed in the output variables. In the Financial Markets example, typical candidate inputs can be macroeconomic variables, such as money supply, unemployment rates, etc. or the performance of other related stocks. It is assumed that the “true” input variables have some time-localized structure, that correspond to the Precursors.

### Problem Set-Up and Objectives

Given the assumptions above, we identify three interrelated problems related to the extraction of knowledge from the dataset  $\mathbb{ID}$ . To simplify our discussion, it is assumed that  $m_1 = 1$ , i.e. the Phenomenon is only observed on a single output.

**Problem 1: Phenomenon Characterization** Given the output time series, Phenomenon Characterization is related to the definition of a suitable Event Space, through Time Series Quantization. Let us return to the example in section 2.1. Define the quantity  $\zeta(\kappa) := z(k) + z(k+1)$ , for  $\kappa = \frac{k}{2}$  and  $k = 0, 2, 4, \dots, N-2$ , and the time series  $\{\mathcal{N}\zeta^j(\kappa)\}, j = 1, 2, \dots, C_{\mathcal{N}}$ , and  $\{\mathcal{A}\zeta^j(\kappa)\}, j = 1, 2, \dots, C_{\mathcal{A}}$ , which are the  $\zeta$  time series belonging to Collections  $\mathcal{N}$  and  $\mathcal{A}$ . Finally, define  $\text{Max}(\mathcal{A}\zeta^j) := \max_{\kappa} \{\mathcal{A}\zeta^j(\kappa)\}$ ,  $\text{Max}(\mathcal{N}\zeta^j) := \max_{\kappa} \{\mathcal{N}\zeta^j(\kappa)\}$ , and the *overall* maxima  $\text{Max}(\mathcal{A}\zeta) := \max_j (\mathcal{A}\zeta)$ ,  $\text{Max}(\mathcal{N}\zeta) := \max_j (\mathcal{N}\zeta)$ . If  $\text{Max}(\mathcal{N}\zeta) = 100$ , and  $\min_j \text{Max}(\mathcal{N}\zeta^j) = 300$ , a threshold of say, 200 separates the two collections. Event sequences constructed using the procedure in the example of section 2.1 will be such that the event type  $E_2$  never occurs on time series belonging to collection  $\mathcal{N}$ , and will occur at least once in all time series belonging to collection  $\mathcal{A}$ . If the occurrences of

$E_2$  are time-localized, i.e.  $A\epsilon^j(\kappa) = E_1$  most of the time, except for few isolated spikes where  $E_2$  is present, then  $E_2$  is a good characterization of the Phenomenon of interest. In many problems, such as Earthquakes, Stock Market Crashes and Computer Attacks the problem of Phenomenon Characterization is very simple. As shown in section 5, the output variables related to traffic counting in machines that are targets of Denial of Service Attacks records readings of 50,000 units during an Attack, compared with about 100 units during normal operation. Also, these bursts are time localized, i.e. the time series in collection  $\mathcal{A}$  remain at readings of about 100 units (similar to collection  $\mathcal{N}$ ), except for the bursts characterizing the presence of the attack.

**Problem 2: Identifying the Input Variables** This is the main focus of interest in this paper. The objective is to select which among the  $m_2$  variables contain precursors for the phenomenon observed in the output. The objective is to obtain time series for which high confidence Precursor Rules of the type  $A \xrightarrow{\tau} B$  exist, where  $B$  is the Phenomenon Characterized in Problem 1, while  $A$  is an event extracted from the candidate time series. Notice that  $\tau$  is not known. Hence, it is not advisable at this stage to do Time Series Quantization at the candidate inputs, as valuable Precursor Information may be destroyed in the process. Clearly, we need a procedure capable of performing the following two tasks: **(1) Detection:** Given an input-output pair  $(\{u(k)\}, \{y(k)\})$  measure the likelihood that a rule of the type  $A \xrightarrow{\tau} B$  exists, where  $A$  is a Precursor extracted from  $\{u(k)\}$  and  $B$  is a Phenomenon extracted from  $\{y(k)\}$  *without knowing the true nature of the Precursor, or the delay between Precursor and Phenomenon*; **(2) Gradation:** Given a fixed output and  $m_2$  candidate inputs, rank candidate input variables according to the likelihood that Precursor Rules exist, *without knowing the true nature of the Precursor, neither the delay between Precursor and Phenomenon*. We show that the GCT is an adequate procedure for addressing both tasks.

**Problem 3: Precursor Characterization** Given the input variables that are most likely to contain Precursors, the problem of Precursor Characterization is to extract the Precursors as time-localized occurrences in the time series through a process of Time Series Discretization of the same nature as Problem 1. The key point however, is that following the solution of Problem 2, one has evidence that these time-localized occurrences give rise to Event Types that are related to the Phenomenon at the output through a rule of the type  $A \xrightarrow{\tau} B$ .

### Procedure

Based on the above, we suggest the following Procedure for extracting Precursor Rules relating Phenomenon in the outputs with Precursors at candidate inputs:

- **Step 1:** Solve Problem 1 through adequate Time Series Quantization at the output time series.
- **Step 2:** Solve Problem 2 by applying the GCT to all input-output pairs  $\{(u_i(k), y(k))\}, i = 1, 2, \dots, m_2$ , and compute the GCI  $g_i$  corresponding to each candidate input. Select candidate inputs either by setting a threshold on  $g_i$ , or by choosing the top  $v$  scores, where typically  $v \ll m_2$ .
- **Step 3:** Solve Problem 3 through adequate Time Series Quantization of the input time series selected on Step 2. The objective is to extract time-localized structures at the selected inputs that precede the Phenomenon in the output. Discard input time series that do not show time-localized structure preceding the Phenomenon <sup>2</sup>. At the end of this step, one has determined the **Phenomenon** and **Precursor** events of interest, as well as a number of candidate Precursor Rules of the form **Precursor**  $\xleftarrow{\tau}$  **Phenomenon**.
- **Step 4:** Compute the confidence of the *associated* Causal Rules **Precursor**  $\xrightarrow{\tau}$  **Phenomenon**, and select the best ones either by thresholding or ranking. At this Step we are verifying if indeed the Precursor events at the inputs are preceding the Phenomenon events at the output.

<sup>2</sup>For the example discussed in section 5, the extraction of time-localized structures at candidate inputs is shown to be very simple. However, this may be a difficult problem in general, as Precursors in the inputs may not be as evident as the Phenomenon in the output. Sophisticated change detection techniques (eg. [4]) may be needed in general cases.

In the next section we will describe the GCT, and demonstrate its suitability as an Exploring Tool for Knowledge Discovery, targeted on Step 2 above. In section 5 we describe the results obtained when applying this procedure in the extraction of Precursor Rules relating Attacking Nodes with Target Nodes in various types of Distributed Denial of Service Attacks.

### 3 Detection and Gradation of Causality in Time Series

#### 3.1 Notation and Definitions

The following notation and nomenclature is used extensively in the following sections. It is commonly used in Statistics and Systems Science. The reader is referred to standard textbooks in these areas for more details - eg. [15] or [17].

**Shift Operators, Transfer Functions and Impulse Responses** Given a time series  $\{z(k)\}$ ,  $k = 0, 1, \dots, N - 1$ , the backward and forward shift operators  $q$  and  $q^{-1}$  are defined as follows:  $qz(k) := z(k + 1)$ ,  $k = 0, 1, \dots, N - 2$  and  $q^{-1}z(k) := z(k - 1)$ ,  $k = 1, 2, \dots, N - 1$ . The backward and forward shift operators are used to describe dynamical input-output relations among variables. In particular, the expression  $y(k) = \frac{\beta(q^{-1})}{\alpha(q^{-1})}u(k) = T(q^{-1})u(k)$  where  $\alpha(q^{-1}) = 1 + \sum_{\ell=1}^p \alpha_{\ell}q^{-\ell}$ ,  $\beta(q^{-1}) = \sum_{\ell=0}^p \beta_{\ell}q^{-\ell}$  denotes  $y(k) = -\sum_{\ell=1}^p \alpha_{\ell}y(k - \ell) + \sum_{\ell=0}^p \beta_{\ell}u(k - \ell)$ , for  $p + 1 \leq k \leq N - 1$ .  $T(q^{-1})$  is called the Transfer Function between  $\{u(k)\}$  and  $\{y(k)\}$ .  $T(q^{-1})$  is a stable Transfer Function if  $\alpha(q^{-1})$  is a Hurwitz polynomial, i.e. all the zeros of  $\alpha(q^{-1})$  belong to the open unit disk. In this case, we can write  $\frac{\beta(q^{-1})}{\alpha(q^{-1})} = t(q^{-1}) = \sum_{\ell=0}^{\infty} t_{\ell}q^{-\ell}$  and  $t(q^{-1})$  is called the Impulse Response associated with the Transfer Function  $T(q^{-1})$ . It well known that  $\lim_{\ell \rightarrow \infty} t_{\ell} = 0$ , and  $\sum_{\ell=0}^{\infty} t_{\ell}^2 =: \|T\|_2^2 < \infty$ , where  $\|T\|_2$  is called the  $\mathcal{L}_2$  norm of  $T(q^{-1})$ . The relationship between  $\{u(k)\}$  and  $\{y(k)\}$  is written in terms of the Impulse Response as  $y(k) = \sum_{\ell=0}^{\infty} t_{\ell}u(k - \ell)$ . If there is a natural number  $L$  such that  $t_{\ell} = 0$  for  $\ell \geq L + 1$ , we say that  $T(q^{-1})$  is a Finite Impulse Response (FIR) of size  $L$ .

**Probability Distributions**  $x \sim X$  indicates that the random variable  $x$  has the distribution  $X$ .  $\mathbb{E}(x)$  denotes the expected value of  $x$ .  $n(\mu, \sigma^2)$  denotes a Gaussian, or Normal Random Variable with mean  $\mu$  and variance  $\sigma^2$ .  $\chi_v^2$  denotes a Chi-Square distribution with  $v$  degrees of freedom.  $F(v_1, v_2)$  denotes an  $F$  distribution with parameters  $v_1$  and  $v_2$ .  $\Gamma^{-1}(v_1, v_2)$  denotes an Inverse-Gamma distribution with parameters  $v_1$  and  $v_2$ . The definitions and properties of these distributions are given in [13].

#### 3.2 The Granger Causality Test as an Exploratory Tool

Testing for causality in the sense of Granger involves using statistical tools for testing whether *lagged* information on a variable  $u$  provides any statistically significant information about the variable  $y$ . If not, then  $u$  does not Granger-cause  $y$ . The Granger Causality Test (GCT - [14]) compares the residuals of an AutoRegressive Model (AR Model) with the residuals of an AutoRegressive Moving Average Model (ARMA Model). Assume a particular lag length  $p$ , and estimate the  $a_i$  and  $b_i$  parameters ( $1 \leq i \leq p$ ) in the following unrestricted equation:

$$y(k) = \sum_{i=1}^p \alpha_i y(k - i) + \sum_{i=1}^p \beta_i u(k - i) + e_1(k) \quad (1)$$

Parameter estimation is performed using Ordinary Least Squares (OLS) - [15]. If  $\{y(k)\}$  and  $\{u(k)\}$  are time series of size  $N$ , it results on a regression with  $T := N - p$  equations, out of which  $2p$  parameters are estimated. The computational cost of the procedure is  $O(T^2)$ . The Null Hypothesis  $H_0$  of the GCT is given by:

$$H_0 : \quad \beta_i = 0, \quad i = 1, 2, \dots, p,$$

i.e.  $u$  does not affect  $y$  up to a delay of  $p$  units. The null hypothesis is tested by estimating the parameters of the following restricted equation

$$y(k) = \sum_{i=1}^p \delta_i y(k-i) + e_0(k) \quad (2)$$

Again, estimation of the  $\delta$  parameters lead to an OLS problem with  $T$  equations. The procedure of the GCT is as follows. Let  $R_1$  and  $R_2$  denote the sum of the squared residuals under the two cases:

$$R_1 = \sum_{k=1}^T e_1^2(k), \quad R_0 = \sum_{k=1}^T e_0^2(k)$$

If the Granger Causality Index (GCI)  $g$  given by:

$$g = \frac{(R_0 - R_1)/p}{R_1/(T - 2p - 1)} \sim F(p, T - 2p - 1) \quad (3)$$

is greater than the specified critical value for the  $F$ -test, then reject the null hypothesis that  $u$  does not Granger-cause  $y$ . As  $g$  increases, the  $p$ -value<sup>3</sup> associated with the pair  $(\{u(k)\}, \{y(k)\})$  decreases, lending more evidence that the Null Hypothesis is *false*. In other words, high values of  $g$  are to be understood as representing strong evidence that  $u$  is causally related to  $y$ . In this work, we utilize the GCT in an exploratory manner, to compare the causality strength of two candidate input time series with respect to a given output. Following the  $p$ -value interpretation, we say that  $\{u_1(k)\}$  is more likely to  $\{u_2(k)\}$  to be causally related with  $\{y(k)\}$  if  $g_1 > g_2$ , where  $g_i$ ,  $i = 1, 2$  denote the GCI for the input-output pair  $(u_i, y)$ . We may be interested in selecting the top 5 or 10 individual candidate input time series that are more likely to be causally related to  $\{y(k)\}$  for more detailed inspection. The GCI is an adequate index to perform this selection.

**Remark 1 (Measuring the causality strength of multiple inputs).** GCT can be extended to the case when one has multiple inputs, and wishes to decide which are the most relevant  $n$ -ples out of a large set of candidates (eg. [15], [6]). As in other Regression Problems, the most relevant  $n$ -ple is not necessarily the one having the top individual scores. One has to compute the combined score of each  $n$ -ple, which in general is not monotonically related to the sum of the individual scores from each element.  $\square$

**Remark 2 (Causality and Data Mining - Related Work).** In [21] the authors investigate the extraction of causal relationships in market basket data. The Causality Tests in their works is essentially different than ours, since time component is not present in their work. Their objective is to disentangle causal relationships from triples of interrelated static variables. The Precursor Rules investigated in the paper are different than the Causal Rules studied in [21].  $\square$

## 4 GCT and the Extraction of Precursor Rules - Modeling and Theoretical Developments

In usual statistical practice, the GCT is utilized to decide if a given  $u$  causes  $y$  for a specified significance level. No assumption is made about the presence (or absence) of localized structure in the time series. In these cases, one is interested in gauging how the time series  $u$  affects the time series  $y$  as a whole. However, in the problem at hand, we are ultimately interested in extracting rules relating time-localized segments of the time series. In our context, the GCT is merely an intermediate step in this process. We argue as follows: if time-localized structures at  $u$  consistently precede time-localized structures at  $y$ , there is good evidence that events in  $u$  are related to events in  $y$ . The localized structures will be examined separately at  $u$  and  $y$  after the existence of a causality relation between the two time series is suggested by GCT. The determination of these structures correspond to Steps 1 and 3 in section 2. In this section, we investigate how GCT behaves

<sup>3</sup>The  $p$ -value of a Statistical Test is the smallest significance level that leads to the rejection of the Null Hypothesis - [10], p. 364.

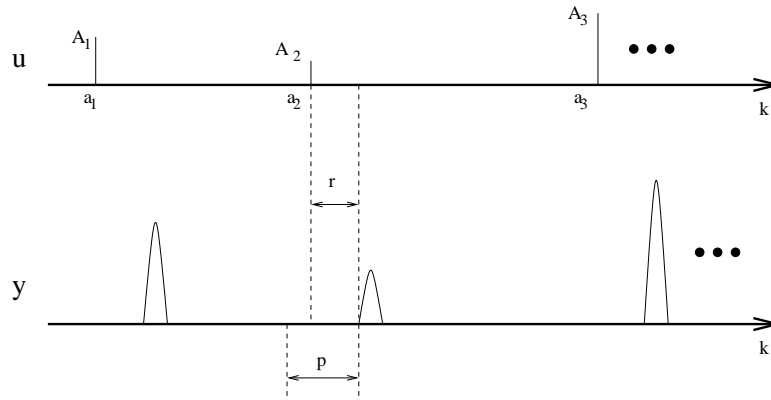
when time series with localized structure are tested for causality. If GCT is used to identify the presence of a causal relationship between two variables, it is implicitly assumed that these variables are related by a model of the form (1), where  $e_1(k)$  is a noise term that captures the mismatch between the recordings and the model output. Let  $y$  represent the output variable that displays the phenomenon, and  $u$  represent the unknown input variable which triggers the phenomenon. Following section 2, we model the relationship between  $u$  and  $y$  as follows:

$$y(k) = q^{-r} \gamma H(q^{-1}) u(k) + w(k) \quad (4)$$

$r$  represents the delay between the input and the output,  $\gamma$  is an amplification gain typically large<sup>4</sup> ( $\gamma \gg 1$ ), while  $H(q^{-1})$  models the dynamic interaction between  $u$  and  $y$ . It is assumed that  $H(q^{-1})$  is a FIR with size  $L$ . It is also assumed that an upper bound  $\rho$  for  $r$  is available, i.e. we are only interested in Precursor Rules of the form  $A \xrightarrow{\rho} B$ , where  $1 \leq r \leq \rho$ . Knowledge of  $\rho$  is needed for selecting the prediction window  $p$  for performing GCT. The obvious choice for  $p$  is to take  $p = \rho + L$ . The event  $A$  is to be mined from  $u$ , while the event  $B$  is to be mined from  $y$ .  $w(k)$  is the noise process, which we assume to be zero mean Gaussian white noise with variance  $\sigma^2$ , i.e.  $w(k) \sim n(0, \sigma^2)$ , for all  $k$ . We expect this model to be a good description for collection  $\mathcal{A}$ . Let  $H(q^{-1}) = \sum_{\ell=0}^{\infty} h_{\ell} q^{-\ell}$  i.e.  $H(q^{-1})$  is written in terms of its Impulse Response as outlined in section 3.1. To complete the modeling, we need to characterize the input signal  $u$ . Since we will be looking for time-localized events, we assume that  $u$  has a time-localized structure, as depicted in Figure 1. In particular,  $u(k)$  is defined as:

$$\begin{aligned} u(a_i) &= A_i, \quad i = 1, 2, \dots, n, \quad \text{where } a_1 < a_2 < \dots < a_n < N - p, \quad \text{and } a_i - a_{i-1} > \rho + L, \quad i = 1, 2, \dots, n \\ u(k) &= 0, \quad \text{for } k \neq a_i \end{aligned} \quad (5)$$

Here,  $N$  is the size of the collected dataset. In the absence of noise ( $w(k) \equiv 0$ ), the output  $y(k)$  essentially follows its impulse response each time inputs are applied at  $a_i$ ,  $i = 1, \dots, n$ , as depicted in Figure 1. The blips at  $u$  happening at each time sample  $a_i$  model the Precursors, while the response at  $y$  according to the Impulse Response of  $\gamma H(q^{-1})$  models the Phenomenon.



**Figure 1.** The idealized inputs and output signals.  $p$  is the length of the window used for parameter estimation when applying GCT. If  $p \geq r + L$ , the representation (1) captures model (4) exactly, and the **Precursor**  $\Rightarrow$  **Phenomenon** rule is “visible” through the model.

If the GCT is applied to time series coinciding with the idealized signals, we will have  $g = \infty$ , since  $R_0$  is finite, while  $R_1 = 0$ , assuming the parameters of  $q^{-r} \gamma H(q^{-1})$  are correctly estimated in equation (1)<sup>5</sup>. Notice that the time elapsed between two consecutive blips ( $a_i - a_{i-1}$ ) is assumed to be larger than the time elapsed between the blip and the entire response due to the blip. This assumption serves to “isolate” each

<sup>4</sup>As described in section 2, it is expected that the Phenomenon will be much more pronounced than the Precursor.

<sup>5</sup>This will be true for the idealized signals under very mild conditions related to the order of the system and the number of blips in the input signal - [18].

individual **Precursor**  $\Rightarrow$  **Phenomenon** occurrence. When  $\sigma > 0$ , one is interested in evaluating how GCT performs, i.e. in determining the relationship between the GCI and the other quantities in the problem.  $g$  is a random variable in this case, so  $\mathbb{E}(g)$  is the quantity of interest. Theorems 1 and 2 are the main results of this section:

**Theorem 1 (The GCI for the idealized signals).** Assume that time series  $\{y^*(k)\}$  and  $\{u^*(k)\}$  with size  $N$  are generated from equation (4) with  $u^*(k)$  given by equation (5),  $w(k) \sim n(0, \sigma^2)$ , for  $k = 0, 1, \dots, N-1$ , and  $H(q^{-1})$  is an FIR with size  $L$ . If  $\gamma^2(\sum_{i=1}^n A_i^2) \gg \sigma^2$  and  $p \geq r + L$ , the expected value of GCI for the pair  $(\{u^*(k)\}, \{y^*(k)\})$  computed using equation (3) can be approximated by:

$$\mathbb{E}(g^*) \approx \frac{N - 3p - 1}{N - p - 2} \gamma^2 \|H(q^{-1})\|_2^2 \frac{(\sum_{i=1}^n A_i^2)}{p\sigma^2} \quad \square \quad (6)$$

**Theorem 2 (The GCI for input signals missing a few blips).** Assume  $\{y^*(k)\}$  and  $\{u^*(k)\}$  with size  $N$  satisfy the conditions in Theorem 1. Let  $\mathcal{I} \subset \{1, 2, \dots, n\}$  denote a non-empty collection of indices. Define  $u^{\mathcal{I}}(k)$  as follows:

$$u^{\mathcal{I}}(a_i) = A_i, \text{ if } i \in \mathcal{I}, \quad u^{\mathcal{I}}(k) = 0, \text{ otherwise} \quad (7)$$

i.e.  $u^{\mathcal{I}}(k)$  has only a fraction of the blips present in  $u^*(k)$ . If  $\gamma^2(\sum_{i \in \mathcal{I}} A_i^2) \gg \sigma^2$  and  $p \geq r + L$ , the expected value of GCI for the pair  $(\{u^{\mathcal{I}}(k)\}, \{y^*(k)\})$  computed using equation (3) can be approximated by:

$$\mathbb{E}(g^{\mathcal{I}}) \approx \frac{N - 3p - 1}{p} \gamma^2 \|H(q^{-1})\|_2^2 \frac{(\sum_{i \in \mathcal{I}} A_i^2)}{(N - p - 2)\sigma^2 + \gamma^2 \sum_{i \notin \mathcal{I}} A_i^2} \quad \square \quad (8)$$

The proof of Theorem 1 is sketched in Appendix A. Due to space limitations, the proof of Theorem 2 is omitted. A number of key observations can be made.

1. The term  $S^* := \frac{(\sum_{i=1}^n A_i^2)}{p\sigma^2}$  can be understood as the Signal-to-Noise Ratio (SNR) between the input blips carrying the Precursors to be extracted, and the noise present in the estimation window of length  $p$ .  $\mathbb{E}(g^*)$  grows with  $S^*$ , which intuitively means that for  $u$  signals of the type shown in Figure 1, if a larger value of  $g$  is observed, it indicates that it is more likely that Precursors could be found. This is certainly a desirable property.
2. The term  $\lambda := \gamma^2 \|H(q^{-1})\|_2^2$  can be understood as an amplification gain between the input and the output. GCI is proportional to  $\lambda$ .
3. As  $N \rightarrow \infty$ ,  $\mathbb{E}(g^*)$  converges to  $\lambda S^*$ , which is a constant. Hence, the corresponding expected  $p$ -value converge to zero. The interpretation is that for a fixed SNR, the certainty that the pair  $(u, y)$  is Granger causal grows with  $N$ . This is also a desirable property.
4. Finally, we turn our attention to equation (8). This expression allows one to relate  $\mathbb{E}(g^{\mathcal{I}})$  with the confidence of Precursor Rules extracted from the pair  $(\{u^{\mathcal{I}}(k)\}, \{y^*(k)\})$ . If  $\{u^{\mathcal{I}}(k)\}$  is selected in Step 2, the resulting Time Series Quantization in Step 3 is trivial: just define  $\epsilon(k) = E_1$  if  $u^{\mathcal{I}}(k) > 0$ ,  $\epsilon(k) = E_2$ , otherwise. Consider now the Precursor Rule  $E_1 \stackrel{\mathcal{I}}{\Leftarrow} B$ , where  $B$  is obtained by performing Step 1 in  $\{y^*(k)\}$ . The confidence of this rule is given by  $\frac{\#(\mathcal{I})}{n}$ , where  $\#(\mathcal{I})$  denotes the number of elements in  $\mathcal{I}$ . This follows from the fact that among the  $n$  times event  $B$  occurs, the  $E_1$  event occurs only  $\#(\mathcal{I})$  times. Consider now two time sequences  $\{u^{\mathcal{I}_1}(k)\}$  and  $\{u^{\mathcal{I}_2}(k)\}$  with  $\mathcal{I}_1 \subset \mathcal{I}_2$ , i.e.  $\{u^{\mathcal{I}_2}(k)\}$  contains all the blips contained in  $\{u^{\mathcal{I}_1}(k)\}$  plus a few more. It is clear from equation (8) that  $\mathbb{E}(g^{\mathcal{I}_2}) > \mathbb{E}(g^{\mathcal{I}_1})$ . It essentially means that higher values of  $g$  are associated with Precursor Rules with higher confidence, which is another property that indicates the suitability of the GCT as an Exploratory Tool for extracting Precursor Rules.

**Remark 3 (Evaluating the approximations for  $\mathbb{E}(g^*)$  and  $\mathbb{E}(g^{\mathcal{I}})$ ).**

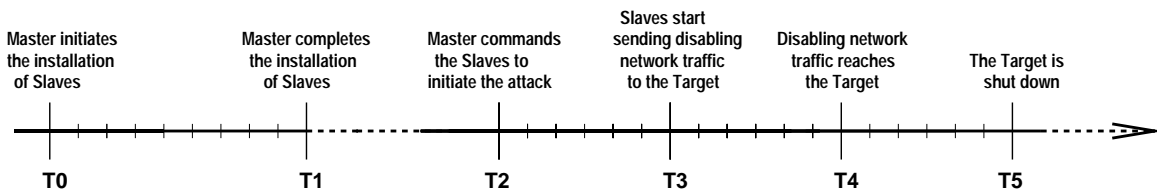
We have conducted several numerical experiments to evaluate the accuracy of the approximations in Theorems 1 and 2. Due to space limitations, only the highlights are given. We have considered transfer functions  $T(q^{-1})$  of order 5 and higher, and datasets including 10 to 20 input blips. Our main experiments were performed for datasets with sizes varying from 1,000 to 10,000.

1. For  $\sigma \rightarrow 0$  and holding the amplitude of the blips constant, we observed that  $\mathbb{E}(g^*)$  and  $\mathbb{E}(g^{\mathcal{I}})$  computed according Theorem 1 and Theorem 2 to become closer and closer to the empirical mean of the GCIs computed through Monte Carlo experiments following the statistical models in the statement of the Theorems. Typically, for  $\gamma \frac{\sum_{i=1}^n A_i^2}{\sigma^2} \approx 100$  we have the approximated values lying within 5% of the sample mean.
2. As  $\sigma$  increases, the approximation tends to give values higher than the empirical mean. But even for  $\gamma \frac{\sum_{i=1}^n A_i^2}{\sigma^2}$  as low as 5, the approximation remain within 20% of the sample mean.
3. The approximation for  $\mathbb{E}(g^{\mathcal{I}})$  was found to be effective for *ranking* candidate input variables, even for large values of  $\sigma$ . By effective we mean the following. We ran Monte Carlo experiments computing the GCI for inputs corresponding to various choices of  $\mathcal{I}$  in Theorem 2. The empirical means were computed in each case. We noticed that the *rank* of the inputs according to the empirical mean of GCI closely matches the *rank* computed following the  $\mathbb{E}(g^{\mathcal{I}})$  approximated by equation (8), even though the individual values of  $\mathbb{E}(g^{\mathcal{I}})$  do not match the corresponding empirical means.  $\square$

## 5 Precursor Rules for Distributed Denial of Service Attacks

### 5.1 DDoS Attacks and the experiments

Distributed Denial of Service (DDoS) attacks have two phases, and involve three classes of systems: the Master, the Slaves, and the Target (eg. [11]). In the first phase of the attack, the Master infiltrates multiple computer systems, and installs the DDoS tools, which are scripts capable of generating large volumes of traffic under command from the Master. We call these infiltrated systems the Slaves. The second phase is the actual DDoS attack. Under command from the Master, the Slaves generate network traffic to bring down the Target system. We assume that the Master is not under monitoring, but the Target and a few Slaves (not all) are. Figure 2 presents a simplified timeline for the DDoS attacks. A Data Set for studying DDoS



**Figure 2.** *DDoS Attacks - A simplified Timeline.*

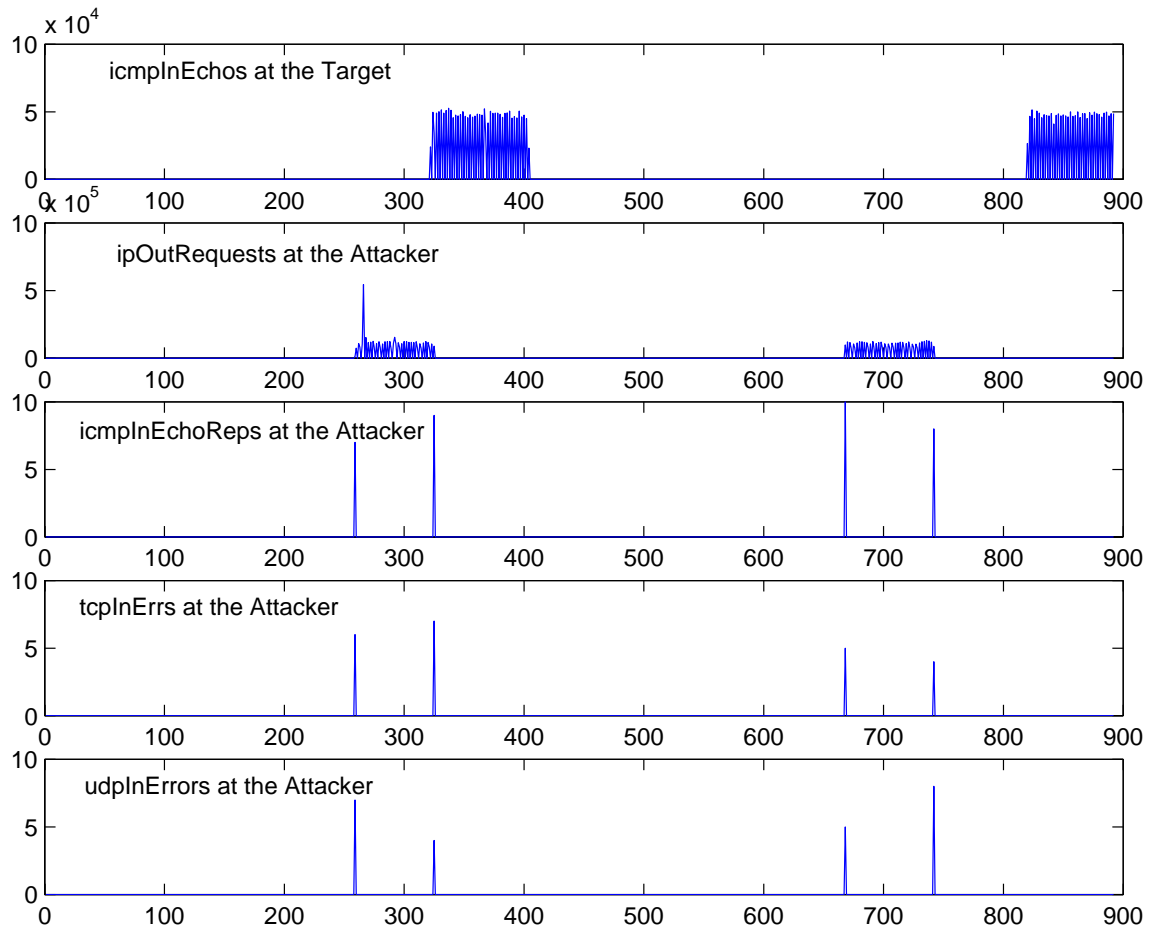
attacks was produced at North Carolina State University (NCSU). All the nodes (attackers and targets) were linked to a common Ethernet. The Network Management System collected 64 MIB variables corresponding to four SNMP MIB (Management Information Base - [22]) groups: `ip`, `icmp`, `tcp` and `udp`. Variables were collected for intervals of 2 hours, at a sample rate of 5 seconds. The details can be found in [9]. We used the data corresponding to both Attack Runs (Collection  $\mathcal{A}$ ) and Normal Runs (Collection  $\mathcal{N}$ ) as described below:

**Attack Runs - Collection  $\mathcal{A}$ :** Three types of DDoS attacks produced by TFN2K (Ping Flood and Targa3) and Trin00 (UDP Flood) were effected. During each of the attacks, MIBs were collected for the Attacking Machine and for the Target. TFN2K and Trin00 are the names of the hacker toolkits, while Ping Flood, Targa3 and UDP Flood are types of DoS attacks they induce. The time series for MIB variables corresponding

to counter variables were differentiated. Two runs were recorded for each type of attack. According to the terminology introduced earlier, Attacker 1 and Attacker 2 are Slaves; the Master is not under monitoring from the Network Management System, so no time series are available for the Master.

**Normal Runs Collection  $\mathcal{N}$ :** MIBs were collected during times when the machines were not being targets of attacks, nor being the attackers. 12 runs are available for the Target Machine, 7 runs are available for Attacker 1, and 14 runs are available for Attacker 2.

The data set includes events starting on T2, defined in Figure 2; the DDoS tools are assumed to be already installed in the Attacking Machines when the Attack Runs start. Hence, prospective Precursor Rules should relate events in T2 or T3 at the Attacker with events in T4 and T5 at the Target. To illustrate the nature of the MIB variables and their relevance for attack detection during a TFN2K Ping Flood Attack, Figure 3 depicts `icmpInEchos` at the Target, aligned with four MIB variables at the Attacker Machine that show remarkable activity before the pings reach the target. These are `ipOutRequests`, `icmpInEchoReps`, `tcpInErrs` and `udpInErrors`. These four variables were obtained from domain knowledge about the TFN2K Ping Flood attack. In practice, we need a procedure to extract these Key Variables for the Attacker automatically, from the entire collection of MIB data at the Attacker Machine. This is exactly the problem addressed in this paper. In the sequel, we show the results obtained using the methodology in section 2 for the case of the TFN2K Ping Flood Attack. Similar results were also verified for the other two types of DDoS attacks ([8], [9]).



**Figure 3.** *TFN2K Ping Flood: Selected MIB variables at the Attacker and at the Target.*

## 5.2 TFN2K Ping Flood - Extracting Precursor Rules

The four steps presented in section 2 were followed:

**Step 1: Phenomenon Characterization** The Ping Flood attack is effected by sending a large amount of ICMPECHOREQUEST packets to the Target. Clearly, `icmpInEchos` is the output in this case. As shown in Figure 3 Phenomenon Characterization is very simple, considering that `icmpInEchos` never rises above a few hundred unit during Normal Runs.

**Step 2: Extracting inputs containing Precursors** In this step, we attempt to determine the variables at the Attacker Machine that contain Precursors. Based on [11], we have domain knowledge (Ground Truth - Table 1) about the Key Variables at the Attacker for TFN2K. The GCT was applied for two runs of TFN2K Ping Flood. T4 events happen more than once in each run, as shown in Figure 3. To test the validity of the GCT for automatically extracting the Key Variables at the Attacker, we consider a scenario in which there are nine potential Attackers against the Target: the true attacker and eight decoys corresponding to the normal runs. We then apply the GCT to measure the causality strength of all MIB variables in the potential attackers, with respect to the Key Variable at the Target in each of the Attacks. MIB variables at potential attackers resulting on a GCI statistic above the threshold for 95% significance level were considered to Granger-cause the Key Variables at the Target, and were kept for analysis in Step 3. The selected variables and scores are shown in Table 2 for one of the Runs. Comparing Tables 1 and 2 it is clear that GCT extracted *all* Ground Truth Variables in this case. We count detections whenever the ground-truth variables described in [11] are correctly picked by the GCT. False alarms correspond to MIB variables being flagged in the decoys. Table 3 summarizes the results for both runs. Notice that at least one “true” MIB variable at the Attacker is detected in each run. The FA (False Alarm) Rate for Decoy MIBs is obtained by computing the total number of significant MIB variables found in all normal runs, divided by the total number of MIB variables.

**Steps 3 and 4: Precursors Characterization and Pruning the Precursor Rules** The Key Variables at the Attacker determined in Step 2 are labeled as causally related with the Attack at the Target, but we still need to find the Precursors. As discussed in section 2, we have a problem of Time Series Quantization. We looked for jumps in the MIB variables, by monitoring the absolute values of the differentiated time series  $z(k) = |y(k) - y(k - 1)|$ . Using 12 Normal Runs, we constructed a *Normal Profile of Jumps* for each of the 64 MIB variables. Given a Key Attacker Variable determined on Step 2, Key Events at the Attacker are defined as jumps larger than the largest jump encountered the *Normal Profile of Jumps*. Key Attacker Variables with no Key Events are discarded. As shown in Table 4, We have found that this procedure led to a substantial reduction of the False Alarms produced on Step 2, with small reductions in the detection rates. Notice that we are still detecting at least one valid precursor at each Attack Run.

MIB	Event
<code>icmpInEchoReps</code>	T2
<code>tcpInErrs</code>	T2
<code>tcpInSegs</code>	T2
<code>udpInErrors</code>	T2
<code>udpInDatagrams</code>	T2
<code>ipOutRequests</code>	T3

**Table 1.** *Key Variables at the Attacker for TFN2K - Ground Truth.*

Rank	MIB	$g$
1	ipOutRequests (T3)	5.26
2	tcpInErrs (T2)	3.50
3	ipInReceives	2.67
4	ipInDelivers	2.65
5	udpInErrs (T2)	2.63
6	udpOutDatagrams	2.58
7	udpInDatagrams (T2)	2.57
8	icmpInEchoReps (T2)	2.04
9	icmpInMsgs	1.99
10	tcpInSegs (T2)	1.31
11	udpNoPorts	1.27

**Table 2.** *TFN2K Ping Flood Run 1: Top MIBs at the Attacker according to the  $g$  statistic.*

Run	Detections	FA per Decoy MIBs (%)
1	6/6	4.49
2	1/6	3.13

**Table 3.** *Results of Step 2: Detection Rates and FA Rates for MIB variables that contain precursors to DDoS Attacks.*

Run	Detections	FA per Decoy MIBs (%)
1	4/6	1.37
2	1/6	0.52

**Table 4.** *Final Results: Detection Rates and FA Rates for Events at MIB variables for TFN2K Ping Flood.*

## 6 Conclusions

A principled methodology was proposed and evaluated for extracting Precursor Rules from Time Series. The paper presents an analytic investigation of the utilization of the Granger Causality Test for time series pairs containing impulsive time-localized structure. The Granger Causality Index is related with the confidence of the Precursor Rules extracted from these time series pairs. While the work has been motivated by our research in Network Security (eg. [9], [8]), we expect that the proposed methodology is general enough to be successfully applied in other fields.

**Acknowledgements** Scientific Systems Company acknowledges the continuing support from the Defensive Information Warfare Branch at the Air Force Research Laboratory in Rome, NY. We are particularly grateful to Mr. Peter J. Radesi and Dr. Leonard J. Popyack, Jr. from AFRL, for their encouragement. We are also grateful to Dr. Lundy Lewis from Aprisma Management Technologies, Prof. Wenke Lee and Mr. Xinzhou Qin from the Georgia Institute of Technology, for several discussions concerning the Computer Security experiments. The datasets used in the experiments described in section 5 were collected under support from Aprisma’s University Fellowship 1999/2000 to Prof. Wenke Lee at North Carolina State University. Finally, we are grateful to Ravi Prasanth, Melvyn Huff and Constantino Rago at SSCI for many enlightening discussions on System Identification.

## A Proof of Theorem 1

Due to the space limitations, we prove the result for the simpler case  $H(q^{-1}) = 1$ , i.e. the model relating the Precursor-carrying input and the Phenomenon-carrying output is given by:

$$y(k) = \gamma u(k-r) + w(k) \quad (9)$$

The general proof follow along the same lines, by replacing the  $\gamma A_i$  terms at the output localized at the  $a_i + r$  by the corresponding Impulse Responses. As described in Remark 3 we have verified the validity of the approximation for moderate values of  $\sigma^2$ . As in the statement of the Theorem,  $\{u^*(k)\}$  is produced by equation (5), and  $\{y^*(k)\}$  is produced as the output of the model given by equation (9) when the input is  $\{u^*(k)\}$ , i.e.  $y^*(k) = \gamma u^*(k-r) + w(k)$ .  $\{u^*(k)\}$  and  $\{y^*(k)\}$  are time series of size  $N$ . We now define the following predictors and prediction errors valid for  $k = p+1, p+2, \dots, N$ , and the corresponding Residual Sum of Squares:

- $\hat{y}_1^*(k)$ : Obtained by fitting an ARMA model (equation (1)) to the pair  $(\{u^*(k)\}, \{y^*(k)\})$ .
- $\hat{y}_0^*(k)$ : Obtained by fitting an AR model (equation (2)) to the pair  $(\{u^*(k)\}, \{y^*(k)\})$ .
- $e_1^*(k) := y^*(k) - \hat{y}_1^*(k)$ ,  $e_0^*(k) := y^*(k) - \hat{y}_0^*(k)$ ,  $R_1^* := \sum_{k=p+1}^N e_1^{*2}(k)$ ,  $R_0^* := \sum_{k=p+1}^N e_0^{*2}(k)$ .

If  $w(k) \equiv 0$ , it is clear that the Least Squares Problem associated with ARMA fitting for  $(\{u^*(k)\}, \{y^*(k)\})$  has the solution  $\alpha_i = 0$ ,  $i = 1, 2, \dots, p$ ,  $\beta_r = \gamma$ ,  $\beta_i = 0$  for  $i \neq r$ , i.e. the ARMA representation gives an exact fit for equation (9). If  $w(k) \neq 0$ , but its variance  $\sigma^2$  is small compared with  $\gamma^2(\sum_{i=1}^n A_i^2)$ , the expected estimated parameters should be close to their true values, by a continuity argument. Hence, the following approximation for  $\hat{y}_1^*(k)$  is warranted:

$$\hat{y}_1^*(k) \approx \gamma u^*(k-r)$$

Similarly, since  $a_i - a_{i-1} > p$  for all  $i$  (recall that  $H(q^{-1}) = 1$ , thus  $L = 0$ ), and  $w(k)$  is white noise, an AR fitting of  $\{y^*(k)\}$  does not differ from AR fitting of white noise. Hence, we approximate  $\hat{y}_0^*(k)$  as:

$$\hat{y}_0^*(k) \approx 0 \quad (10)$$

which is the least square estimator for gaussian white noise sequences with zero mean. Therefore, by making the approximations above we have:

$$\begin{aligned} e_1^*(k) &= y^*(k) - \hat{y}_1^*(k) = w(k) \\ e_0^*(k) &= y^*(k) - \hat{y}_0^*(k) = \gamma u^*(k-r) + w(k) \end{aligned}$$

which gives:

$$R_1^* = \sum_{k=p+1}^N w^2(k), \text{ and } R_0^* = \sum_{k=p+1}^N [\gamma u^*(k-r) + w(k)]^2$$

and therefore, using the fact that  $\{u^*(k)\}$  is defined as equation (5):

$$\begin{aligned} R_0^* - R_1^* &= \sum_{k=p+1}^N [\gamma u^*(k-r) + w(k) - w(k)][\gamma u^*(k-r) + w(k) + w(k)] \\ &= \sum_{k=p+1}^N \gamma^2 u^{*2}(k-r) + \sum_{k=p+1}^N 2[\gamma u^*(k-r)w(k)] \approx \gamma^2 \sum_{i=1}^n A_i^2 \end{aligned}$$

where we approximate the random variable  $R_0^* - R_1^*$  by its mean  $\gamma^2 \sum_{i=1}^n A_i^2$ , which is warranted by the fact that its mean is much larger than its standard deviation. Now, from equation (3) with  $T = N - p$ :

$$\mathbb{E}(g^*) = \frac{N - 3p - 1}{p} \mathbb{E}(R_0^* - R_1^*) \mathbb{E}\left(\frac{1}{R_1^*}\right) \quad (11)$$

Since  $w(k) \sim n(0, \sigma^2)$  it follows that  $\frac{1}{\sigma^2} R_1^* \sim \chi_{(N-p)}^2$ , and  $\sigma^2 \frac{1}{R_1^*} \sim \Gamma^{-1}(v_1, v_2)$ , where  $v_1 = \frac{N-p}{2}$  and  $v_2 = \frac{1}{2}$ . Since  $\mathbb{E}[\Gamma^{-1}(v_1, v_2)] = \frac{v_2}{v_1 - 1}$  (eg. [13]), we have  $\mathbb{E}\left(\frac{1}{R_1^*}\right) = \frac{1}{\sigma^2(N-p-2)}$ , which substituted in equation (11) gives equation (6), completing the proof.

# Bibliography

- [1] R. Agrawal, T. Imielinski, and A. Swami. Database Mining: A Performance Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5(6):914–925, December 1993.
- [2] R. Agrawal, K.-I. Lin, H. P. Sawhney, and K. Shim. Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases. In *Proceedings of the 21st VLDB Conference*, Zurich, Switzerland, 1995.
- [3] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the 11th International Conference on Data Engineering*, pages 3–14, 1995.
- [4] M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Applications*. Prentice Hall, 1993.
- [5] C. Bettini, S. Jajodia, and X. S. Wang. *Time Granularities in Databases, Data Mining and Temporal Reasoning*. Springer-Verlag, Berlin, 2000.
- [6] H. Boudjellaba, J.-M. Dufour, and R. Roy. Testing Causality Between Two Vectors in Multivariate Autoregressive Moving Average Models. *Journal of the American Statistical Association*, 87(420):1082–1090, 1992.
- [7] G. E. P. Box and G. M. Jenkins. *Time Series Analysis, Forecasting and Control*. Holden Day Series, 1976.
- [8] J. B. D. Cabrera, L. Lewis, X. Qin, W. Lee, and R. K. Mehra. Proactive Intrusion Detection of Distributed Denial of Service Attacks - A Case Study in Security Management. *Journal of Network and Systems Management*, June 2002. In Press.
- [9] J. B. D. Cabrera, L. Lewis, X. Qin, W. Lee, R. K. Prasanth, B. Ravichandran, and R. K. Mehra. Proactive Detection of Distributed Denial of Service Attacks using MIB Traffic Variables - A Feasibility Study. In *Proceedings of the Seventh IFIP/IEEE International Symposium on Integrated Network Management*, pages 609–622, Seattle, WA, May 2001.
- [10] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, Belmont, CA, 1990.
- [11] P. J. Criscuolo. Distributed Denial of Service - Trin00, Tribe Flood Network, Tribe Flood Network 2000, and Stacheldraht. Technical Report CIAC-2319, Department of Energy - CIAC (Computer Incident Advisory Capability), February 2000.
- [12] G. Das, K.-I. Lin, H. Mannila, G. Renganathan, and P. Smyth. Rule discovery from time series. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 16–22, 1998.
- [13] M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. John Wiley and Sons, Inc., New York, Second edition, 1993.
- [14] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 34:424–438, 1969.

- [15] J. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.
- [16] F. Höppner. Discovery of Temporal Patterns - Learning Rules about the Qualitative Behaviour of Time Series. In *Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Data Bases*, September 2001.
- [17] T. Kailath. *Linear Systems*. Prentice-Hall, Inc., 1980.
- [18] L. Ljung. *System Identification - Theory for the User*. Prentice Hall, Second edition, 1999.
- [19] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3):259–289, 1997.
- [20] J. J. McGuire, P. F. Ihmlé, and T. H. Jordan. Time-Domain Observations of a Slow Precursor to the 1994 Romanche Transform Earthquake. *Science*, 274, October 1996. Issue of October 4, 1996.
- [21] C. Silverstein, S. Brin, R. Motwani, and J .D. Ullman. Scalable techniques for mining causal structures. In A. Gupta, O. Shmueli, and J. Widom, editors, *VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases*, pages 594–605, New York City, NY, USA, August 1998. Morgan Kaufmann.
- [22] W. R. Stevens. *TCP/IP Illustrated, Volume 1: The Protocols*. Addison-Wesley, 1994.
- [23] N. Vandewalle, M. Ausloos, P. Boveroux, and A. Minguet. Visualizing the log-periodic pattern before crashes. *The European Physical Journal B*, 9:355–359, 1999.