A STATISTICAL FRAMEWORK FOR ANALYZING CYBER ATTACKS

APPROVED BY SUPERVISING COMMITTEE:

Shouhuai Xu, Ph.D.

Hugh Maynard, Ph.D.

Kay A. Robbins, Ph.D.

Ravi Sandhu, Ph.D.

Maochao Xu, Ph.D.

Accepted:

Dean, Graduate School

DEDICATION

I lovingly dedicate this dissertation to my family...

in memoriam of my father, who will always be with me.

to my mother, for your selfless support and incitement.

to my two sisters, my best friends.

A STATISTICAL FRAMEWORK FOR ANALYZING CYBER ATTACKS

by

ZHENXIN ZHAN, M.Sc.

DISSERTATION
Presented to the Graduate Faculty of
The University of Texas at San Antonio
In Partial Fulfillment
Of the Requirements
For the Degree of

DOCTOR OF PHILOSOPH YIN COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT SAN ANTONIO
College of Sciences
Department of Computer Science
May 2014

ACKNOWLEDGEMENTS

First, I would like to take the opportunity to acknowledge and extend my deepest gratitude to my advisor, Dr. Shouhuai Xu. Without his guidance, this research could not be accomplished. He has guided me and also challenged me during this journey. His enthusiasm, insights and patience has helped me learn not only how to approach a research problem but also taught me how to be a better person in life.

Second, I would like to thank my co-advisor Dr. Maochao Xu. His expertise and insights have helped me work out statistical details of this dissertation.

Third, I would like to thank my dissertation committee members Dr. Hugh Maynard, Dr. Kay A. Robbins and Dr. Ravi Sandhu for their comments.

The research described in the dissertation was partly supported by Dr. Shouhuai Xu's ARO Grant # W911NF-13-1-0141. The studies were approved by IRB. We also thank CAIDA for providing us a blackhole-collected dataset that is analyzed in the disseration.

May 2014

A STATISTICAL FRAMEWORK FOR ANALYZING CYBER ATTACKS

Zhenxin Zhan, Ph.D.
The University of Texas at San Antonio, 2014

Supervising Professor: Shouhuai Xu, Ph.D.

Data-driven cyber security analytics is one important approach to understanding cyber attacks. Despite its importance, there are essentially no systematic studies on characterizing the statistical properties of cyber attacks. The present dissertation introduces a systematic statistical framework for analyzing cyber attack data. It also presents three specific results that are obtained by applying the framework to analyze some honeypot- and blackhole-captured cyber attack data, while noting that the framework is equally applicable to other data that may contain richer attack information. The first result is that honeypot-captured cyber attacks often exhibit Long-Range Dependence (LRD). The second result is that honeypot-captured cyber attacks can exhibit Extreme Values (EV). The third result describes spatial and temporal characterizations that are exhibited by blackhole-captured cyber attacks. The dissertation shows that by exploiting the statistical properties exhibited by cyber attack data, it is possible to achieve certain "gray-box" predictions with high accuracy. Such prediction capability can be exploited to guide the proactive allocation of resources for effective defense.

TABLE OF CONTENTS

Vita

LIST OF TABLES

LIST OF FIGURES

Chapter 1: INTRODUCTION

1.1 Research Motivation: Understanding and Characterizing Cyber Attacks

Data-driven analysis is an important approach to understanding cyber attacks. While the importance is well recognized, there are not many studies in this sub-field perhaps because it demands both real data and advanced statistical skills. This dissertation studies cybersecurity phenomena, which are manifested by cyber attack data that can be captured by passive network instruments such as honeypots and blackholes. Honeypots aim to monitor attacks by emulating (to certain degrees) vulnerable network services [55]. Blackholes (also known as network telescopes, darkspaces, Internet sinks) aim to monitor attacks without setting up any network services and without interacting with the attackers. Since honeypots need to emulate network services while blackholes do not, blackholes often have much larger IP address spaces. For example, CAIDA's blackhole (`www.caida.org`) is a /8 network, or 0.4% of the entire Internet IP v4 address space.

1.2 Framework

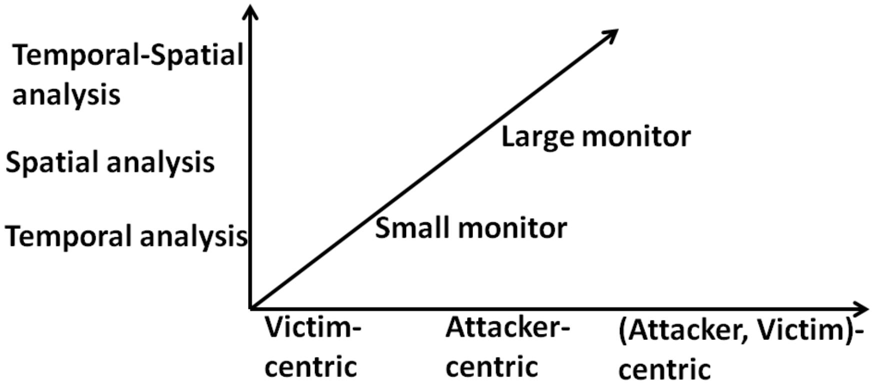


Figure 1.1: Overview of problem space.

Figure 1.1 hilights the problem space. There are possible 18 combinations of scenarios, and the dissertation covers some of them. For honeypot-captured data, we aim to identify statistical properties and exploit the properties for prediction. For blackhole data, we aim to infer the global cyber security posture.

Specifically, the statistical framework is characterized as follows. The framework is centered on the concept we call Stochastic Cyber Attack Processes, a new kind of mathematical objects that can be instantiated at multiple resolutions. This abstraction can naturally represent cyber attacks. Empowered by this abstraction, the framework has three orthogonal perspectives:

- **Identifying Statistical Properties Exhibited by Cyber Attack Data:** What statistical properties do the attack processes possess? For example, honeypot-captured attack data can exhibit the Long-Range Dependence (LRD) property and Extreme Value phenomenon. LRD property is also known as long-memory and is in sharp contrast to the memoryless of Poisson processes. This is for the first time that LRD is found to be relevant in the cybersecurity domain, despite that it has been known for two decades to be relevant in the benign traffic domain (where no attacks present) [37, 38, 72].

- **Explaining Cause of the Properties (if possible):** Why do the attack processes have those properties? Answering this question not only will deepen our understanding of mathematical nature of cyber attacks, but also will lead to effective exploitations of the properties — especially for prediction. Causes of the properties can be mysterious, but are important to know. For example, we have found evidences supporting that the cause of LRD in the cybersecurity domain is probably different from the cause of LRD in the benign-traffic domain.

- **Exploiting the Properties for Better Prediction:** How can we predict attacks hours (or even days) ahead of time? In contrast to the folklore that cyber attacks are *not* predictable, our preliminary analysis already shows how we can exploit the properties of the attack processes to predict the number of incoming attacks hours ahead of time. This would give the defenders sufficient early-warning time for proactively allocating defense resources. The prediction power can be attributed to the *gray-box* prediction methods, which explicitly accommodates the relevant statistical properties. This is in sharp contrast to the practice of *black-box* predictions.

1.3 Data sources

Stochastic cyber attack processes are manifested by various kinds of data. For studying them, we have identified the following data sources:

- Honeypot-captured cyber attack data: The data is from a honeypot infrastructure. For various (especially, legal) reasons, the infrastructure is based on low-interaction honeypots, including Dionaea [2], Mwcollector [3], Amun [1], and Nepenthes [11]. Basically, low-interaction honeypots simulate services and passively wait for connections from compromised computers in the wild. The term “low-interaction" refers to that the simulated service does not cover the entire protocol software stack, which means that the captured/collected data may not be sufficient for precisely pinning down the specific attacks.

  Each honeypot IP address was assigned to one of these programs and was completely isolated from the other honeypot IP addresses. To save resources, a single honeypot computer was assigned with multiple IP addresses and thus ran multiple honeypot software programs. A dedicated computer was used to collect the raw network traffic as `pcap` files, which are timestamped at the resolution of microsecond. The vulnerable services offered by all four honeypot programs are SMB, NetBIOS, HTTP, MySQL and SSH, each of which is associated to a unique TCP port. This means that each IP address (i.e., honeypot software) opens the ports corresponding to these services. We call these ports *production ports*, and the other ports *non-production ports* (because they are associated to no services). The concrete attacks targeting the production ports can be dependent upon the specific vulnerabilities emulated by the honeypot programs (e.g., the Microsoft Windows Server Service Buffer Overflow MS06040 and Workstation Service Vulnerability MS06070 for the SMB service).

- Blackhole-captured cyber attack data: This kind of data is collected at large blackholes, which are routable IP spaces but with no services. We have got access to some data from CAIDA. Blackhole-captured data is complementary to honeypot-captured data because on the positive side, blackhole can be much larger (e.g., /8) than honeypot, and on the downside, blackhole-captured data does not have any interaction information with the attackers.

Chapter 2: ANALYZING LONG-RANGE DEPENDENCE EXHIBITED BY CYBER ATTACKS

2.1 Introduction

Characterizing cyber attacks not only can deepen our understanding of cyber threats but also can lead to important implications for effective cyber defense. Honeypot is an important tool for collecting cyber attack data, which can be seen as a “birthmark" of the cyber threat landscape as observed from a certain IP address space. Therefore, studying this kind of data allows us to extract useful information about cyber attacks/threats. However, this perspective of cyber security has not been understood well, perhaps because it requires both real data and fairly involved statistical techniques. Motivated by the need to better understand cyber attacks, this chapter initiates the study of rigorous statistical properties of cyber attacks as exhibited by honeypot-collected data.

2.1.1 Our Contributions

In this chapter, we aim to rigourously study the statistical properties of cyber attacks as exhibited by honeypot-collected data. We make two contributions. First, we propose a framework for identifying, characterizing and exploring statistical properties of honeypot-captured cyber attacks. The framework is centered on a new concept we call *stochastic cyber attack processes*, which are a new kind of mathematical objects for modeling cyber attacks. The framework is geared towards answering the following questions: (i) What statistical properties do the stochastic cyber attack processes exhibit? (ii) What are the implications of these properties? (iii) What is the cause of these properties?

Second, we conduct a case study by applying the framework to analyze a dataset that is collected by a honeypot of 166 IP addresses for five periods of time (220 days cumulative). Findings of the case study are: (i) Stochastic cyber attack processes are not Poisson. Instead, they might exhibit Long-Range Dependence (LRD) — a property that is not known to be relevant in the cyber security domain until now. (ii) LRD can be exploited to better predict the incoming attacks. This shows the power of “gray-box" (rather than “black-box") prediction. (iii) The cause of LRD in cyber security domain might be fundamentally different from the case of LRD in the setting of benign traffic.

We plan to make our statistical analysis software code publicly available so that other researchers can use it to analyze their data of the same or similar kind.

2.1.2 Related Work

In the literature, honeypot-collected cyber attack data has been studied from the following perspectives: analyzing honeypot-observed probing activities [39], characterizing/grouping attacks [8–10, 20, 21, 42, 53, 54, 67], and identifying methods to detect attacks such as DoS (denial-of-service) [28], scans [34], worms [25,26], and botnets [41,63]. These perspectives are different from ours because we study statistical properties that can be exploited to better predict cyber attacks. On the other hand, LRD was first observed in benign traffic about two decades ago [37,38,61,72]. To our knowledge, LRD is not known to be relevant in the cyber security domain until now.

There have been studies on characterizing blackhole-collected traffic data (see, for example, [51, 73]) or one-way traffic in live networks [31]. Our study is different from theirs because of the following. First, honeypot-collected data includes two-way communications; whereas blackhole-collected data mainly corresponds to one-way communications. Second, we rigorously explore statistical properties such as LRD; whereas their studies do not pursue such characteristics. Nevertheless, we believe that our analysis framework can be adapted to analyze blackhole-collected data as well.

The rest of the chapter is organized as follows. Section 2.2 presents the framework for analyzing honeypot-collected attack data. Section 3.2 briefly reviews some statistical preliminaries, while some detailed statistical techniques are deferred to the Appendix. Section 2.4 describes a case study by applying the framework to analyze a specific dataset. Section 4.6 concludes the chapter with future research directions.

2.2 Statistical Sub-Framework for Characterizing Honeypot-Captured Cyber Attacks

The framework is centered on the new concept of *stochastic cyber attack processes* (or *attack processes* for short).

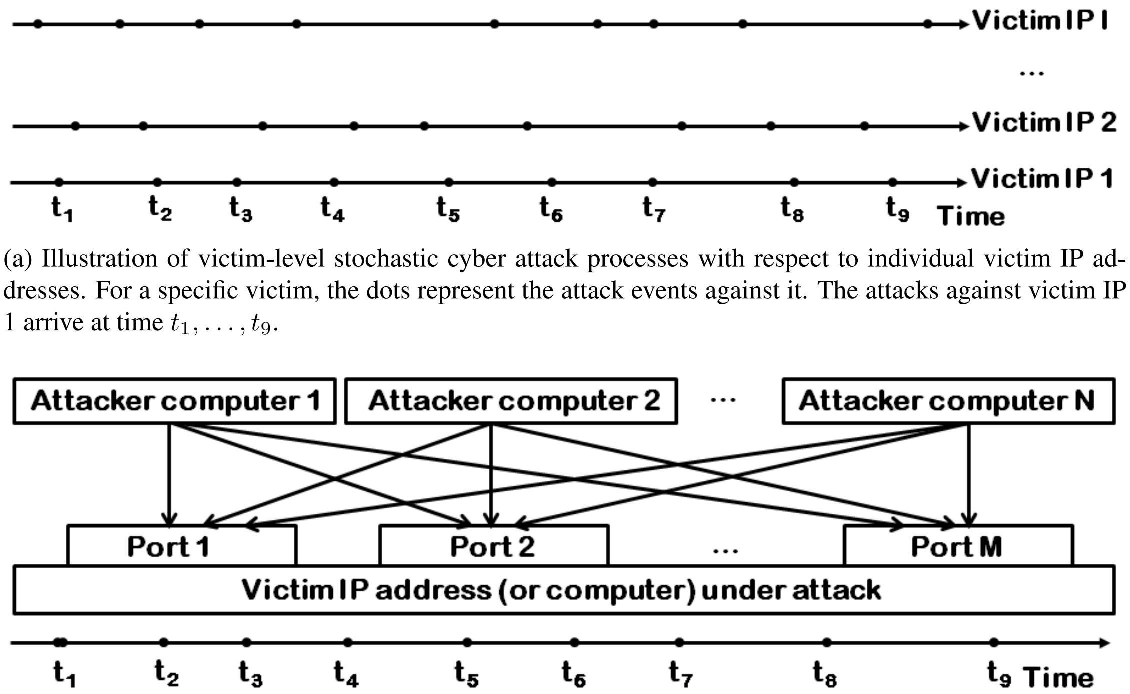
2.2.1 The Concept of Stochastic Cyber Attack Processes

Cyber attacks can be naturally modeled as *stochastic cyber attack processes*, which are in principle Point Processes [24]. Stochastic cyber attack processes can be instantiated at multiple levels (or resolutions). Network-level attack processes model cyber attacks against a network of interest; victim-level attack processes model cyber attacks against individual victim computers or IP addresses; port-level attack processes model cyber attacks against individual ports of a victim computer; attacker-level attack processes model cyber attacks launched by distinct attackers against a victim computer. Further, port-level attack processes can be defined with respect to the *production ports* that are associated to some services, or with respect to the *non-production ports* that are not associated to any services. The distinction of model resolution is important because a high-level (i.e., low-resolution) attack process may be seen as the superposition of multiple low-level (i.e., high-resolution) attack processes, which may help explain the cause of a particular property exhibited by the high-level processes.

For example, Figure 2.1a illustrates the attacks against individual victim IP addresses, where the dots on the time line formulate a victim-level attack process. Figure 2.1b further shows that a victim is attacked by N attackers (or attacking computers) at some ports and the attacks arrive at time $t_1, \ldots, t_9$.

2.2.2 The Statistical Analysis Sub-Framework

The framework consists of 5 steps and is geared toward answering the afore-mentioned three questions.



(a) Illustration of victim-level stochastic cyber attack processes with respect to individual victim IP addresses. For a specific victim, the dots represent the attack events against it. The attacks against victim IP 1 arrive at time $t_1, \ldots, t_9$.

(b) Elaboration of a victim-level attack process where the attacks arrive at time $t_1, \ldots, t_9$.

Figure 2.1: Illustration of victim-level stochastic cyber attack processes

Step 1: Data pre-processing Honeypot-collected cyber attack data is often organized according to the honeypot IP addresses. Because the data involves two-way communications between the honeypot and the remote attackers, we need a pre-processing procedure to take care of two issues. First, we may need to differentiate the traffic corresponding to the *production ports* that are associated to some honeypot programs/services, and the traffic corresponding to the *non-production ports* that are not associated to any services.

Second, we need to treat each incoming TCP flow as an attack, especially when the honeypot-emulated vulnerable services are based on TCP. A TCP flow can be uniquely identified from honeypot-collected raw `pcap` data via the attacker's IP address, the port used by the attacker, the victim IP address in the honeypot, and the port that is under attack. An unfinished TCP handshake can also be treated as a flow (attack) because an unsuccessful handshake can be caused by events such as: the port in question is busy (i.e., the connection is dropped). For flows that do not end with the FIN flag (which would indicate safe termination of TCP connection) or the RST flag (which would indicate unnatural termination of TCP connection), we need to choose two parameters in the pre-process. One parameter is the *flow timeout time*, meaning that a flow is considered expired when no packet of the flow is received during a time window (e.g., 60 seconds would be reasonable for low-interaction honeypots that provide limited interactions [7], but a longer time may be needed for high-interaction honeypots). The other parameter is the *flow lifetime*, meaning that a flow is considered expired when a flow lives longer than a pre-determined lifetime, which can be set as 300 seconds for low-interaction honeypots [7] but a longer time may be needed for high-interaction honeypots.

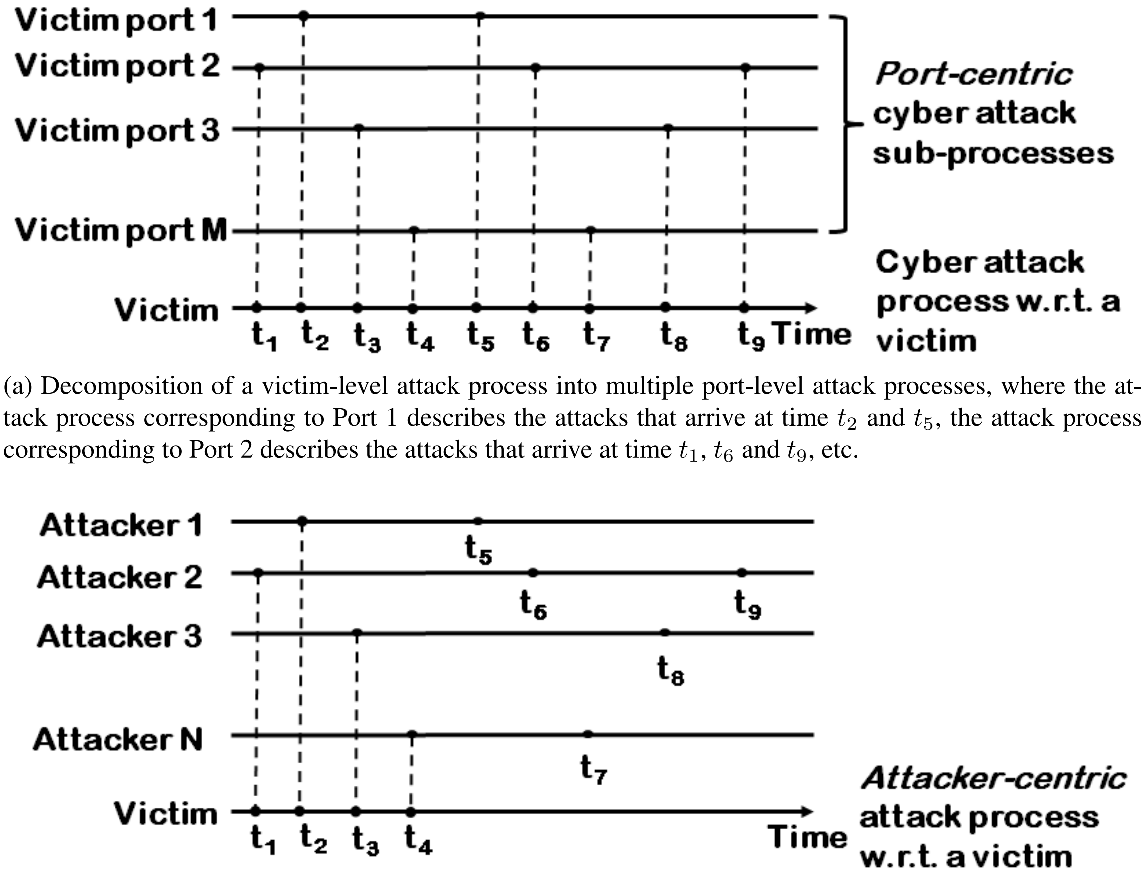
Step 2: Basic statistical analysis The basic statistics of cyber attack data can offer hints for advanced statistical analysis. For stochastic cyber attack processes, the primary statistic is the *attack rate*, which describes the number of attacks that arrive at unit time (e.g., minute or hour or day). The secondary statistic is the *attack inter-arrival time*, which describes the time intervals between two consecutive attack events. By investigating the min, $mean$, $median$, $variance$ and max of these statistics, we can identify outliers and obtain hints about the properties of the attack processes. For example, if the attack events are bursty, an attack process may not be Poisson, which can serve as a hint for further advanced statistical analysis.

Step 3: Advanced statistical analysis: Identifying statistical properties of attack processes This step is to identify statistical properties of attack processes at resolutions of interest. A particular question that should be asked is: Are the attack processes Poisson? If not Poisson, what properties do they exhibit? It would be ideal that the attack processes are Poisson because we can easily characterize Poisson processes with very few parameters, and because there are many mature methods and techniques for analyzing them. For example, we can use the property — the superposition of Poisson processes is still a Poisson process [27] — to simplify problems when we consider attack processes at multiple resolutions/levels. In many cases, attack processes may not be Poisson. For characterizing such processes, we need to use advanced statistical methods, such as Markov process, Lévy process, and time-series methods [24,52,60]. This step is crucial because identifying advanced statistical properties of attack processes can pave the way for answering the next questions.

Step 4: Exploiting the statistical properties This step addresses the following question: How can we exploit the statistical properties of stochastic cyber attack processes to do useful things? One exploitation is to predict the incoming attacks in terms of attack rate. This is so because if the processes exhibit a certain property (e.g., Long-Range Dependence [61, 71] or Short-Range Dependence [15, 52, 69]), the prediction model should accommodate the property in order to achieve better predictions. We note that although prediction is geared toward honeypot-oriented traffic, it can be useful for defending production networks as well. This is true because when honeypot-captured attacks are increasing (or decreasing), the attack rate with respect to production networks might also be increasing (or decreasing) as long as the honeypots are approximately uniformly deployed at all IP address space. Moreover, it is possible to characterize the relations between the attack traffic with respect to a honeypot and the attack traffic with respect to a production network. Although many honeypots are currently deployed at consecutive IP addresses (including the dataset we use for case study), it is doable in practice to blend honeypot IP addresses into production networks. Since being able to predict incoming attacks (especially hours ahead of time) is always appealing, this would give incentives to deploy honeypot as such, or to study the relations between the attack traffic against honeypots and the attack traffic against production networks.

Step 5: Exploring cause of the statistical properties This step aims to address the following question: What caused the statistical properties of stochastic cyber attack processes? This question is interesting because it reflects a kind of “natural" phenomenon in cyberspace. In order to answer the question, we propose to study two approaches. One approach is to study the decomposed lower-level (i.e., higher-resolution) stochastic cyber attack processes. For example, in order to investigate whether or not a certain property is caused by another certain property of the low-level (i.e., high-resolution) processes, we can decompose a victim-level attack process into port-level attack processes that correspond to the individual ports of the victim. This is illustrated in Figure

2.2a, where the victim-level attack process is decomposed into M port-level attack processes.


Victim port 1
Victim port 2
Victim port 3
Victim port M
Victim
t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 Time
Port-centric cyber attack sub-processes
Cyber attack process w.r.t. a victim
Attacker 1
Attacker 2
Attacker 3
Attacker N
Victim
t_1 t_2 t_3 t_4
t_5 t_6 t_9 t_8 t_7
Time
Attacker-centric attack process w.r.t. a victim

(a) Decomposition of a victim-level attack process into multiple port-level attack processes, where the attack process corresponding to Port 1 describes the attacks that arrive at time t_2 and t_5, the attack process corresponding to Port 2 describes the attacks that arrive at time t_1, t_6 and t_9, etc.

(b) Attacker-level attack process can be derived from victim-level attack process by ignoring the subsequent attacks launched by the same attacker. In this example, the attacker-level attack process corresponding to the victim describes the attacks that arrive at time t_1, t_2, t_3, t_4.

Figure 2.2: Two approaches to exploring causes of statistical properties

The other approach is to investigate whether or not a certain property is caused by the intense (consecutive) attacks that are launched by individual attackers. For this purpose, we can consider the attacks against each victim that are launched by *distinct* attackers. As illustrated in Figure 2.2b, even though an attacker launched multiple consecutive attacks against a victim, we only need to consider the first attack. If the attacker-level attack processes do *not* exhibit the property that is exhibited by the victim-level attack processes, we can conclude that the property is caused by the intensity of the attacks that are launched by individual attackers.

2.3 Statistical Preliminaries

2.3.1 Long-Range Dependence (LRD)

A stationary time sequence $\{X_i, i \geq 1\}$ is said to possess LRD [61, 71] if its autocorrelation function

$$\rho(h) = \mathrm{Cor}(X_i, X_{i+h}) \sim h^{-\beta} L(h), \quad h \to \infty, \tag{2.1}$$

for $0 < \beta < 1$, where $L(\cdot)$ is a slowly varying function meaning that $\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1$ for all $t > 0$. The degree of LRD is expressed by Hurst parameter (H), which is related to the parameter β in Eq. (2.1) as $\beta = 2 - 2H$. This means that for LRD, we have $1/2 < H < 1$ and the degree of LRD increases as $H \to 1$. In the Appendix, we briefly review six popular Hurst-estimation methods that are used in this chapter.

LRD can be caused by the following: non-stationarity [45], short-range dependent time series with change points in the mean, slowly varying trends with random noise, stationary parametric time series with time-varying parameters [56,62]. These are called "spurious LRD" and are not the focus of the present study. We will remove spurious LRD processes by testing the null hypothesis that a given time series is a stationary LRD process against the alternative hypothesis that it is affected by change points or a smoothly varying trend [56]. Specifically, one test is:

$$H_0\text{: } X_t \text{ is stationary with LRD}$$

vs

$$H_a\text{: } X_t = Z_t + \mu_t \text{ with } \mu_t = \mu_{t-1} + \psi_t \eta_t$$

where Z_t is a stationary short-memory process, η_t is a white noise process and ψ_t is a Bernoulli random variable which takes value 1 with probability p_n. The other alternative is:

$$H_a\text{: } X_t = Z_t + h(t/n),$$

where $h(\cdot)$ is a Lipschitz continuous function on $[0, 1]$.

2.3.2 Two Statistical Models for Predicting Incoming Attacks

We call a model *LRD-less* if it cannot accommodate LRD and *LRD-aware* if it can accommodate LRD. Let ϵ_t be independent and identical normal random variables with mean 0 and variance σ_ϵ^2. We consider two popular models.

- LRD-less model ARMA(p, q): This is the autoregressive moving average process of orders p and q with
$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}.$$
It is one of the most popular models in time series [23].

- LRD-aware model FARIMA(p, d, q): This is the well-known Fractional ARIMA model where $0 < d < 1/2$ and $H = d + 1/2$ $[5, 15, 71]$. Specifically, a stationary process X_t is called FARIMA(p, d, q) if
$$\phi(B)(1 - B)^d X_t = \psi(B)\epsilon_t,$$
for some $-1/2 < d < 1/2$, where
$$\phi(x) = 1 - \sum_{j=1}^{p} \phi_j x^j \quad \text{and} \quad \psi(x) = 1 + \sum_{j=1}^{q} \psi_j x^j,$$
B is the back shift operator defined by $BX_t = X_{t-1}$, $B^2 X_t = X_{t-2}$, and so on.

2.3.3 Measurement of Prediction Accuracy

Suppose $X_m, X_{m+1}, \ldots, X_z$ are observed data (all non-negative), and $Y_m, Y_{m+1}, \ldots, Y_z$ are the predicted data. We can define prediction error $e_t = X_t - Y_t$ for $m \leq t \leq z$. Recall the popular statistic PMAD (Percent Mean Absolute Deviation):

$$\mathsf{PMAD} = \frac{\sum_{t=m}^{z} |e_t|}{\sum_{t=m}^{z} X_t},$$

which can be seen as the overall prediction error. We also define a variant of it, called *underestimation error*, which counts only the underestimations as follows:

$$\mathsf{PMAD}' = \frac{\sum_{t=m}^{z} e_t}{\sum_{t=m}^{z} X_t} \ \ for\ e_t > 0\ and\ corresponding\ X_t.$$

This is relevant because if the defender is willing to over-provision some defense resources, the predicted results are perhaps more useful because underestimation error corresponds to the attacks that can be overlooked due to insufficient defense resources.

It is also convenient to use the following measurement of *overall accuracy* (OA for short):

$$\mathsf{OA} = 1 - \mathsf{PMAD}.$$

Correspondingly, we can define the following measurement of *underestimation accuracy* (UA for short):

$$\mathsf{UA} \ = \ 1 - \mathsf{PMAD}'.$$

2.4 Applying the Sub-Framework to Analyze Some Real Data

In order to demonstrate the usefulness of the framework, we now conduct a case study by applying it to analyze a specific dataset. The framework and analysis can be applied to other datasets of the same or similar kind.

2.4.1 Step 1: Data Pre-Processing

We use hoenypot dataset as described in section 1.3. Table 2.1 summarizes the dataset, which corresponds to 166 victim/honeypot IP addresses for five periods of time. These periods are not strictly consecutive because of network/system maintenance etc.

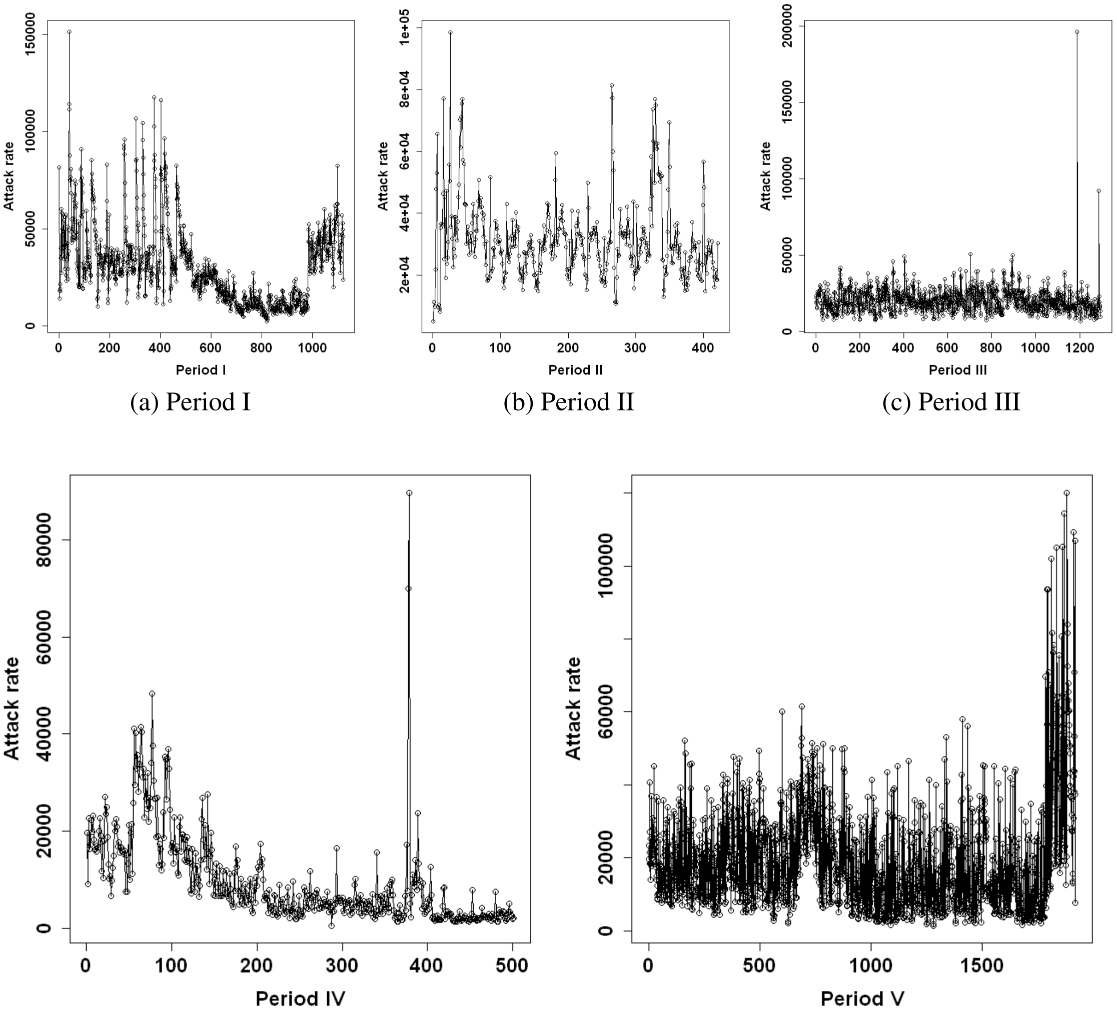
Table 2.1: Data description

Period	Dates	Duration (days)	# victim IPs
I	11/04/2010 - 12/21/2010	47	166
II	02/09/2011 - 02/27/2011	18	166
III	03/12/2011 - 05/06/2011	54	166
IV	05/09/2011 - 05/30/2011	21	166
V	06/22/2011 - 09/12/2011	80	166

In our pre-processing, we resolve the two issues described in the pre-processing step of the framework as follows. First, we disregard the attacks against the non-production ports because such TCP connections are often dropped. The vulnerable services offered by all four honeypot programs are SMB, NetBIOS, HTTP, MySQL and SSH, each of which is associated to a unique TCP port. These are the production ports. Note that the specific attacks against the production ports are dependent upon the vulnerabilities emulated by the honeypot programs (e.g., Microsoft Windows Server Service Buffer Overflow MS06040 and Workstation Service Vulnerability MS06070 for the SMB service). Since low-interaction honeypots do not capture sufficient information for precisely recognizing the specific attacks, we do not look into specific attack types. Second, for flows that do not end with the FIN flag (indicating safe termination of TCP connection) or the RST flag (indicating unnatural termination of TCP connection), we use the following two parameter values: 60 seconds for the *flow timeout time* and 300 seconds for the *flow lifetime*.

2.4.2 Step 2: Basic Statistical Analysis

We consider the *per-hour* attack rate with respect to the honeypot network, with respect to each individual victim IP address, and with respect to each production port of each individual victim. The choice of *per-hour* is natural, while noting that *per-day* attack rate is not appropriate because



(a) Period I

(b) Period II

(c) Period III

(d) Period IV

(e) Period V

Figure 2.3: Time series plots of the network-level attack processes corresponding to the five periods. The x-axis indicates the relative time with respect to the start time for each period (unit: hour). The y-axis indicates the number of attacks (per hour) arriving at the honeypot network during a period.

each period is no more than 80 days. Because the numbers of victim-level and port-level attack processes are substantially larger than the number of network-level attack processes, different methods are used to represent their basic statistics.

Basic statistics of network-level attack processes For network-level attack processes, it is feasible and appropriate to plot the time series of the attack rate (per hour), namely the total number of attacks against the honeypot network of 166 victims.

Figure 2.3 plots the time series of attacks. We make the following observations. First, the five periods exhibit different attack patterns. For example, Periods I, II and V are relatively stationary. Second, there are some extremely intense attacks during some hours in Periods III and IV. The specific hour corresponding to the extreme value in Period III is Apr 01, 2011, 12 Noon (US Eastern Time); the attacks are against the SSH services. It is evident that the attacks are brute-forcing password. The peak of attacks during Period IV occurs at May 16, 2011, 3 AM (US Eastern Time). The intense attacks are against the HTTP service. We find no information from the Internet whether or not the peaks correspond to (for example) outbreaks of some worm or botnet. Third, although the five plots exhibit some change-points, a formal statistical analysis (using the method for removing spurious LRD, which is reviewed in Section 3.2) shows that there are some change-points only in Period III, which correspond to the largest attack rate. This means that visual observations can be misleading.

Table 2.2 describes the basic statistics of the attack rate as exhibited by the network-level attack processes. We observe that on average, the victim network is least intensively attacked during Period IV because the average per-hour attack rate is about 9861, which is substantially smaller than the average attack rate during the other periods. We observe that the variances of attack rates are much larger than the corresponding mean attack rates of the network-level attack processes. This hints that these processes are not Poisson. As we will see in Section 2.4.3, these processes actually exhibit LRD instead.

Table 2.2: Basic statistics of network-level attack processes corresponding to the five periods of time.

Period	MIN	Mean	Median	Variance	Max
I	2572	30963.2	28263	401243263.2	151189
II	5155	31576.8	29594	167872819.0	98527
III	6732	20382.3	19579	72436071.5	196210
IV	637	9861.1	6528	93209085.3	89718
V	1417	18960.2	15248.5	205276388.4	120221

Basic statistics of victim-level attack processes For victim-level attack processes, we consider the attack rate or the number of attacks (per hour) arriving at a victim. Since there are 166 victims in each period, we cannot afford to plot time series of victim-level attack processes.

Table 2.3: Basic statistics of victim-level attack processes: attack rate (per hour). For a specific period and a specific statistic $X \in \{\mathsf{Mean}, \mathsf{Median}, \mathsf{Variance}, \mathsf{MAX}\}$, LB (UB) stands for the lower-bound or minimum (upper-bound or maximum) of statistic X among all the victims and all the hours. In other words, the LB and UB values represent the minimum and maximum per-hour attack rate observed during an entire period and among all the victims.

Period	Mean(·)		Median(·)		Variance(·)		MAX(·)	
	LB	UB	LB	UB	LB	UB	LB	UB
I	32.1	1810.4	8	1327	1589.9	3219758.8	247	14403
II	49.8	1412.0	43	1112	1466.5	1553585.6	335	10995
III	11.5	1513.5	3	1490	254.0	676860.7	125	5287
IV	3.5	1663.4	1	1184	29.7	2808045.2	41	7793
V	34.0	2228.8	8.5	1526.5	1225.6	4639659.1	274	12267

Table 2.3 summaries the observed lower-bound (minimum) and upper-bound (maximum) values of per-hour attack rate for each statistic among the 166 victims. By taking Period I as an example, we observe the following. The average per-hour attack rate (among all the victims and among all the hours) is some number between 32 and 1810 attacks per hour; the median per-hour attack rate is some number between 8 and 1327 attacks per hour; the maximum number of attacks against a single victim can be up to 14403. Boxplots of the four statistics, which are not included for the sake of saving space, show that the five periods exhibit somewhat similar (homogeneous) statistical properties. For example, each statistic has many outliers in each period. By looking into all individual victim-level attack processes, we find that among all the 830 victim-level attack

processes (166 victims/period × 5 periods = 830 victims), the variance of attack rate is at least 3.5 times greater than the mean attack rate corresponding to the same victim. This fact — the variance is much larger than the mean attack rate — hints that Poisson models may not be appropriate for describing victim-level attack processes. This suggests us to conduct formal statistical tests, which will be presented in Section 2.4.3.

Basic statistics of port-level attack processes For port-level attack processes, Table 2.4 summarizes the lower-bound (minimum value) and upper-bound (maximum value) for each statistic. By taking Period I as an example, we observe the following. There can be no attacks against some production ports during some hours, which explains why the Mean per-hour attack rate can be 0. On the other hand, a port (specifically, port 445 at Nov 6, 2010, 9 AM US Eastern time) can be attacked by 14363 attacks within one hour. Like what is observed from the victim-level attack processes, we observe that the variance of attack rate is much larger than the mean attack rate. This means that the port-level attack processes are not Poisson. Indeed, as we will see in Section 2.4.5, many port-level attack processes are actually heavy-tailed.

Table 2.4: Basic statistics of port-level attack processes: attack-rate (per hour). As in Table 2.3, LB and UB values represent the minimum and maximum per-hour attack rate observed during an entire period and among all production ports of the victims.

Period	Mean(·)		Median(·)		Variance(·)		MAX(·)	
	LB	UB	LB	UB	LB	UB	LB	UB
I	0	1740.7	0	1196	0	3249318.9	1	14363
II	0	1251.5	0	948	0	1545078.5	1	10992
III	0	1482.1	0	1458	0	661847.3	1	5275
IV	0	1613.4	0	1142	0	2588396.6	1	6961
V	0	2169.8	0	1448.5	0	4629744.3	1	12267

2.4.3 Step 3: Identifying Statistical Properties of Attack Processes

We now characterize the statistical properties exhibited by network-level and victim-level attack processes. In particular, we want to know they exhibit similar (if not exactly the same) or different properties. In the above, we are already hinted that the attack processes are not Poisson. In what follows we aim to pin down their properties.

Network-level attack processes exhibit LRD The hint that network-level attack processes are not Poisson suggests us to identify their properties. It turns out that the network-level attack processes exhibit LRD as demonstrated by their Hurst parameters. Table 2.5 describes the six kinds of Hurst parameters corresponding to the network-level attack processes. Although the Hurst parameters suggest that they all exhibit LRD, a further analysis shows the LRD exhibited in Period III is spurious because it was caused by the non-stationarity of the process. Therefore, 4 out of the 5 network-level attack processes exhibit (legitimate) LRD.

Table 2.5: The estimated Hurst parameters for network-level attack processes. The six estimation methods are reviewed in Appendix 6.1.1. Note that a Hurst parameter value being negative or being greater than 1 means that either the estimation method is not suitable or the attack process is non-stationary.

Period	RS	AGV	Peng	Per	Box	Wave	LRD?
I	0.80	0.95	0.88	1.03	1.00	0.75	Yes
II	0.74	0.59	0.86	0.75	0.97	0.84	Yes
III	0.74	0.52	0.65	0.63	0.63	0.65	No
IV	1.05	0.97	0.95	1.07	0.97	1.22	Yes
V	0.74	0.78	0.74	1.03	0.80	0.80	Yes

Victim-level attack processes exhibit LRD For the 830 (166 victims/period $\times 5$ periods =830) victim-level attack processes, we first rigorously show that they are not Poisson. Assume that the attack inter-arrival times are independent and identically distributed exponential random variables with distribution

$$F(x) = 1 - e^{-\lambda x}, \, \lambda > 0, x \geq 0.$$

To test the exponential distribution, we first estimate the unknown parameter λ by the maximum likelihood method. Then, we compute the Kolmogorov-Smirnov (KS), Cramér-von Mises (CM), and Anderson-Darling (AD) test statistics [32, 57] (cf. Appendix 6.1.3 for a review) and compare them against the respective critical values.

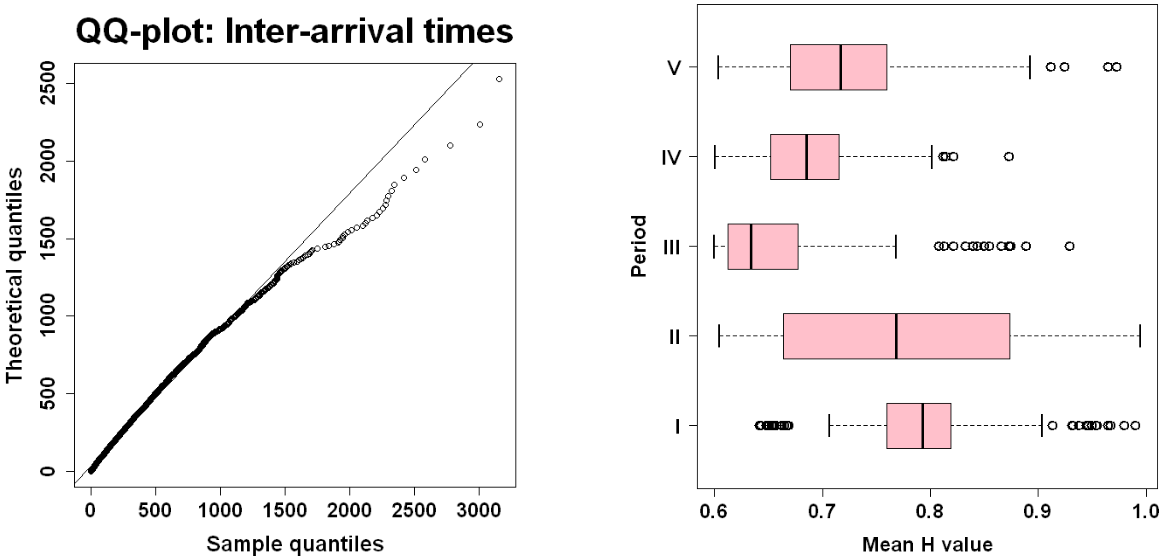
Table 2.6: Minimum values of the three test statistics for attack inter-arrival time (unit: second) corresponding to the victim-level attack processes, where min and max represent the minimal and maximal minimum values among all victim-level attack processes in a period, and Inf means the value is extremely large.

Period	KS		CM		AD	
(days)	min	max	min	max	min	max
I	0.13	0.54	482.30	59543.87	inf	inf
II	0.06	0.50	47.08	20437.82	298.73	inf
III	0.06	0.65	163.71	51434.32	1103.70	inf
IV	0.04	0.81	3.44	31376.27	22.83	inf
V	0.08	0.65	323.39	214543.54	inf	inf
CV	0.01		0.22		1.13	

Table 2.6 reports the minimum test statistics, where the critical values for the test statistics are based on significance level .05 and obtained from [18, 19]. Since the values are far from the critical values, there is no evidence to support the exponential distribution hypothesis. Because the minimum test statistics violate the exponential distribution assumption already, greater test statistics must violate the exponential distribution assumption as well.

We also use QQ-plot to evaluate the goodness-of-fit of exponential distributions for the attack inter-arrival time of victim-level attack processes that simultaneously exhibit the minimum test statistics in Table 2.6. This is the victim from Period IV with $H_{\mathsf{KS}} = 0.04$, $H_{\mathsf{CM}} = 3.44$ and $H_{\mathsf{AD}} = 22.83$. If the attack inter-arrival time corresponding to this particular victim does not exhibit the exponential distribution, we conclude that no attack inter-arrival time in this dataset exhibits the exponential distribution. The QQ plot is displayed in Figure 2.4a. We observe a large deviation in the tails. Hence, exponential distribution cannot be used as the distribution of attack inter-arrival times, meaning that all the victim-level attack processes are not Poisson.

Given that the victim-level attack processes are not Poisson, we suspect they might exhibit LRD as well. Figure 2.4b shows the boxplots of Hurst parameters of attack rate. We observe that Periods I and II have relatively large Hurst parameters, suggesting stronger LRD. Table 2.7 summarizes the minimums and maximums of the estimated Hurst parameters of attack rates. Consider Period I as an example, we observe that the attack processes corresponding to 163 (out of the 166) victims



(a) QQ-plot of inter-arrival time of victim-level attack process that exhibits the minimum KS, CM and AD value simultaneously

(b) Boxplot of Hurst parameters of attack rate of the victim-level attack processes corresponding to the 5 periods

Figure 2.4: Victim-level attack processes are not Poisson but exhibit LRD

Table 2.7: The estimated Hurst parameters for attack rate (per hour) of the victim-level attack processes. The six estimation methods are reviewed in Appendix 6.1.1. Note that a Hurst value being negative or being greater than 1 means that either the estimation method is not suitable or the process is non-stationary. The column “# of victims w/ $\bar{H} \in [.6, 1]$" represents the total number of victim-level attack processes whose average Hurst parameters $\in [.6, 1]$ (where average is among the six kinds of Hurst parameters), which suggests the presence of LRD. The column “# of victims w/ LRD" indicates the total number of victim-level attack processes that exhibit LRD rather than spurious LRD. (The same notations will be used in the description of Tables 2.8 and 2.13.)

Period	RS		AGV		Peng		Per		Box		Wave		# victims w/	# victims w/
	min	max	min	max	min	max	min	max	min	max	min	max	$\bar{H} \in [.6, 1]$	LRD
I	0.53	1.01	0.46	0.98	0.66	1.14	0.73	1.39	0.55	1.15	0.40	0.96	163	159
II	0.49	0.94	0.40	0.98	0.56	1.37	0.53	1.69	0.33	1.32	-0.55	1.33	130	116
III	0.65	0.95	0.30	0.96	0.53	1.06	0.44	1.22	0.43	0.98	0.33	1.02	93	87
IV	0.40	1.13	0.12	1.00	0.49	1.45	0.33	1.74	0.42	1.32	-0.34	1.47	126	125
V	0.52	1.01	0.14	0.99	0.45	1.22	0.47	1.43	0.57	1.30	-0.16	1.18	158	89

have average Hurst parameters falling into $[.6, 1]$ and thus suggest LRD, where the average is taken over the six kinds of Hurst parameters. However, only 159 (out of the 163) victim-level attack processes exhibit legitimate LRD because the other 4 (out of the 163) victim-level attack processes are actually spurious LRD (i.e., caused by the non-stationarity of the processes). We also observe that in Period III, there are only 87 victim-level attack processes that exhibit LRD. Overall, 70%

victim-level attack processes, or $159 + 116 + 87 + 125 + 89 = 576$ out of $166 \times 5 = 830$ attack processes, exhibit LRD.

Port-level attack processes exhibit LRD Table 2.8 summarizes the Hurst parameters of port-level attack processes. We observe that there are respectively 316, 397, 399, 328, 406 port-level attack processes that exhibit LRD. Since there are 5 production ports per victim and 166 victims, there are 830 port-level attack processes per period. Since there are 5 periods of time, there are 4150 port-level attack processes in total (830 ports/period × 5 periods=4150 ports). This means that $(316 + 397 + 399 + 328406)/4150 = 44.5\%$ port-level attack processes exhibit LRD.

Table 2.8: The estimated Hurst parameters for port-level attack rate (per hour) of the port-level attack processes.

Period	RS		AGV		Peng		Per		Box		Wave		total # of	# ports w/	# ports w/
	min	max	min	max	min	max	min	max	min	max	min	max	ports	$\bar{H} \in [.6, 1]$	LRD
I	0.41	1.01	-0.18	0.98	-0.15	1.23	0.38	1.55	0.39	1.48	-0.18	1.00	830 + 0	349	316
II	0.23	1.50	0.04	0.97	0.18	1.51	0.32	1.68	0.26	1.45	-0.60	1.38	829 + 1	419	397
III	0.14	1.01	-0.02	0.96	0.27	1.08	0.38	1.28	0.34	1.07	0.08	1.00	830 + 0	422	399
IV	0.25	1.17	0.05	1.00	0.24	1.57	0.18	1.70	0.29	1.50	-1.10	1.72	828 + 2	339	328
V	0.43	1.14	0.12	0.99	0.42	1.40	0.45	1.52	0.40	1.41	-1.07	1.43	830 + 0	528	406

Summary In summary, we observe that 80% (4 out of 5) network-level attack processes exhibit LRD, 70% victim-level attack processes exhibit LRD, and 44.5% port-level attack processes exhibit LRD. This means that defenders should expect that the burst of attacks will sustain, and that cyber attack processes should be modeled using LRD-aware stochastic processes.

2.4.4 Step 4: Exploiting LRD to Predict Attack Rates

Assuming that the attacks arriving at honeypots are representative of, or related to, the attacks arriving at production networks (perhaps in some non-trivial fashion that can be identified given sufficient data), being able to predict the number of incoming attacks hours ahead of time can give the defenders sufficient early-warning time to prepare for the arrival of attacks. Intuitively, the model that is good at prediction in this context should accommodate the LRD property. This is confirmed by our study described below.

Prediction algorithm Let $\{X_1, \ldots, X_n\}$ be the time series of observed attack rates. The basic idea of prediction is to use some portion of the observed data to build a model (training or model fitting), which is then used to predict the attack rates corresponding to the rest/future portion of the observed data. In order to build a reliable model, 50% of the observed data is used as the training data for building models. Let h be an input parameter indicating the number of steps (i.e., hours) we will predict ahead of time, and p be another input parameter indicating location of the prediction starting point. The algorithm operates as follows:

1. Divide $\{X_1, \ldots, X_n\}$ into two parts, $\{X_1, \ldots, X_m\}$ and $\{X_{m+1}, \ldots, X_n\}$, where $m = \lfloor pn \rfloor$.
2. Repeat the following steps until $m > n - h$.
   (a) Fit $\{X_1, \ldots, X_m\}$ to obtain a model denoted by M_m.
   (b) Use M_m to predict the number of attacks, denoted by Y_{m+l}, that will arrive during the $(m + l)$th step, where $l = 1, \ldots, h$.
   (c) Compute prediction error $e_{m+l} = X_{m+l} - Y_{m+l}$ for $l = 1, \ldots, h$.
   (d) Set $m = m + h$.

Prediction results for network-level attack processes Now we report the prediction results, while comparing the LRD-aware FARIMA model and the LRD-less ARMA model. Table 2.9 describes the prediction error of the network-level attack processes. We observe the following. First, for Periods I and II, both 1-hour ahead and 5-hour ahead FARIMA prediction errors are no greater than 22%. However, the 10-hour ahead FARIMA prediction is pretty bad. This means that LRD-aware FARIMA can effectively predict the attack rate even five hours ahead of time. This would give the defender enough early-warning time.

Second, Period III network-level attack process exhibits spurious LRD. However, both the LRD-aware FARIMA and the LRD-less ARMA models can predict incoming attacks up to 5 hours ahead of time. Indeed, the prediction error of FARIMA is slightly greater than the prediction error

Table 2.9: Prediction error of network-level attack processes using the LRD-aware FARIMA and the LRD-less ARMA, where prediction errors are defined in Section 3.2. $p = 0.5$ means that we start predicting in the midpoint of each network-level attack process.

	PMAD		PMAD$'$	
Period	FARIMA	ARMA	FARIMA	ARMA
1-hour ahead prediction ($h = 1,\ p = 0.5$)				
I	0.179	0.446	0.173	0.157
II	0.217	0.363	0.149	0.149
III	0.298	0.273	0.305	0.312
IV	0.548	0.526	0.126	0.106
V	0.517	0.529	0.424	0.411
5-hour ahead prediction ($h = 5,\ p = 0.5$)				
I	0.206	0.556	0.292	0.314
II	0.212	0.351	0.420	0.411
III	0.297	0.272	0.246	0.250
IV	0.847	0.838	0.226	0.207
V	0.526	0.555	0.414	0.417
10-hour ahead prediction ($h = 10,\ p = 0.5$)				
I	0.869	0.801	0.314	0.281
II	1.024	1.034	0.277	0.284
III	1.00	1.002	0.202	0.201
IV	0.648	0.627	0.282	0.490
V	0.982	0.952	0.402	0.412

of ARMA. This reiterates that if an attack process does not exhibit LRD, it is better not to use LRD-aware prediction models; if an attack process does exhibit LRD, LRD-aware prediction models should be used instead. This highlights the advantage of “gray-box" prediction over “black-box" prediction.

Third, although Period IV exhibits LRD, even its 1-hour ahead FARIMA prediction is not good enough, with prediction error greater than 50%. While it is unclear what caused this effect, we note that the underestimation error PMAD' for 5-hour ahead prediction is still reasonable for Period IV (22.6$\%$ for FARIMA and ARMA). This means that if one is willing to over-provision defense resources to some extent, then the prediction for Period IV is still useful.

Table 2.10: Number of victim-level attack processes that can be predicted by the LRD-aware FARIMA model, which is more accurate than the LRD-less ARMA model. For the column “total # of victims $((x_1, x_2)/(y))$," y is the total number of victims that exhibited LRD (or non-LRD), x_1 (or x_2) is total number (out of the y) of victims for which the Maximum Likelihood Estimator (MLE) used in the FARIMA (ARMA) algorithm actually converges (i.e., $y - x_1$ and $y - x_2$ victims cannot be predicted because the MLE does not converge). The column “# of victims w/ average OA (or UA) $\geq z\%$" represents the average number of victims (out of the x_1 or x_2 victims that can be predicted), which lead to average prediction accuracy in terms of overall-accuracy OA (or underestimation-accuracy UA) at least $z\%$, where average is taken over all predictions.

Period	total # of victims $((x_1, x_2)/(y))$	# of victims w/ average $\mathsf{OA} \geq 80\%$		# of victims w/ average $\mathsf{OA} \geq 70\%$		# of victims w/ average $\mathsf{UA} \geq 80\%$		# of victims w/ average $\mathsf{UA} \geq 70\%$	
		FARIMA	ARMA	FARIMA	ARMA	FARIMA	ARMA	FARIMA	ARMA
I	LRD: (152,152)/(159)	2	1	29	13	13	4	40	35
	non-LRD: (7,7)/(7)	0	0	4	4	1	4	7	6
II	LRD: (109,109)/(116)	0	0	3	2	2	1	12	6
	non-LRD: (50,49)/(50)	0	0	0	0	4	1	6	2
III	LRD: (82,82)/(87)	0	0	4	4	9	5	23	19
	non-LRD: (79,79)/(79)	0	0	0	0	0	0	10	7
IV	LRD: (118,118)/(125)	0	0	2	2	2	3	4	6
	non-LRD: (41,39)/(41)	0	0	0	0	1	0	2	0
V	LRD: (73,73)/(89)	0	0	0	0	0	1	2	3
	non-LRD: (77,61)/(77)	0	0	0	0	0	0	1	0

Fourth, Period V resists both prediction models in terms of both overall prediction error PMAD and underestimation error PMAD'. The fundamental cause of the effect is unknown at the moment, and is left for future studies. Nevertheless, we suspect that Extreme Value Theory could be exploited to address this problem.

Prediction results for victim-level attack processes Since there are 166 victims per period, there are 830 victim-level attack processes for which we will do prediction. Recall that 70% victim-level attack processes exhibit LRD. We use Table 2.10 to succinctly present the prediction results, which are with respect to 10-hour ahead predictions during the last 100 hours of each time period. We make the following observations. First, the LRD-aware FARIMA model performs better than the LRD-less ARMA model. For example, among the 152 (out of the 159) victim-level attack processes in Period I that exhibit LRD and are amenable to prediction (i.e., the Maximum Likelihood Estimator actually converges; the Estimator does not converge for 159-152=7 LRD processes though), FARIMA can predict for 29 victim-level attack processes about their 10-hour ahead attack rates with at least 70% overall accuracy (OA), while ARMA can only predict for 13 victim-level attack processes at the same level of accuracy. If the defender is willing to over-provision some resources and mainly cares about the underestimation error (which could cause overlooking of attacks), FARIMA can predict for 40 victim-level attack processes while ARMA can predict for 35.

Second, the victim-level attack processes in Period I exhibit LRD and render more to prediction when compared with the victim-level attacks processes in the other periods, which also exhibit LRD. Moreover, for non-LRD processes, neither FARIMA nor ARMA can provide good predictions. This may be caused by that (some of) the non-LRD processes are non-stationary. We plan to investigate into these issues in the future.

Summary It is feasible to predict network-level attacks even 5 hours ahead of time. For attack processes that exhibit LRD, LRD-aware models *can* predict their attack rates better than LRD-less models do. However, there are LRD processes that can resist the prediction of even LRD-aware models. This hints that new prediction models are needed.

Table 2.11: For each victim-level attack process that exhibits LRD, some port-level attack processes (called “sub-processes" for short) exhibit heavy-tails.

Period	total # of victims exhibiting LRD	# of victims w/ sub-processes exhibiting heavy-tail	# of victims with certain # of sub-processes exhibiting heavy-tail					total # of ports exhibiting heavy-tail	Shape mean value	# of ports w/ shape value $\in (.5, 1)$	# of ports w/ shape value ≥ 1	Standard deviation
			1	2	3	4	5					
I	159	56	50	6	0	0	0	62	.11	1	0	.11
II	116	80	78	11	1	0	0	103	.40	50	0	.22
III	87	47	39	6	2	0	0	57	.22	2	0	.18
IV	125	3	3	0	0	0	0	3	.43	1	0	.35
V	89	32	29	1	2	0	0	37	.30	5	1	.25

2.4.5 Step 5: Exploring Causes of LRD

Despite intensive studies in other settings, the fundamental cause of LRD is still mysterious. One known possible cause of LRD is the superposition of heavy-tailed processes [37, 38, 72]. Another candidate cause of LRD is that some attackers launch intense (consecutive) attacks (e.g., bruteforcing SSH passwords). In what follows we examine the two candidate causes.

LRD exhibited by network-level attack processes is not caused by heavy-tailed victim-level attack processes We want to know whether or not the LRD exhibited by the 4 network-level attack processes during Periods I, II, IV and V are caused by the superposition of heavy-tailed victim-level attack processes. That is, we want to know how many victim-level attack processes during each of the four periods are heavy-tailed. We find that among the vector of $(166, 166, 166, 166)$ victim-level attack processes during Periods I, II, IV and V, the vector of victim-level attack processes that exhibit heavy-tails is correspondingly $(101, 0, 24, 31)$, by using the POT method that is reviewed in Appendix A-B. This means that Period I is the only period during which majority of victim-level attack processes exhibit heavy-tails. A few or even none processes in the three other periods exhibited heavy-tails. This suggests that LRD exhibited by the network-level attack processes does not have the same cause as what is believed for benign traffic [59].

LRD exhibited by victim-level attack processes is not caused by heavy-tailed port-level attack processes Now we investigate whether or not the LRD exhibited by victim-level attack processes is caused by that the underlying port-level attack processes exhibit heavy-tails, a property briefly reviewed in Appendix 6.1.2. Table 2.11 shows that only 8% port-level attack processes, or $56 + 80 + 47 + 3 + 32 = 218$ out of the $(159 + 116 + 87 + 125 + 89 = 576)$ victims × 5 ports/victim = 2880 port-level attack processes, exhibit heavy-tails. Moreover, only 29 (out of the 576) victim-level attack processes have 2 or 3 port-level attack processes that exhibit heavy-tails. Further, there is only 1 port-level attack process that exhibits infinite mean because the shape value ≥ 1, and there are $1 + 50 + 2 + 1 + 5 = 59$ port-level attack processes that exhibit infinite variance because their shape values $\in (.5, 1)$. The above observations also hint that unlike in the setting of benign traffic [59], LRD exhibited by victim-level attack processes is not caused by the superposition of heavy-tailed port-level attack processes.

LRD exhibited by victim-level attack processes is not caused by individual intense attacks Now we examine whether or not LRD is caused by the individual attackers that launch intense attacks. For this purpose, we consider *attacker-level attack processes*, which model the attacks against each victim that are launched by *distinct* attackers. In other words, we only consider the first attack launched by each attacker, while disregarding the subsequent attacks launched by the same attacker.

Table 2.12 describes the observed lower-bound and upper-bound of the four statistics regarding the attacker-level processes, where the bounds are among all victims within a period of time. By taking Period II as an example, we observe the following: on average there are between 48 and 100 attackers against one individual victim within one hour, and there can be up to 621 attackers against one individual victim within one hour. Further, attacks in Periods III and IV exhibit different behaviors from the other three periods. From the boxplots of the basic statistic, which are not presented for the sake of saving space, we observe that the attackers' behaviors are actually very different in the 5 periods. In particular, the attacker-level attack processes in Period II have many outliers in terms of the four statistics, meaning that the attack rate during this period varies a lot.

In order to see whether or not the attacker-level attack processes still exhibit LRD, we describe their Hurst parameters in Table 2.13. Using Period I as an example, we observe that there are 153

Table 2.12: Basic statistics of attack rate of the attacker-level attack processes (per hour).

Period	Mean(·)		Median(·)		Variance(·)		MAX(·)	
	LB	UB	LB	UB	LB	UB	LB	UB
I	30.2	67.8	4	45	1498.1	4094.3	225	432
II	48.6	100.8	42	93	1195.1	6298.3	306	621
III	11.1	33.0	2	29	223.6	270.8	64	100
IV	1.9	23.8	1	23	26.32	92.7	40	65
V	33.4	127.9	8	105	1132.7	7465.2	266	605

Table 2.13: The estimated Hurst parameters of the attack rate of attacker-level attack processes (per hour).

Period	RS		AGV		Peng		Per		Box		Wave		# victims w/	# victims w/
	min	max	min	max	min	max	min	max	min	max	min	max	$\bar{H} \in [.6, 1]$	LRD
I	0.593	0.977	0.851	0.958	0.896	1.111	1.174	1.334	0.942	1.185	0.582	0.843	153	153
II	0.570	0.883	0.616	0.950	0.689	1.070	0.710	1.152	0.663	1.242	-0.360	0.728	92	77
III	0.776	0.994	0.364	0.747	0.630	0.748	0.460	0.679	0.608	0.746	0.389	0.668	163	103
IV	0.657	0.920	0.273	0.955	0.690	0.872	0.559	1.206	0.612	0.952	0.288	1.004	166	165
V	0.495	0.758	0.563	0.727	0.499	0.806	0.898	1.114	0.660	0.977	0.567	0.931	166	77

(out of the 166) victims whose corresponding attacker-level attack processes suggest LRD because their average Hurst parameter $\in [.6, 1]$, where the average is taken over the six kinds of Hurst parameter methods. Moreover, none of the 153 attacker-level processes exhibit spurious LRD. Using Period V as another example, we observe that all 166 attacker-level attack processes have average Hurst parameter $\in [.6, 1]$, but only 77 attacker-level attack processes exhibit LRD while the other 89 attacker-level attack processes exhibit spurious LRD (caused by non-stationarity of the processes). The above discussion suggests that LRD exhibited by victim-level attack processes is not caused by the intense (consecutive) attacks launched by individual attackers, simply because most (or many) attacker-level attack processes also exhibit LRD.

Summary The LRD exhibited by the attack processes is neither caused by the superposition of heavy-tailed sub-processes, nor caused by the intense attacks that are launched by individual attackers. While we ruled out the two candidate causes of LRD, it is an interesting and challenging future work to precisely pin down the cause of LRD in this context.

2.4.6 Limitation of the Case Study

Although our statistical framework is widely applicable, our case study has three limitations that are imposed by the specific dataset. First, the dataset, albeit over $47 + 18 + 54 + 21 + 80 = 220$ days in total (5 periods of time), only corresponds to 166 honeypot IP addresses. We wish to have access to significantly larger dataset of this kind. Nevertheless, it is notoriously difficult to get such data for various reasons. For example, it appears that even the PREDICT project (`https://www.predict.org/`) does not offer this kind of data. Still, this chapter explores an important direction for cyber security research because of the potential reward in understanding the statistical properties of cyber attacks and in possibly predicting the incoming attacks with good accuracy.

Second, the dataset is attack-agnostic in the sense that we know the ports/services the attackers attempt to attack, but not the specific attack details because the data was collected using low-interaction honeypots. Although this issue can be resolved by using high-interaction honeypots, there are legitimate concerns about high-interaction honeypots from a legal perspective.

Third, the data is collected using honeypot rather than using production network. For real-life adoption of the prediction capability presented in the chapter, attack traffic would be blended into the production traffic. Whether or not the blended traffic also exhibits LRD is an interesting future study topic. The main challenge again is the legal and privacy concerns. (It may not be a good idea to simply blend the honeypot traffic with production traffic because this would disrupt the attack process structure.)

2.5 Conclusion and Future Work

We introduced the concept of stochastic cyber attack processes, which offers a new perspective for studying cyber attacks. We also proposed a statistical framework for analyzing such processes. By applying the framework to some honeypot-collected attack data, we found that majority of the attack processes exhibit LRD. We demonstrated that LRD-aware models can better predict the attack rates 5 hours ahead of time, especially for network-level attack processes. This hints that attacks against enterprise-level networks are probably more amenable to prediction than attacks against individual computers. The prediction power comes from “gray-box" (rather than “black-box") models.

The present study introduces a range of interesting problems for future research. First, we need to further improve the prediction accuracy, despite that the LRD-aware FARIMA model can predict better than the LRD-less ARMA models. For this purpose, we plan to study some advanced models with high volatilities. Second, it is important to rigorously explain the fundamental cause of LRD as exhibited by honeypot-captured cyber attacks. Our study only ruled out two candidate causes.

Third, the victim-level attack processes and network-level attack processes exhibited similar phenomena (i.e., LRD). This hints a sort of *scale-invariance* that, if turns out to hold, would have extremely important implications (for example) in achieving scalable analysis of cyber attacks.

Chapter 3: ANALYZING EXTREME VALUES EXHIBITED BY CYBER ATTACKS

3.1 Introduction

Data-driven analytics can deepen our understanding about the statistical phenomena/properties of cyber attacks, and can potentially help tackle the fundamental feasibility of predicting cyber attacks (possibly at some some level of aggregation). Any significant progress in cyber attack predictability, even for minutes (if not hours) ahead of time and even at an aggregate level, would give the defender enough earlywarning time to prepare for adequate defense (e.g., the defender can dynamically allocate sufficient resources for deep packet inspection or flow-level assembly and analysis). Despite its clear importance, there has been little progress, perhaps because of the lack of real data and readily usable statistical methodologies.

Recently, we made a first such effort, by proposing a statistical framework to formulate and answer the relevant questions [77]. In the present chapter, we make a further substantial step toward the ultimate goal, by studying the *extreme-value phenomenon* exhibited by honeypot-captured cyber attacks. The extreme-value phenomenon refers to the many outliers above certain thresholds, namely extremely large attack rates (per unit time) against a target of interest (e.g., honeypot in the context of this chapter). It is important to investigate the extreme-value phenomenon because of the following.

First, the extreme-value phenomenon is robust because it cannot be filtered out by SNORT. This means that studying the extreme-value phenomenon exhibited by honeypot-captured cyber attacks is useful to the defense of production networks, provided that the honeypot is not bypassed by the attacker (which can be assured by randomizing the locations of honeypots). For example, when the attacker launch intense new attacks that cannot be detected, the distribution of the extreme attack rates observed at the honeypots are about the same as the distribution of the extreme attack rates observed at production networks. This distribution information can be used to adjust the defense at the production networks (e.g., using resource-consuming behavior-based detection, rather than resource-effective signature-based detection). This is especially useful when we can predict the attack rates even hours ahead of time, which is possible as shown in the chapter.

Second, one might think that the extreme values are caused by denial-of-service attacks. Our analysis shows that denial-of-service attack is indeed among the most often seen attacks (as recognized by SNORT) for the largest clusters of extreme attack rates (i.e., intense attacks that sustain for hours). However, the most often seen attack at the hour of highest attack rate (i.e., the hour of most intense attacks) is *not* necessarily denial-of-service, but attacks that can be buffer-overflow exploits (as recognized by SNORT). This further confirms the value of studying the extreme phenomenon.

3.1.1 Our Contributions

We propose a novel methodology for investigating the extreme-value phenomenon exhibited by cyber attacks. To the best of our knowledge, we are the first to study the extreme-value phenomenon in the domain of cyber security analytics, while noting that the methodology can be seamlessly incorporated into, and therefore enhance, the framework described in [77]. More specifically, we make two contributions.

First, we propose a methodology for systematically characterizing the extreme-value phenomenon exhibited by (honeypot-captured) cyber attacks. The methodology aims to integrate two complementary statistical approaches: the Extreme Value Theory (EVT) and the Time Series Theory (TST). We investigate a connection between the two approaches, when applying them to predict extreme attack rates. We conclude that these two predictive approaches should be used together in practice, because EVT-based methods are more appropriate for long-term predictions and TST-based methods are more appropriate for short-term predictions. A combination of the two kinds of predictions can lead to more useful results for guiding the defender's resource allocation decision-making. As we will see, a resource allocation strategy based on EVT-predicted magnitude of attack rates (24 hours ahead of time), but with adjustments based on TST-predicted maximum attack rates (1 hour ahead of time), can cope with the worst-case scenario (i.e., the largest attack rate). This strategy gives the defender more earlywarning time. On the other hand, a more cost-effective strategy can be based on TST-predicted maximum attack rates (1 hour ahead of time), while taking into consideration EVT-predicted magnitude of attack rates (24 hours ahead of time), would be able to cope with the average-case scenarios. This strategy also requires the defender to be more agile than the previous strategy.

Second, we propose a novel statistical technique for analyzing cyber attack data. Specifically, we propose a family of time-series FARIMA+GARCH models, which can accommodate both the extreme-value phenomenon and the Long-Range Dependence (LRD) phenomenon exhibited by cyber attacks, where GARCH (Generalized AutoRegressive Conditional Heteroskedasticity) accommodates the extreme-value phenomenon, and FARIMA (Fractional AutoRegressive Integrated Moving Average) accommodates the LRD phenomenon. We show that FARIMA+GARCH can indeed better fit and predict than existing popular methods. This further highlights the power of “gray-box" fitting and predictions.

3.1.2 Related Work

Although this chapter is the first to analyze the extreme-value phenomenon exhibited by cyber attacks, the study is inspired by, and substantially extends, the statistical framework we recently proposed [77]. The framework formulates a systematic way of thinking in cyber attack analytics, which is centered on the novel concept of *stochastic cyber attack process*. A key finding in [77] is that cyber attacks (i.e., stochastic cyber attack processes) exhibits the LRD phenomenon, and the FARIMA model (which accommodates LRD) can better fit and predict the attack-rate time series than the ARMA model (which cannot accommodate LRD). Since we here analyze the extreme-value phenomenon that is not analyzed in [77], the analysis methodology and techniques described in the present chapter can be seamlessly incorporated into the framework [77] to enhance its statistical power.

On the other hand, honeypot-captured cyber attack data has been studied from several different perspectives, such as: classifying the captured attacks into known and unknown attacks [8], identifying the traffic characteristics of the same attacks [67], and others [9,10,20,21,29,42,53,54,67]. Loosely related investigations include the analysis of blackhole-captured traffic data (e.g., [51,73]) or one-way traffic [31], which emphasize on classifying the data into classes (e.g. scanning, peer-to-peer applications, unreachable services, misconfigurations, worms etc). There also have been studies that aim to extract useful information from honeypot-captured data, such as: probing activities [39], Denial of Service attacks [28], scans [34], worms [25,26] and botnets [41,63].

Our study is at a higher level of abstraction, by considering the statistical properties of the aggregate data (i.e., attacks coming to an entire honeypot) and focusing on their prediction utility (i.e., exploiting the exhibited statistical properties for possibly better detection).

The chapter is organized as follows. Section 3.2 briefly reviews some preliminary statistical knowledge. Section 3.3 describes our statistical analysis methodology. Section 3.4 uses EVT to analyze the extreme attack rates. Section 3.5 uses TST to analyze the time series data. Section 3.6 discusses a connection between EVT-based and TST-based predictions. Section 4.5 discusses limitations of the present study and directions for further research. Section 4.6 concludes the chapter.

3.2 Statistical Preliminaries

We now review the main statistical concepts and techniques that are used in the present chapter. Throughout the chapter, we use the more intuitive terms “extreme values", “extreme events", “extreme-value events" and the EVT jargon “exceedances" interchangeably. We also use the intuitive term “average inter-arrival time" between consecutive extreme values and the EVT jargon “return period" interchangeably.

3.2.1 Statistics of Extreme-Value Phenomena

Figure 3.1 illustrates a time series of attack rates (per some unit time), where the threshold line corresponds to a threshold value μ such that the green dots (above the threshold line) are *extreme attack rates* or *extreme values*. At a high level, the extreme-value phenomenon can be characterized from a “spatial" perspective (i.e., the *distribution* of the magnitude of extreme values), a “time" perspective (i.e., the *inter-arrival time* between extreme values), and a “spatial-time" perspective (i.e., the concept of *return level* described below).

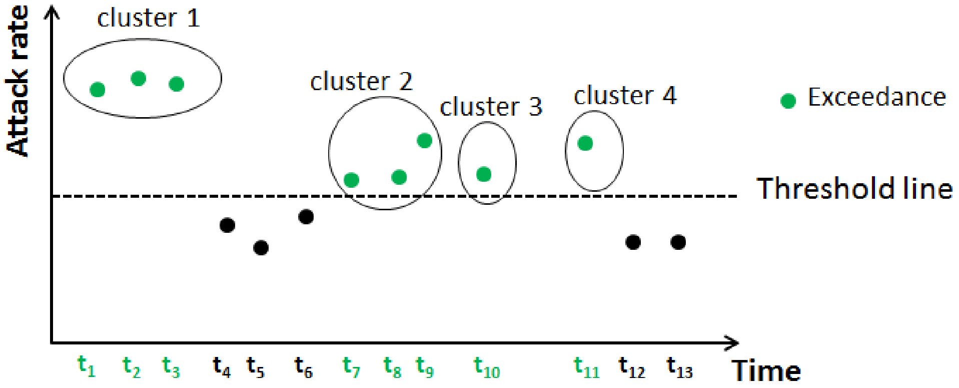


Figure 3.1: Illustration of extreme-value phenomena: dashed line represents a threshold; green dots are extreme values or *exceedances*; black dots are non-extreme values; extremal index θ indicates the clustering degree of extreme values (where a cluster is a set of consecutive extreme values).

Distribution of extreme values

It is known that if $X_1, \ldots, X_n$ are stationary, then $[X_i - \mu | X_i > \mu]$ may follow the standard GPD (Generalized Pareto Distribution) with survival function

$$\bar{G}_{\xi,\sigma(\mu)}(x) = 1 - G_{\xi,\sigma(\mu)} = \begin{cases} \left(1 + \xi \dfrac{x}{\sigma}\right)^{-1/\xi}, & \xi \neq 0, \\ \exp\{-x/\sigma\}, & \xi = 0. \end{cases},$$

where $x \in \mathbb{R}^+$ if $\xi \in \mathbb{R}^+$ and $x \in [0, -\sigma/\xi]$ if $\xi \in \mathbb{R}^-$, ξ and σ are respectively called *shape* and *scale* parameters. If $X_1, \ldots, X_n$ are from a non-stationary process, then $[X_i - \mu | X_i > \mu]$ may follow a non-stationary GPD with time-dependent parameters, namely

$$\bar{G}_{\xi(t),\sigma(t)}(x) = \begin{cases} \left(1 + \dfrac{\xi(t)x}{\sigma(t)}\right)^{-1/\xi(t)}, & \xi(t) \neq 0, \\ \exp\{-x/\sigma(t)\}, & \xi(t) = 0. \end{cases}$$

To fit the distribution of extreme values, we use the POT (Point Over Threshold) method [27,59].

Extremal index $0 < \theta \leq 1$

Intuitively, this index captures the degree that the extreme values are clustered as follows (cf. Figure 3.1): $1/\theta$ indicates the mean size of the clusters of extreme values; $\theta = 1$ means that each cluster has one extreme value (i.e., the extreme values do not exhibit the clustering behavior). Formally, let $\{X_1, X_2, \ldots, \}$ be a sequence of random variables from a stationary process. Let $M_n = \max\{X_1, \ldots, X_n\}$. Under certain regularity conditions, it holds that

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = H_\xi^\theta(x),$$

where a_n and b_n are normalizing constants,

$$H_\xi(x) = \begin{cases} \exp\left\{-(1+\xi\frac{x-\mu}{\sigma})^{-1/\xi}\right\}, & \xi \neq 0 \\ \exp\{-e^{-\frac{x-\mu}{\sigma}}\}, & \xi = 0 \end{cases} \tag{3.1}$$

is a non-degenerate distribution function with $1 + \xi\frac{x-\mu}{\sigma} > 0$, and $\theta \in (0, 1]$ is the *extremal index*.

Return level

Intuitively, return level captures the expected *magnitude* of extreme values (but not necessarily the *maximum* value). Let T be the average inter-arrival time between consecutive extreme values, which is also called *return period*. The probability that an extreme event occurs is $p = 1/T$. The concept of *return level* identifies a special threshold such that there is, on average, a single extreme event during each return period. Formally, suppose random variable X has a stationary GPD with shape parameter ξ and scale parameter σ. Then,

$$P(X > x) = \zeta_\mu \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}, \quad \xi \neq 0,$$

where $\zeta_\mu = P(X > \mu)$. The *return-level* x_m, which is exceeded on average once per m observations, is given by $x_m = \mu + \sigma/\xi \left[(m\zeta_\mu)^\xi - 1\right]$. For non-stationary GPD, the return level is given by $x_m = \mu + \sigma(m)/\xi(m) \left[(m\zeta_\mu)^{\xi(m)} - 1\right]$.

3.2.2 Properties and Models of Time Series

Long-Range Dependence (LRD)

Unlike the Poisson process that exhibits the memoryless property, a LRD process exhibits persistent correlations that the autocorrelation decays slowly (i.e., slower than the exponential decay). Formally, a stationary time series $\{X_t\}$ exhibits LRD if its autocorrelation function is $r(h) = \mathrm{Cor}(X_i, X_{i+h}) \sim h^{-\beta} L(h)$ as $h \to \infty$, where $0 < \beta < 1$ and $L(\cdot)$ is a slowly varying function [27]. Note that $\lim\limits_{x\to\infty} \frac{L(tx)}{L(x)} = 1$ for all $t > 0$, and $\sum_h r(h) = \infty$. The degree of LRD is quantified by the Hurst parameter $H = 1 - \beta/2$, meaning that $1/2 < H < 1$ and the degree of LRD increases as $H \to 1$.

FARIMA and GARCH Time Series Models

FARIMA (Fractional AutoRegressive Integrated Moving Average) and GARCH (generalized autoregressive conditional heteroskedasticity) are two widely used time-series models [23]. FARIMA can accommodate LRD and GARCH can accommodate the extreme-value phenomenon. Let $\phi(x) = 1 - \sum_{j=1}^{p} \phi_j x^j$, $\psi(x) = 1 + \sum_{j=1}^{q} \psi_j x^j$, and ϵ_t be independent and identical normal random variables with mean 0 and variance σ_ϵ^2.

A time series $\{X_t\}$ is called a FARIMA(p, d, q) process if $\phi(B)(1 - B)^d X_t = \psi(B)\epsilon_t$, where $-1/2 < d < 1/2$, and B is the back shift operator $BX_t = X_{t-1}$, $B^2 X_t = X_{t-2}$, etc.

On the other hand, a time series $\{X_t\}$ is called a GARCH process [16] if $X_t = \sigma_t \epsilon_t$, where the noises (also called *innovations*) ϵ_t's are the standard white noise distribution.

We consider two variants of GARCH. For the Standard GARCH (SGARCH) model, we have $\sigma_t^2 = w + \sum_{j=1}^{q} \alpha_j \epsilon_{t-j}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$.

For the Integrated GARCH (IGARCH) model, we have

$\phi(B)(1-B)\epsilon_t^2 = w + (1-\psi(B))v_t,$

where $v_t = \epsilon_t^2 - \sigma_t^2$.

To accommodate more general noise distributions, we will consider skewed Student-T distribution (SSTD) or skewed Generalized Error distribution (SGED).

3.3 Data and Analysis Methodology

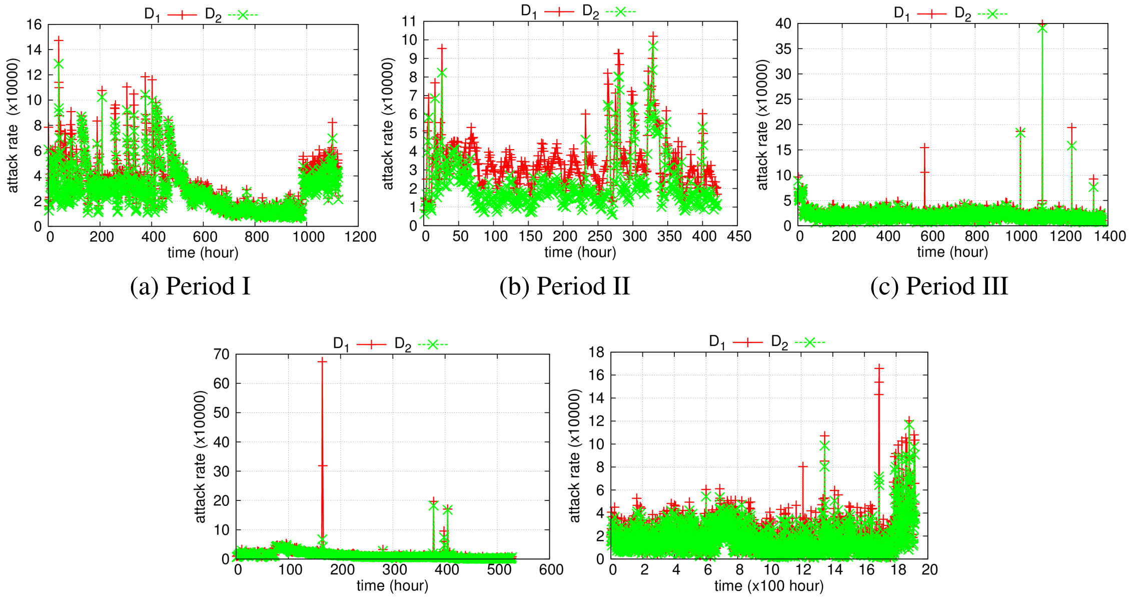
3.3.1 Data Description

The raw data (in `pcap` format), denoted by D_0, was collected by a honeypot of 166 IP addresses, during five periods in 2010-2011 of time respectively 47, 18, 54, 21 and 80 days (the same as in [77]). The honeypot ran four low-interaction honeypot programs: Amun [1], Dionaea [2], Mwcollector [3] and Nepenthes [11]. Each program was associated to a unique IP address. Each physical computer monitored multiple IP addresses. As we elaborate below, we derived D_1 and D_2 from D_0, where D_1 describes what is observed at the honeypot (i.e., the view at the attacker's end) and D_2 is closer to what can be observed at production networks with SNORT filtering (i.e., the view at the defender's end).

Specifically, D_1 was derived from D_0 as follows: We extracted the attack traffic with respect to production ports (because the traffic corresponding to nonproduction ports contains little useful information), and treated each remote-end initiated TCP flow as an attack (because the honeypot does not offer any legitimate service). An unsuccessful TCP handshake was also deemed as attack, as the handshake could have been dropped by the honeypot computer. For assembling TCP flows, we set the flow lifetime as 300 seconds (i.e., an attack/flow does not span over 300 seconds) and the flow timeout time as 60 seconds (i.e., an attack/flow expires if there is no activity for 60 seconds); these parameters were suggested in [7].

D_2 was derived from D_0 as follows: We first replayed D_0 against the widely used SNORT intrusion detection system version 2.9.3 [4], which was released on July 20th, 2012 (i.e., about

8 months after the data was collected as we wanted to know, in a sense, the best-case effect of SNORT filtering). As such, D_2 would be closer to the view of the attacks reaching production networks (with perimeter defense like SNORT). The output of the replaying procedure is then assembled into TCP flows as in the case of deriving D_1 from D_0.



(a) Period I

(b) Period II

(c) Period III

(d) Period IV

(e) Period V

Figure 3.2: Time series plots of attack rates (number of attacks per hour): non-filtered D_1 (red) vs. SNORT-filtered D_2 (green): The percentages of attacks filtered by SNORT for Periods I-V are respectively 14.0%, 33.5%, 11.3%, 19.4% and 16.4%, meaning that SNORT has limited success. The extreme-value phenomenon is robust and prolific in D_1 and D_2.

3.3.2 The Extreme-Value Phenomenon

Figure 3.2 plots D_1 and D_2 in the five periods of time. We observe the extreme-value phenomenon, namely the many extreme attack rates (i.e., extreme values or spikes) caused by intense attacks in D_1 of all five periods. Some extreme values in Periods III-V are filtered by SNORT. For example, the spikes at the 527th hour in Period III and the 165th hour in Period IV are detected by SNORT as “Malware suspicious FT 200 Banner on Local Port"; the two spikes at the 1237th and 1335th hours in Period III are detected as “SSH scan"; the spike at the 1693th hour in Period V is detected as “SIP Invite Message Remote Denial of Service Vulnerability". However, there are still many extreme values in D_2. This means the extreme-value phenomenon is robust and important for the defense of production networks, provided that the honeypot is not bypassed by the attacker (which can be justified by the extreme-value phenomenon itself in this case, and achieved by randomizing the locations of honeypots in general).

3.3.3 Extreme-Value Analysis Methodology

Our methodology is centered on analyzing statistical properties of extreme attack rates and the feasibility of exploiting the statistical properties to predict the extreme values (for dynamic proactive defense). For this purpose, we aim to integrate two complementary statistical approaches. The first approach is based on EVT (Extreme Value Theory), which deals with extreme attack rates. This approach is appropriate for relatively long-term prediction of extreme events (e.g., 24-hour ahead of time), because (i) extreme events would not occur often (otherwise, they may not be extreme events any more), and (ii) the analysis only considers extreme attack rates. The second approach is based on TST (Time Series Theory), which does not differentiate the extreme values (above a threshold) and the non-extreme values (below the threshold). This explains why the TST approach is appropriate for short-term prediction (e.g., 1-hour ahead of time). The two approaches are complementary to each other because (i) the data/information they use is different (i.e., proper subset vs. super set), and (ii) the predictions they make are different (i.e., return levels or expected magnitude of extreme attack rates vs. concrete attack rates). Therefore, it is interesting to seek connection between the two approaches, especially from the prediction perspective.

For both EVT- and TST-based analyses, we propose to proceed as follows: First, identify the statistical properties exhibited by D_1 and D_2, such as: the stationarity of the processes that drive D_1 and D_2 and the clustering behavior of the extreme values. Second, exploit the identified statistical properties to fit the relevant data (i.e., “gray-box" fitting). Although there are generic methods for analyzing the extreme-value phenomenon, our intuition is that the generic methods are not sufficient because they do not consider the properties exhibited by cyber attacks. This motivates us to propose new statistical techniques that are relevant to the cyber security domain (e.g., the family of FARIMA+GARCH models). Third, predict attack rates by using EVT- and TST-based methods (i.e., “grad-box" prediction), and explore the relation (especially, the consistency) between the predicted results.

3.4 EVT-based Extreme-Value Analysis

In order to fit the distribution of extreme values, we need to determine the extreme values are driven by a stationary process (i.e., the distribution does not change) or non-stationary process, because we need to use time-invariant or time-dependent distribution parameters. This suggests us to consider four candidate distributions/models: $M_1, \ldots, M_4$, where M_1 corresponds to the stationary process case and the others correspond to the non-stationary process case. Specifically,

- M_1: The standard GPD.
- M_2: GPD with time-invariant shape parameter ξ but time-dependent scale parameter $\sigma(t) = \exp\left(\beta_0 + \beta_1 \log(t)\right)$.
- M_3: GPD with time-invariant scale parameter σ but time-dependent shape parameter $\xi(t) = \gamma_1 + \gamma_2 t$.
- M_4: GPD with time-dependent parameters $\sigma(t) = \exp\left(\beta_0 + \beta_1 \log(t)\right)$ and $\xi(t) = \gamma_1 + \gamma_2 t$.

Since we do not know the stationarity a priori, we first use M_1 to fit the extreme attack rates. If M_1 cannot fit well, we use non-stationary distributions/models $M_2, \ldots, M_4$ to fit the extreme attack rates. We use standard goodness-of-fit statistics and QQ-plot for evaluating the quality of fitting.

3.4.1 Fitting stationary extreme attack rates

We use Algorithm 3.1 to fit stationary extreme attack rates. The algorithm uses QQ-plot and two goodness-of-fit statistics called CM and AD, where CM and AD measure the goodness-of-fit of a distribution. If the p-values of both CM and AD statistics are greater than $.1$ (which is more conservative than the textbook criterion $.05$), and the QQ-plot also confirms the goodness-of-fit, the algorithm concludes that the extreme attack rates are stationary and follow the standard GPD; otherwise, the extreme attack rates are non-stationary and will be fitted via Algorithm 3.2 described below.

Algorithm 3.1 Fitting stationary extreme attack rates via M_1

```
INPUT: attack-rate time series
OUTPUT: $M_1$ fitting result
initialize $quantileSet$ {assuring $\geq 30$ extreme values}
for $q \in quantileSet$ (from the minimum to the maximum in the increase order) do
    use the standard GPD to fit the extreme attack rates that are greater than threshold quantile $q$
    evaluate goodness-of-fit statistics CM, AD, QQ-plot
    if fitting is good then
        estimate GPD parameters $(\xi, \sigma)$, extremal index $\theta$
        return $(q, \xi, \sigma, \theta)$ {the first successful fitting}
return -1 {stationary distribution fitting failed}
```

A key ingredient in Algorithm 3.1 is the threshold quantile, namely the threshold specified by a quantile such that an attack rate above the threshold is an extreme value. Specifically, $quantileSet$ is an ordered set of quantiles, where the maximum quantile is chosen to guarantee there are at least 30 extreme attack rates (because, as a rule of thumb, 30 is required for the sake of reliable fitting), and the minimum quantile is 20% difference from the maximum quantile with step-length 5%. For example, suppose there are 1000 attack rates (corresponding to observations during 1000 hours). The maximum threshold quantile is $1-\frac{30}{1000}\times 100\% = 97\%$ and the minimum threshold quantile is 77%, leading to $quantileSet = \{77\%, 82\%, 87\%, 92\%, 97\%\}$. The algorithm starts with threshold quantile 77%, and then 82% etc. (i.e., according to the increase order), and halt until the first successful fitting (in which case, parameters are obtained) or all the fitting attempts fail (i.e., the process is non-stationary).

Table 3.1 summarizes the fitting results using M_1. For non-filtered data D_1, Periods III and V are from stationary processes: the former has threshold quantile $q = 0.90$ (i.e., there are 130 extreme attack rates that are above the 90% quantile) and extremal index $\theta = 0.60$ (i.e., a cluster contains, on average, $1/.60 = 1.67$ extreme values, or extensive attacks sustain on average 1.67 hours); the latter has threshold quantile $q = 0.95$ and extremal index $\theta = .33$. For SNORT-filtered

Table 3.1: EVT-based fitting of stationary attack rates, where (q, ξ, σ, θ) are output by Algorithm 3.1, “# of EV" means “number of extreme values (i.e., extreme attack rates)", “# of C" means “number of clusters". The last column indicates the best fitting model.

Period	q	# of EV	# of C	θ	ξ	σ	model
D_1: non-filtered attack-rate time series							
III	0.90	130	95	0.73	0.36	3778.19	M_1
V	0.95	96	31	0.33	0.16	13553.5	M_1
D_2: SNORT-filtered attack-rate time series							
III	0.95	69	28	0.40	0.95	5722.72	M_1

data D_2, only Period III is from a stationary process, with threshold quantile $q = 0.95$, extremal index $\theta = 0.40$, and shape parameter $\xi = 0.95 > 0.5$ (meaning infinite variance of extreme attack rates, namely heavy-tailed extreme attack rates). Combining the above observations and the fact that the lengths of the five periods are respectively 47, 18, 54, 21, and 80 days, we suspect that stationary process may not be observed for a period of time shorter than 50 days, which may be necessary but not sufficient (noting that Period V corresponds to 80 days).

Algorithm 3.2 Fitting non-stationary extreme attack rates

```
INPUT: non-stationary extreme attack rates
OUTPUT: fitting result
initialize $quantileSet$ {same as in Algorithm 3.1}
for $q \in quantileSet$ (from the minimum to the maximum in the increase order) do
    use models $M_2$, $M_3$ and $M_4$ to fit the attack rates that are greater than threshold quantile $q$
    evaluate goodness-of-fit via AIC and QQ-plot
    if any of the three models fits well then
        choose the model with the minimum AIC value, or choose the simpler/simplest model
        whose AIC value is fairly close to the minimum AIC value
        return $(q, \text{AIC value})$ of the selected model $M_j$
return -1 {failed in fitting extreme attack rates}
```

3.4.2 Fitting non-stationary extreme attack rates

We use Algorithm 3.2 to select the best fitting model for the non-stationary extreme attack rates in D_1 and D_2, where $quantileSet$ is the same as in Algorithm 3.1. We use the AIC (Akaike information criterion) statistic [23] and QQ-plot to evaluate the goodness-of-fit, where AIC captures the fitting loss (i.e., the smaller the AIC value, the better the fitting). As a thumb of rule, two AIC values are considered *fairly close* when their difference is less than 10. If a model $M_j \in \{M_2, M_3, M_4\}$ incurs the minimum AIC value that is not fairly close to any of the other two AIC values, we choose M_j as the best fitting model; otherwise, we choose the simpler/simplest model whose AIC value is fairly close to the minimum AIC value (model M_2 is considered simpler than M_3, which is simpler than M_4).

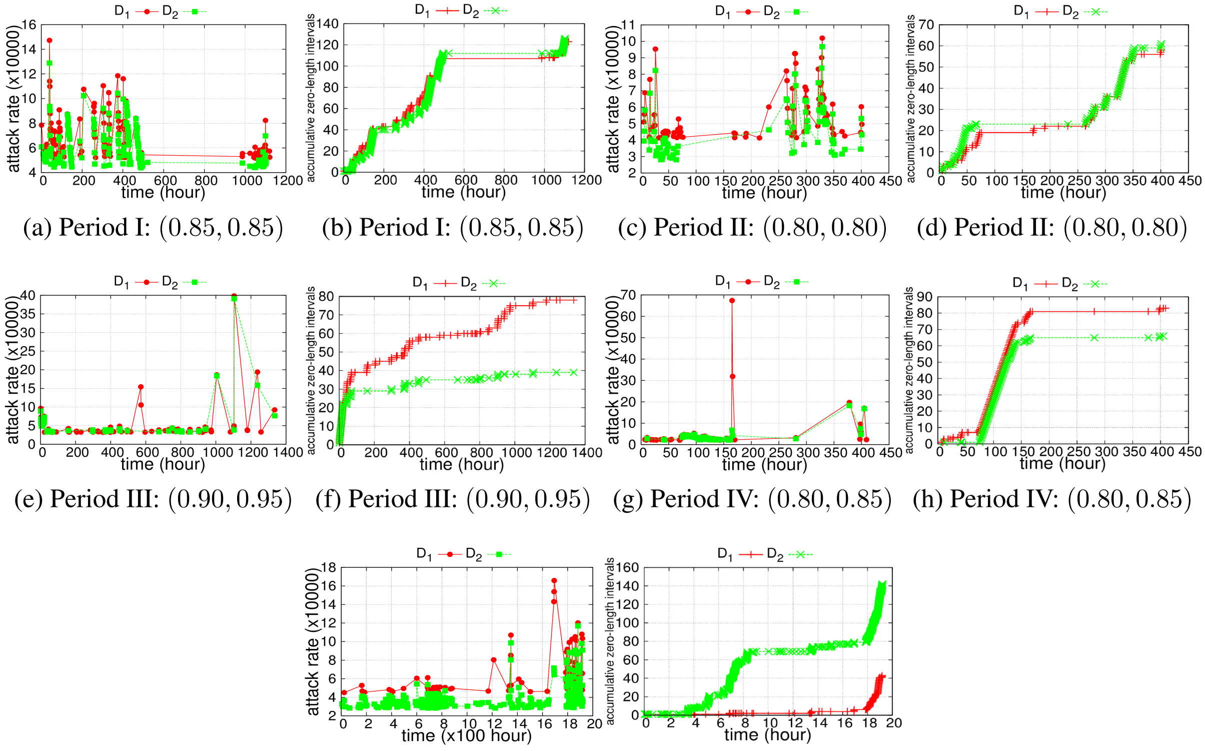
Table 3.2: EVT-based fitting of non-stationary extreme attack rates, where column M_j $(2 \leq j \leq 4)$ represents the AIC value of model M_j, $\surd$ indicates that QQ-plot confirms fitting well (QQ-plots are omitted for saving space), and the other notations are the same as in Table 3.1.

Period	q	# of EV	M_2	M_3	M_4	QQ-plot	model
AIC values for non-filtered extreme attack rates in D_1							
I	0.85	168	3656	3653	3656	$\surd$	M_2
II	0.80	84	1774	1774	1776	$\surd$	M_2
IV	0.80	105	1774	2014	2016	$\surd$	M_2
AIC values for SNORT-filtered extreme attack rates in D_2							
I	0.85	168	3632	3629	3632	$\surd$	M_2
II	0.80	84	1816	1821	1823	$\surd$	M_2
IV	0.85	79	1672	1649	1651	$\surd$	M_3
V	0.85	288	5908	5901	5906	$\surd$	M_2

Table 3.2 summarizes the fitting results. For non-filtered data D_1, the AIC values of all the three models for Periods I and II are fairly close, and thus we choose the simpler model M_2. Since the AIC value of M_2 in Period IV is the smallest, we choose M_2 as the best fitting model for Period IV. For SNORT-filtered D_2, the AIC values of all the three models in Periods I, II and V are fairly close, and therefore we choose the simpler M_2 as the fitting model. For Period IV, models M_3 and M_4 have smaller AIC values that are fairly close to each other, and therefore we choose M_3 as the best fitting model.

3.4.3 Effect of SNORT-filtering on extreme attack rates

The left-hand column in Figure 3.3 plots extreme attack rates with respect to the threshold quantile values identified in Tables 3.1-3.2. We observe that SNORT-based filtering does not change the distributions/models of extreme attack rates in Periods I, II and III. This is because as shown in Tables 3.1-3.2, Period I in D_1 and D_2 is fitted by M_2, Period II in D_1 and D_2 is fitted by M_2, and



(a) Period I: $(0.85, 0.85)$ (b) Period I: $(0.85, 0.85)$ (c) Period II: $(0.80, 0.80)$ (d) Period II: $(0.80, 0.80)$

(e) Period III: $(0.90, 0.95)$ (f) Period III: $(0.90, 0.95)$ (g) Period IV: $(0.80, 0.85)$ (h) Period IV: $(0.80, 0.85)$

(i) Period V: $(0.95, 0.85)$ (j) Period V: $(0.95, 0.85)$

Figure 3.3: Time series plots of extreme attack rates (left-hand column) and accumulative number of zero-length intervals between extreme attack rates (right-hand column) with respect to threshold quantiles (q_1, q_2), where q_1 is the threshold quantile for D_1 and q_2 is the threshold quantile for D_2. SNORT-based filtering does not make significant difference to the thresholds for D_1 and D_2, meaning that the extreme-value phenomenon is robust and is not destroyed by SNORT-based filtering. The implication is (see text for reasoning): When the attacker launch intense new attacks (i.e., attacks that are not detected), the distribution of the extreme attack rates observed at the honeypots are about the same as the distribution of the extreme attack rates with respect to a production network, provided that the attacker launch attacks without differentiating honeypots and production networks. This distribution information can be used to adjust the defense at production networks (e.g., using resource-consuming behavior-based detection, rather than resource-effective signature-based detection).

Period III in D_1 and D_2 is fitted by M_1. This can be visually confirmed because the red-colored curves and the green-colored curves in Figures 3.3a, 3.3c and 3.3e exhibit similar patterns. One the other hand, SNORT-based filtering does have an impact on Periods IV and V. This is because as shown in Tables 3.1-3.2, Period IV in D_1 is fitted by M_2 but in D_2 is fitted by M_3, and Period V in D_1 is fitted by M_1 but in D_2 is fitted by M_2. This also can be visually confirmed because SNORT filtered the highest spikes in Figures 3.3g and 3.3i. Nevertheless, all these distributions are in the GPD family.

The right-hand column in Figure 3.3 plots the accumulative numbers of zero-length intervals between extreme attack rates, meaning that intense attacks sustain for multiple hours. For each period, we observe that for both D_1 and D_2, there are many zero-length intervals. Now we look into the most often seen attacks in D_1 that are recognized by SNORT during the longest consecutive hours of intense attacks (i.e., largest clusters of extreme attack rates). Period I has 16 consecutive hours of intense attacks (from the 129th hour to the 144th hour). Each of the 16 consecutive hours has one of the following attacks as the most intense attack (as identified by SNORT): (i) denial-of-service vulnerability exploit, (ii) Worm spread attempt, (iii) Apache Buffer Overflow Vulnerability exploit, (iv) MS SQL exploit attempt, and (v) SSH Scan. Period II has 17 consecutive hours of intense attacks (from the 322th to the 338th hour). Each of the 17 consecutive hours has one of the following attacks as the most intense attack: (i) denial-of-service vulnerability exploit, and (ii) Worm spread attempt. Period III has 24 consecutive hours of intense attacks (from the 1st hour to the 23th hour). Each of the 24 consecutive hours has the following attack as the most intense attack: (i) denial-of-service vulnerability exploit. Period IV has 65 consecutive hours of intense attacks (from the 76th hour to the 140th hour). During majority of the 65 hours, the most often seen attack is (i) denial-of-service vulnerability, (ii) Worm spread attempt, and (v) SSH scan. Period V has 18 consecutive hours of intense attacks (from the 1882th hour to the 1899th hour). The most often seen attacks during the 18 hours are (i) denial-of-service vulnerability exploit, (ii) Worm spread attempt, (iv) MS SQL exploit attempt, and (v) SSH scan.

The preceding analysis shows that denial-of-service is the most often detected attack (by

SNORT) during the longest consecutive hours of intense attacks (i.e., the largest cluster of extreme attack rates) in each of the five periods. This makes us wonder whether this is also true for the hour of largest attack rate (i.e., highest spike). In order to answer this question, we look into the attacks that are detected by SNORT during the hour of the largest attack rate. The left-hand column of Figure 3.4 plots the number of attackers vs. the number of attacks they launch. We observe that a significant number of attackers only launched a single attack, but a small number of attackers launched a large number of attacks. This hints the possibility of power-law distribution. A further statistical analysis (details omitted) concludes that Periods I, IV and V in both D_1 and D_2 follow power-law distributions, but Periods II and III in both D_1 and D_2 do not follow the power-law distribution. The right-hand column of Figure 3.4 further plots the five most often seen attacks (as detected by SNORT) corresponding to the highest spike in D_1. Moreover, the most frequently seen attack is detected by SNORT as “Apache buffer overflow vulnerability" for Period I and “suspicious FTP scan" for Period IV, which corresponds to majority or even more than 95$\%$ of the detected attacks (as shown in Figures 3.4b and 3.4h). Therefore, denial-of-service is not necessarily the most often seen attack during the hour of largest attack rate; instead, it can be buffer-overflow (Period I).

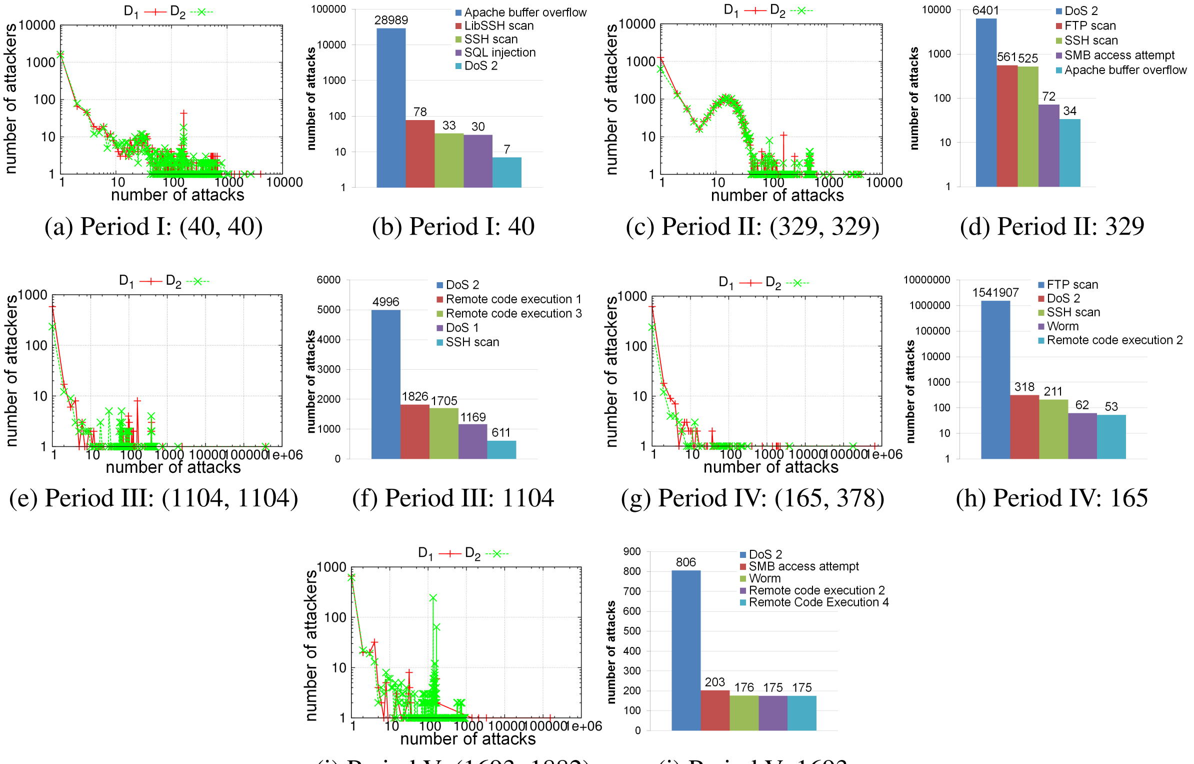
3.5 TST-based Extreme-Value Analysis

Now we study how TST-based models can fit the extreme values.

Since the attack-rate time series exhibit the LRD phenomenon [77] and the extreme-value phenomenon, we need models that can accommodate both. Since the GARCH model can accommodate the extreme-value phenomenon [27] and the FARIMA model can accommodate LRD, we propose to use the following FARIMA+GARCH model:

$$\phi(B)(1-B)^d(y_t-\mu_t)=\psi(B)\epsilon_t,$$

where ϕ and ψ are the same as in Section 3.2.2, B is the lag operator, $(1-B)^d$ is the LRD process



(a) Period I: (40, 40)

(b) Period I: 40

(c) Period II: (329, 329)

(d) Period II: 329

(e) Period III: (1104, 1104)

(f) Period III: 1104

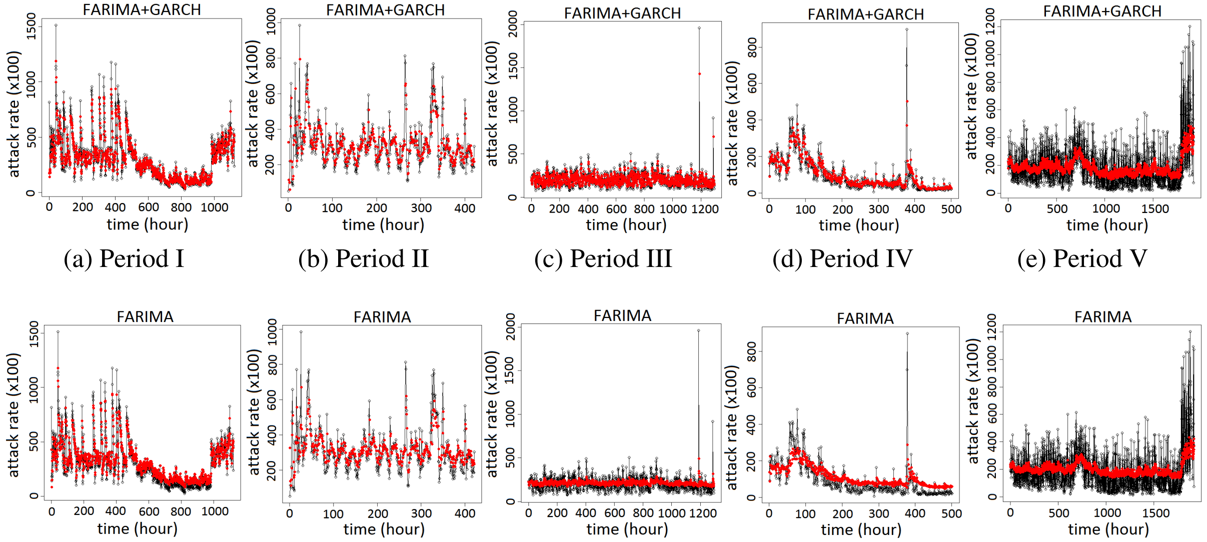
(g) Period IV: (165, 378)

(h) Period IV: 165

(i) Period V: (1693, 1882)

(j) Period V: 1693

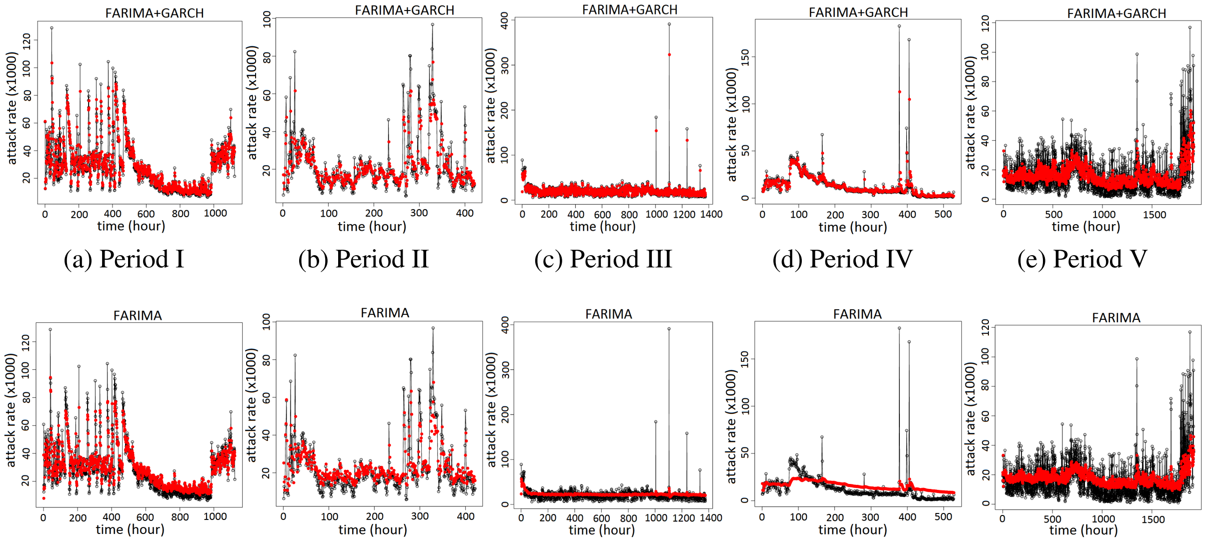
Figure 3.4: Left-hand column: number of attackers vs. the number of attacks they launch (as detected by SNORT) with respect to (a, b), where a indicates that the highest spike in D_1 occurs at the ath hour, and b indicates that the highest spike in D_2 occurs at the bth hour. Further analysis shows that Periods I, IV and V in both D_1 and D_2 follow power-law distributions (i.e., many attackers only launch small number of attacks and a small number of attackers launch many attacks), but Periods II and III in both D_1 and D_2 do not follow any power-law distribution. Right-hand column: The five most frequently seen attacks as detected by SNORT with respect to a, where a indicates the highest spike occurs at the ath hour in D_1, “DoS 1" is short for “Microsoft Windows SSL library denial-of-service vulnerability", “DoS 2" is short for “SIP (Session Initiation Protocol) invite message remote denial-of-service vulnerability", “Remote code execution 1/2/3/4" are respectively short for “Microsoft MS06-041/MS06-040/MS05-021/MS08-067 vulnerabilities."



(a) Period I (b) Period II (c) Period III (d) Period IV (e) Period V

(f) Period I (g) Period II (h) Period III (i) Period IV (j) Period V

Figure 3.5: TST-based model fitting of non-filtered attack rate data D_1, where black circles represent the observed attack rates and red dots represent the fitted values. FARIMA+GARCH fits Periods I-III better than FARIMA (especially for the extreme attack rates), but not Periods IV-V.



(a) Period I (b) Period II (c) Period III (d) Period IV (e) Period V

(f) Period I (g) Period II (h) Period III (i) Period IV (j) Period V

Figure 3.6: TST-based model fitting of SNORT-filtered attack rate data D_2, where black circles represent the observed attack rates and red dots represent the fitted values. FARIMA+GARCH fits Periods I-III better than FARIMA (especially the extreme attack rates) but not Periods IV-V.

with Hurst parameter H satisfying $0 < d = H - .5 < 1$, and $\mu_t = \mu + \xi\sigma_t$ with variance σ_t following SGARCH (i.e., Standard GARCH) or IGARCH (i.e., Integrated GARCH) with noise distribution SSTD or SGED (as reviewed in Section 3.2.2). This actually leads to a family of FARIMA+GARCH models:

- M'_1: FARIMA+SGARCH+SSTD;
- M'_2: FARIMA+SGARCH+SGED;
- M'_3: FARIMA+IGARCH+SSTD;
- M'_4: FARIMA+IGARCH+SGED.

For comparison, we also consider the FARIMA mode, which can better fit/predict the time series than the LRD-less ARMA model [77]. To select the best fitting model, we use two model selection criterions: PMAD (Percent Mean Absolute Deviation) and AIC. To select the best prediction model, we use PMAD. Suppose $X_m, \ldots, X_h$ are the observed attack rates and $X'_m, \ldots, X'_h$ are the fitted (predicted) attack rates. we have $\mathrm{PMAD} = \sum_{t=m}^{m+h} |X_t - X'_t| / \sum_{t=m}^{m+h} X_t$, which captures the overall fitting (prediction) error (i.e., the smaller the PMAD value, the better the model fitting or prediction).

Table 3.3 summarizes the fitting results. For Periods I-III in D_1, we observe that FARIMA+GARCH has the smaller AIC values as well as the smaller PMAD values (all < 0.2). For Periods IV-V in D_1, FARIMA+GARCH still has the smaller PMAD and AIC values, but the PMAD values are greater than 0.3. Therefore, FARIMA+GARCH can fit Periods I-III better than FARIMA, but not Periods IV-V. For D_2, we can draw the same conclusion.

To have a better understanding on what caused the fitting inaccuracy, Figures 3.5-3.6 plot the actual fitting results of D_1 and D_2, respectively. We observe that FARIMA+GARCH can indeed fit the data (especially the extreme attack rates) better than FARIMA. For Periods I-III in both D_1 and D_2, we observe that the slight inaccuracy of FARIMA+GARCH is mainly caused by that the extreme values (i.e., spikes) are not fitted 100%.

Table 3.3: TST-based fitting of attack rates (per hour).

Period	FARIMA + GARCH			FARIMA	
	model	PMAD	AIC	PMAD	AIC
D_1: non-filtered attack-rate time series					
I	M_1'	**0.170**	11.0	0.192	11.8
II	M_1'	**0.185**	11.4	0.195	11.9
III	M_3'	**0.196**	10.7	0.239	11.3
IV	M_3'	0.363	9.6	0.439	10.7
V	M_4'	0.441	11.8	0.482	12.0
D_2: SNORT-filtered attack-rate time series					
I	M_1'	**0.159**	10.6	0.188	11.6
II	M_4'	**0.197**	10.8	0.239	11.6
III	M_3'	**0.211**	10.5	0.297	12.0
IV	M_3'	0.258	9.3	0.545	11.6
V	M_3'	0.426	11.5	0.477	11.8

For Period IV, Figures 3.5d-3.6d show that FARIMA+GRACH should fit D_1 and D_2 well enough. However, the significant fitting inaccuracy (as shown in Table 3.3) is still caused by the missing of the extreme attack rates. For Period V, FARIMA+GARCH clearly cannot fit D_1 and D_2 well, possibly because there exist some complex statistical patterns that change rapidly. We are not aware of any readily available statistical tools that can cope with such time series, and expect to develop some advanced tools for this purpose in the future.

3.6 EVT- and TST-based Joint Analysis

For prediction, we consider all the models rather than the best fitting models only, because the best fitting models often are, but not always, the best prediction models.

3.6.1 EVT-based prediction of extreme attack rates

We use EVT-based methods to predict the return level, which gives the expected *magnitude* of extreme attack rates (but not necessarily the expected *maximum* attack rate) within a future period of time. This is the best EVT can predict. In order to evaluate the accuracy of the return-level predictions, we use the last 120 hours in each period for prediction. We use Algorithm 3.3 to predict return levels for return period $h = 24$ hours ahead of time, where ℓ is chosen such that $m = \lfloor n\ell \rfloor = n - 120$. We use the *binomial test* [43] to measure the prediction accuracy, such that the p-value greater than .05 means that prediction is accurate. The prediction results are described in Table 3.5 and discussed below (for comparison with TST-based predictions).

Algorithm 3.3 EVT-based prediction of return levels

```
INPUT: extreme attack rates $\{X_1, \ldots, X_n\}$ with respect to threshold quantiles described in Tables 3.1-3.2, EVT model family $\{M_1, M_2, M_3, M_4\}$ described in Section 3.4, $0 < \ell < 1$, $h$ (# of hours as return period)
OUTPUT: prediction model $M_j \in \{M_1, M_2, M_3, M_4\}$
$m = \lfloor n\ell \rfloor$
for $i = 1$ to $4$ do
    while $m + h \leq n$ do
        Using $\{X_1, \ldots, X_m\}$ to fit the $M_i$-type model
        Use $M_i$ to predict the return level between the $(m+1)$th and the $(m+h)$th hours
        $m = m + h$
    Evaluate prediction accuracy using the binomial test
return $M_j \in \{M_1, M_2, M_3, M_4\}$ with the highest $p$-value (or simpler model with the same $p$-value)
```

3.6.2 TST-based prediction of (extreme) attack rates

We showed (in Table 3.3) FARIMA+GARCH provides better fitting than FARIMA. Now we investigate the prediction power of FARIMA+GARCH. We use Algorithm 3.4 to find the best prediction model $M'_j \in \{M'_1, \ldots, M'_4\}$, where we use the last 100 hours for h-hour ahead prediction. In order to select the h that leads to best predictions, we consider $h = 1, 4, 7, 10$, which lead to different values for ℓ in Algorithm 3.4.

Table 3.4 summarizes the prediction results of the best model. For D_1, we observe that for Periods I-III and IV, the PMAD value of $h = 1$ in each period is the smallest among the four prediction resolutions: $h = 1, 4, 7, 10$. In particular, for Periods I-III, the PMAD value of 1-hour ahead prediction is smaller than 0.2, indicating accurate prediction. However, Periods IV and V have PMAD values 0.339 and 0.378 at 1-hour ahead prediction, meaning that the predictions are inaccurate. For D_2, we have similar observations. Therefore, we will compare TST-based *1-hour*

```
Algorithm 3.4 TST-based prediction of attack rates
INPUT: attack-rate time series $\{X_1, \dots, X_n\}$, FARIMA-GARCH model family $\{M'_1, M'_2, M'_3, M'_4\}$, $0 < \ell < 1$, $h$ (# of hours ahead prediction)
OUTPUT: best prediction model $M' \in \{M'_1, \dots, M'_4\}$
for $i = 1$ to $4$ do
    $m = \lfloor n\ell \rfloor, j = 0$
    while $m + h \leq n$ do
        use $\{X_1, \dots, X_m\}$ to fit a $M'_i$-type model
        use $M'_i$ to predict attack rates $\{X'_{m+1}, \dots, X'_{m+h}\}$
        $m = m + h$
    evaluate PMAD value of the predictions
return $M' \in \{M'_1, M'_2, M'_3, M'_4\}$ with the smallest PMAD values
```

ahead predictions of concrete attack rates with EVT-based *24-hour ahread* predictions of expected magnitude of extreme attack rates. This is summarized in Table 3.5 and discussed below.

Table 3.4: TST-based h-hour ahead predictions: $h = 1, 4, 7, 10$.

Period	ℓ	Selected Model	PMAD			
			h=1H	h=4H	h=7H	h=10H
D_1: non-filtered attack-rate time series						
I	0.90	M'_3	**0.138**	0.172	0.255	0.300
II	0.70	M'_4	**0.121**	0.343	0.390	0.386
III	0.90	M'_3	**0.140**	0.276	0.316	0.282
IV	0.80	M'_3	0.339	0.409	0.535	1.152
V	0.95	M'_3	0.378	0.388	0.470	0.288
D_2: SNORT-filtered attack-rate time series						
I	0.90	M'_2	**0.133**	0.204	0.269	0.302
II	0.70	M'_1	**0.232**	0.480	0.545	0.469
III	0.90	M'_1	**0.154**	0.299	0.386	0.221
IV	0.80	M'_3	0.436	1.062	1.084	1.415
V	0.95	M'_3	0.346	0.494	0.589	0.319

3.6.3 Making use of EVT- and TST-based predictions

Table 3.5 reports EVT-predicted return levels as well as the corresponding p-values of binomial test, the observed maximum attack rates, and the TST-predicted maximum attack rates as well as the corresponding PMAD values. We observe the following. First, EVT-based best prediction models (described in Table 3.5) are respectively simpler than the EVT-based best fitting models

Table 3.5: Comparison between EVT- and TST-based predictions, where "Ha-b" means the predictions correspond to time interval between the ath and the bth hours (among the last 120 hours of each period). For each period, there are three rows: The first row represents EVT-based predictions of return levels (i.e., expected magnitude of extreme attack rates) and the corresponding p-value of the predictions, where prediction model M_j is selected by Algorithm 3.3. The second row, denoted by "obs.", describes the observed (i.e., actual) maximum attack rates. The third row describes the maximum attack rates derived from TST-predicted attack rates by model M_j' (selected by Algorithm 3.4 with $h = 1$) and the corresponding PMAD value.

Per.		H1-24	H25-48	H49-72	H73-96	H97-120	p or PMAD
D_1: Non-filtered attack-rate time series							
I	M_2	76056	76824	77088	78333	82275	**0.07**
	obs.	53197	60203	57370	62868	82576	
	M_3'	50656	53744	52427	58183	71719	**0.13**
II	M_2	60910	60668	63572	62750	60752	**0.17**
	obs.	83157	101937	45186	38218	60274	
	M_4'	72457	83073	43942	37267	51853	**0.18**
III	M_1	33263	32916	32836	32733	32654	**0.07**
	obs.	32993	30194	21476	92379	29722	
	M_3'	30869	28505	21382	77747	27921	**0.21**
IV	M_2	29747	29622	28622	30048	30514	0.01
	obs.	23224	11671	6634	6263	8225	
	M_3'	*20329*	*9443*	*6674*	*5341*	*5605*	0.40
V	M_1	40129	41265	42360	43526	45996	0.00
	obs.	101971	105171	114462	120221	109216	
	M_3'	40341	43581	50053	54644	55007	0.41
D_2: SNORT-filtered attack-rate time series							
I	M_1	69002	68472	67849	67318	67105	**0.07**
	obs.	45123	44673	50116	56714	69795	
	M_2'	39125	40730	42293	49227	58758	**0.13**
II	M_1	50621	53794	61225	59442	58960	**0.33**
	obs.	74838	96757	37897	23722	53100	
	M_1'	60426	82960	37656	22967	44048	**0.22**
III	M_1	35373	35232	34998	34894	34922	**0.07**
	obs.	27783	25717	18412	76543	24815	
	M_1'	26315	24514	17880	66544	23557	**0.22**
IV	M_2	25692	25605	24355	24546	24929	0.01
	obs.	19420	9221	6634	4949	6317	
	M_3'	*22525*	*9772*	*5791*	*4731*	*5558*	0.43
V	M_1	34795	35782	36538	37437	40006	0.00
	obs.	60703	17634	18132	52651	35481	
	M_3'	37864	37935	42951	53202	48988	0.41

(described in Tables 3.1-3.2). However, TST-based best prediction models (described in Table 3.5) are respectively the same as the TST-based best fitting models (described in Table 3.4). This means that the presence of defense may simplify the prediction of *extreme* attack rates, but not necessarily the prediction of attack rates.

Second, we observe consistency between the predictions based on the two approaches. Specifically, EVT-predicted return levels (i.e., expected magnitude of extreme attack rates) are accurate for Periods I-III in both D_1 and D_2 because the p-values are greater than 0.05, while TST-predicted attack rates are also accurate for Periods I-III in both D_1 and D_2 because the PMAD values are smaller than or equal to 0.22. On the other hand, EVT-predicted return levels are inaccurate for Periods IV-V in both D_1 and D_2 because the p-values are smaller than 0.02, and TST-predicted attack rates are also inaccurate for Periods IV-V in both D_1 and D_2 because the PMAD values are greater than or equal to 0.40. However, there is a significant difference between Periods IV and V. For Period IV in both D_1 and D_2, we observe the TST-predicted maximum attack rates, which are extracted from TST-predicted attack-rate time series plotted in Figures 3.7g-3.7h, are actually accurate with respect to the observed maximum attack rates. This means that although TST-based predictions are not accurate overall, their predicted maximum attack rates can be accurate. This is useful because the predicted maximum attack rate can be the most important factor for the defender's resource allocation decision-making. Unfortunately, TST-predicted maximum attack rates are inaccurate for Period V. A possible cause for Period V is that there may exist some complex time series pattern, such as cyclical trends or seasonal trends (i.e., some repeated patterns).

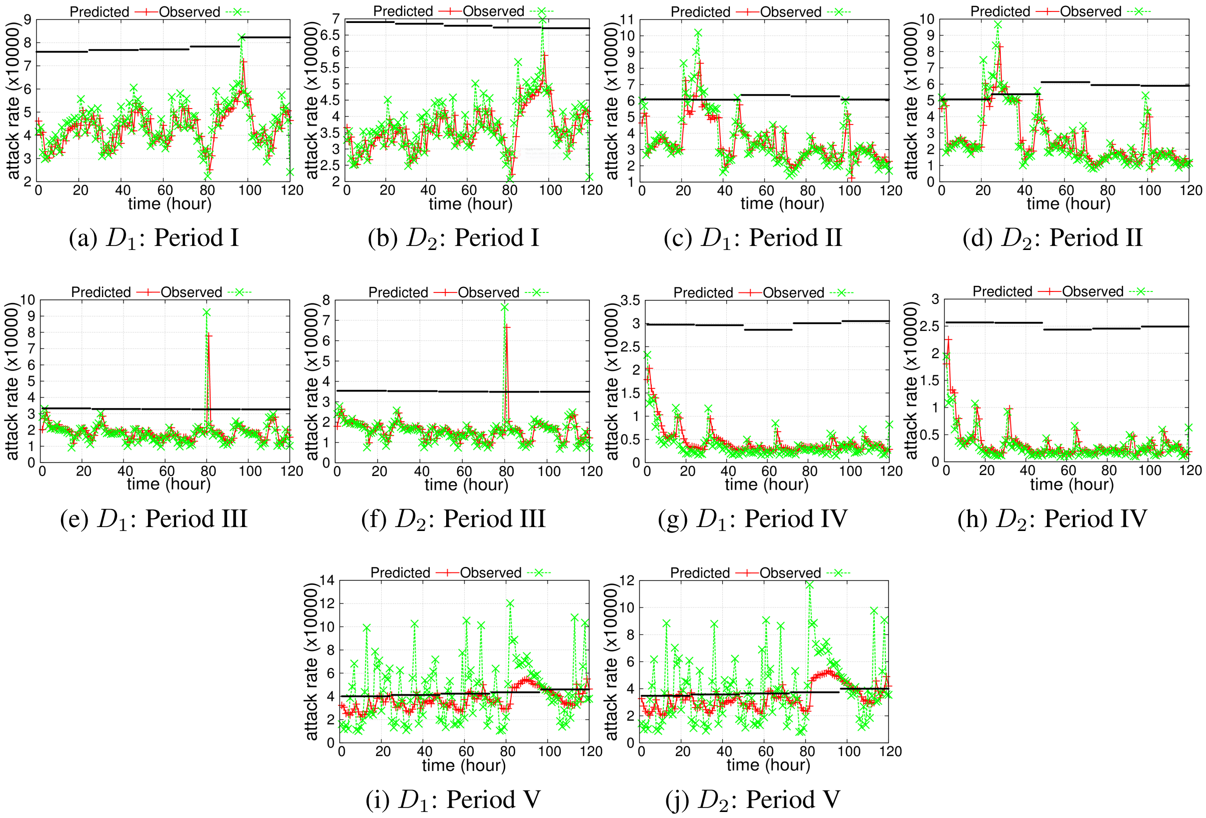
Third, for Periods I-III where both EVT and TST provide overall accurate predictions, we observe that EVT-predicted return levels, which are given to the defender 24 hours ahead of time, can be used as an evidence for resource allocation. Moreover, EVT-based resource allocations could be dynamically adjusted by taking into account TST-based 1-hour ahead predictions. Specifically, Figure 3.7 suggests the following: when a TST-predicted maximum attack rate (which is obtained only 1 hour ahead of time) is above the EVT-predicted return level (which is obtained 24 hours ahead of time), the defender can dynamically allocate further resources for the anticipated attack rates predicted by TST-based methods (e.g., the highest spikes in Periods II-III as shown in Figure 3.7). This strategy is of course conservative because it (on average) overprovisions resources so as to cope with the worst-case scenario (i.e., matching the largest attack rates). Nevertheless, this strategy gives the defender more earlywarning time. An alternate strategy is to use TST-based predicted maximum attack rates as the initial evidence for allocating defense resources, then take into consideration EVT-based long-term predictions (e.g., some weighted average). This strategy might prevent resource overprovisioning, but may not provision sufficient resources to cope with the largest attack rates (e.g., the highest spike in Period III as shown in Figure 3.7). This strategy requires the defender to be more agile (than in the preceding strategy).

3.7 Limitations and Future Research Directions

First, the analysis results are limited by the datasets. Nevertheless, the analysis results are sufficient for justifying the value of the methodology and the newly introduced family of FARIMA+GARCH models, which are equally applicable for analyzing other datasets (e.g., larger datasets, or datasets collected by high-interaction honeypots).

Second, the connection between EVT and TST is a good starting point. Especially, we observed that EVT-predicted return levels are often above the actual maximum attack rates, but TST-predicted maximum attack rates are often below the actual maximum attack rates. We suggested the possibility of using some weighted average of EVT-predicted return level and TST-predicted maximum attack rate as the predicted maximum attack rate. This heuristics needs to be justified rigorously. Moreover, there might be some deeper connections that can be exploited to formulate more powerful prediction techniques. Finally, there may be some fundamental trade-off between the earlywarning time we can give to the defender and the prediction accuracy. These connections have not be investigated by the theoretic statistics community, and our engineering-driven demand would give statistical theoreticians enough motivation to explore this exciting topic.

Third, Period V cannot be fitted and predicated accurately (even for maximum attack rates only) possibly because there exist some properties other than LRD and extreme events. Further



(a) D_1: Period I (b) D_2: Period I (c) D_1: Period II (d) D_2: Period II

(e) D_1: Period III (f) D_2: Period III (g) D_1: Period IV (h) D_2: Period IV

(i) D_1: Period V (j) D_2: Period V

Figure 3.7: Comparing EVT-predicted return levels (i.e., expected magnitudes of extreme attack rates), observed attack rates during the last 120 hours in each period, and TST-based predictions of attack rates. EVT-predicted return levels are produced by Algorithm 3.3 and summarized in Table 3.5, and plotted as horizontal lines during the respective intervals of 24 hours. TST-based predictions are produced by Algorithm 3.4. For Periods I-III, EVT-predicted return levels are accurate, and TST-predicted attack rates as well as *maximum attack rates* are also accurate. For Period IV, EVT-predicted extreme attack rates are about one order of magnitude above the observed attack rates, but TST-predicted maximum attack rates are accurate. For Period V, neither EVT nor TST can predict accurately.

studies are needed for exploring if these are some *seasonal* or *cyclical* trends or the extreme values are generated by some self-exciting process [44].

3.8 Conclusions

We have presented a novel methodology for analyzing the extreme-value phenomenon exhibited by (honeypot-captured) cyber attacks. Our methodology is based on a novel integration of EVT (Extreme Value Theory) and TST (Time Series Theory), and can be seamlessly incorporated into the framework we recently proposed [77]. For TST-based analysis, we proposed a family of FARIMA+GARCH models for fitting and predicting both stationary and non-stationary time series. We believe that this study will inspire other researchers to devise a complete families of statistical frameworks and techniques that can adequately satisfy the needs of cyber security.

Chapter 4: ANALYZING CYBER SECURITY POSTURE

4.1 Introduction

Blackhole [14, 22] (aka darknet [12], network telescope [49], network sink [75]) is a useful instrument for monitoring unused, routeable IP address space. Since there are no legitimate services associated to these unused IP addresses, traffic targeting them is often caused by attacks. This allowed researchers to use blackhole-collected data (together with other kinds of data) to study, for example, worm propagation [13, 46, 48] and denial-of-service (DOS) attacks [33, 47]. Despite that blackhole-collected data can contain unsolicited, but not necessarily malicious, traffic that can be caused by misconfigurations in remote computers or by Internet background radiation [30, 50, 74]. Analyzing blackhole-collected data could lead to better understanding of *cybersecurity posture* (i.e., security-related situation in Internet).

4.1.1 Our Contributions

In this chapter, we aim to empirically characterize the cybersecurity posture based a dataset collected by CAIDA's /8 blackhole during the month of March 2013. Our analysis emphasizes on (i) identifying interesting cybersecurity phenomena and (ii) explaining their (hypothetic) cause. We analyze both the "as is" data and the data obtained after heuristically filtering some rarely seen attackers (as an approximation to misconfiguration-caused traffic). Our findings are highlighted as follows. First, we define the notion of *sweep-time*, namely the time it takes for most blackhole IP addresses to be attacked at least once. We find that the sweep-time follows the power-law distribution. Second, we find that the *total* number of distinct attackers that are observed by the blackhole is largely determined by the number of distinct attackers from a certain country code.[1]

We expect to publish more detailed analysis and (hypothetical) explanations of the newly iden-

[1] We are fortunate to see the real, rather than anonymized, attacker IP addresses, which allow us to aggregate the attackers based on their country code. In this chapter, we will not disclose any specific IP address. Our study is approved by IRB.

tified phenomena elsewhere shortly.

4.1.2 Related Work

Investigations based on blackhole-collected data can be classified into two categories. The first category of studies analyze blackhole-collected data *only*. These studies include the characterization of Internet background radiation [50, 74], the characterization of scan activities [6], and the characterization of backscatter for estimating global DOS activities [33,47], The present study falls into this category as we analyze blackhole-data only. However, we aim to analyze cybersecurity posture especially attacks that are likely caused by malicious worm, virus and bot activities. This explains why we exclude the backscatter data (which is filtered as noise in the present chapter), with or without filtering the traffic that may be caused by misconfigurations. Due to the lack of ground truth — a fundamental limitation of blackhole-collected data, our analysis methodology allows to draw robust statistical conclusions.

The second category of studies aim to analyze blackhole traffic together with other kinds of relevant data. These studies includes using blackhole data and network-based intrusion detection and firewall logs to analyze Internet intrusion activities [76], using out-of-band informaiton to help analyze worm propagation [13, 46, 48], and using active interactions with remote IP addresses to filter misconfiguration-caused traffic [50]. Somewhat related studies include the identification of one-way traffic from data where two-way traffic is well understood [6, 17, 36, 68].

All these studies are loosely related to the effort of the present chapter, as we neither assume the availability of, nor use, any out-of-band information.

The rest of the chapter is organized as follows. Section 4.2 describes the data we analyze. Section 4.3 analyzes cybersecurity posture from the perspective of victims. Section 4.4 analyzes cybersecurity posture from the perspective of attackers.

Section 4.5 discusses the limitation of the present study. Section 4.6 concludes the chapter.

4.2 Data Description and Representation

4.2.1 Data Description

The data we analyze was collected between 3/1/2013 and 3/31/2013 by CAIDA's Blackhole, which is a passive monitoring system based on a globally routeable but unused /8 network (i.e., $1/256$ portion of the entire Internet IP v4 address space) [65]. Since blackhole collects unsolicited traffic, meaning that the collected traffic would contain *malicious traffic* that reaches the blackhole (e.g., automated malware spreading), but may also contain *non-malicious traffic* — such as Internet Background Radiation (e.g., backscatter caused by the use of spoofed source IP addresses that happen to belong to the blackhole) and misconfiguration-caused traffic (e.g. mistyping an IP address by a remote computer). This means that pre-processing the raw data is necessary. We will analyze D_1 and D_2 that are obtained after applying the pre-processing procedures described below.

Data D_1. Based on CAIDA's standard pre-processing [66], the collected IP packets are organized based on eight fields: source IP address, destination IP address, source port number, destination port number, protocol, TTL (time-to-live), TCP flags and IP length. The flows are classified into three classes: *backscatter*, *ICMP request* and *other*. At a high level, backscatter traffic is identified via TCP SYN+ACK, TCP RST, while ICMP request is identified via ICMP type 0/3/4/5/11/12/14/16/18. (A similar classification method is used in [74].) We are more interested in analyzing cybersecurity posture corresponding to attacks that are launched through TCP/UDP protocols. Since (i) backscatter-based analysis of DOS attacks has been studied elsewhere [33, 47] and (ii) ICMP has been mainly used to launch DOS attacks such as *ping flooding* and *smurf or fraggle* attacks [35, 47, 70], we disregard the traffic corresponding to *backscatter* and *ICMP request*. This means that we focus on the TCP/UDP traffic in the *other* category mentioned above. We call the resulting data D_1, in which each TCP/UDP flow is treated as a distinct attack.

Data D_2. Although (i) D_1 already excludes the traffic corresponding to *backscatter* and *ICMP request*, and (ii) we consider only TCP/UDP flows in the *other* category mentioned above, D_1

may still contain traffic caused by misconfigurations. Eliminating misconfiguration-cased traffic in blackhole-collected data is a hard problem because blackhole is passive, namely that blackhole does not interact with remote computers to collect rich information about attacks. Indeed, we have discussed in Section 4.1.2 (Related Work) that existing studies on recognizing misconfiguration-caused traffic had to use payload information (e.g., [40]), which is however beyond the scope of blackhole-collected data. Moreover, it is worth mentioning that recognizing one-way traffic already requires to using extra information such as two-way traffic [30], and that recognizing misconfiguration-caused traffic is an even harder problem because misconfiguration can cause both one-way *and* two-way traffics. These observations suggest us to use some heuristics to filter possible misconfiguration-caused traffic from D_1. Specifically, we obtain D_2 by filtering from D_1 the flows that correspond to remote IP addresses that initiated fewer than 10 flows/attacks during the month. This heuristic method filters possibly many misconfiguration-caused flows in blackhole-collected data, as well as possibly some number of malicious attacks. Even though the ground truth (i.e., which TCP/UDP flows correspond to malicious attacks) is not known — a fundamental limitation of balckhole, D_2 might be closer to the ground truth than D_1.

4.2.2 Data Representation

In order to analyze the TCP/UDP flow data D_1 and D_2, we represent the flows through time series at some *time resolution* from the perspectives of *victims* (i.e., blackhole IP addresses that are "hit" by some remote attacking IP addresses contained in D_1 or D_2), from the perspective of *attackers* (i.e., the remote attacking IP addresses contained in D_1 or D_2), and from the perspective of *attacks* (i.e., TCP/UDP flows initiated from remote attacking IP addresses in D_1 or D_2 are treated as attacks). We consider two time resolutions (because a higher resolution may lead to more precise statistics): hour, denoted by "H"; minute, denoted by "m". For a given time resolution of interest, the total time interval $[0, T]$ is divided into short periods $[i, i+1)$ according to time resolution $r \in \{H, m\}$, where $i = 0, 1, \ldots$, and $T = 744$ hours (or $T = 4,464$ minutes) in this case.

Let V be CAIDA's fixed set of blackhole IP addresses and A be the rest of IP addresses in

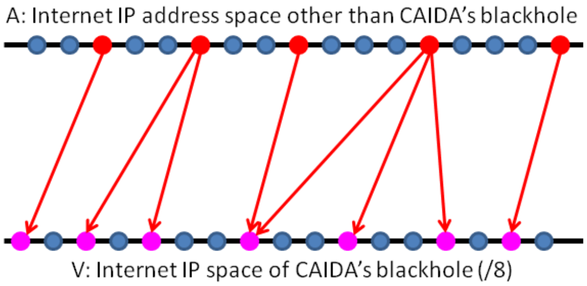

A: Internet IP address space other than CAIDA's blackhole
V: Internet IP space of CAIDA's blackhole (/8)

Figure 4.1: Illustration of attacker-victim relation during time interval $[i, i+1)$ at time resolution $r \in \{H, r\}$: each dot represents an IP address, a red-colored dot represents an attacking IP address (i.e., attacker), a pink-colored dot represents a blackhole IP address (i.e., victim), the number of attackers is $|A(r; i, i+1)| = 5$, the number of victims is $|V(r; i, i+1)| = 7$, and the number of attacks is $y(r; i, i+1) = 8$.

cyberspace, where $|A| = 2^{32} - |V|$. As illustrated in Figure 4.1, for each time interval $[i, i+1)$ at time resolution r, let $V(r; i, i+1) \subseteq V$ be the set of *distinct victims* that are attacked at least once during time interval $[i, i+1)$, $A(r; i, i+1) \subseteq A$ be the set of *distinct attackers* that launched attacks against some $v \in V(r; i, i+1)$, and $y(r; i, i+1)$ be the number of *distinct attacks* launched by the attackers belonging to $A(r; i, i+1)$ against victims belonging to $V(r; i, i+1)$. Note that a victim may be attacked by the same attacker via multiple attacks (i.e., flows), in which case we treat each flow as a distinct attack.

4.3 Characteristics of Sweep-Time

In this section, we ask and address the following question: How long does it take for most blackhole IP addresses to be attacked at least once? More precisely, we can naturally extend the notations introduced above as follows: For time resolution $r \in \{H, m\}$ and any time interval $[i, j)$ with $i + 1 < j$, we can naturally use $V(r; i, j) = \bigcup_{\ell=i}^{j-1} V(r; \ell, \ell+1)$ to represent the cumulative set of distinct victims that are attacked at some point during time interval $[i, j)$. As a result, $V(r; 0, T)$ is the set of distinct victims that are attacked at least once during time interval $[0, T)$, where $V(r; 0, T)$ is actually independent of the time resolution r but we keep r for notational consistence. It is possible that $V(r; 0, T) \approx V$, meaning that some blackhole IP address is never attacked during the time interval $[0, T)$. As such, the question we ask becomes: How long does it take for $\tau \times$

$|V(r;0,T)|$ victims to be attacked at least once, where $0 < \tau < 1$? This suggests us to define the following notion of *sweep-time*, which is relative to the starting point of observation time.

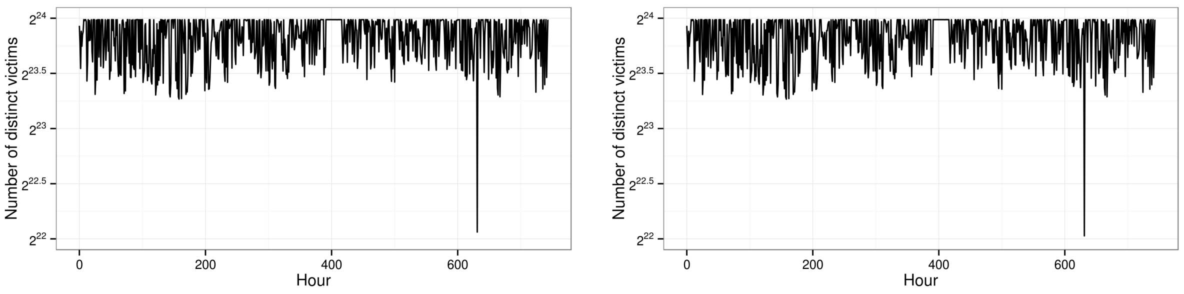
Definition 1. *The* sweep-time *starting at the ith time unit of time resolution r, denoted by I_i, is defined as:*

$$\left|\bigcup_{\ell=i}^{I_i-1} V(r;\ell,\ell+1)\right| < \tau \times |V(H;0,733)| \leq \left|\bigcup_{\ell=i}^{I_i} V(r;\ell,\ell+1)\right|.$$

By taking into consideration the starting point of observation time i, we naturally obtain a time series of sweep-time $I_0, I_1, \ldots$. Our focus is to characterize the time series of sweep-time I_i.

4.3.1 Distribution of Sweep-Time

For data D_1, we have $|V(H;0,744)| = 16,657,796 \approx 2^{23.99}$ and $|V(H;0,744)|/|V| = 16,657,796/2^{24} \approx 99.29\%$ of the entire blackhole IP address space are attacked at least once during the month. For data D_2, we have $|V'(H;0,744)| = 16,657,726 < |V(H;0,744)|$ because D_2 filtered some rarely seen attackers in D_1.

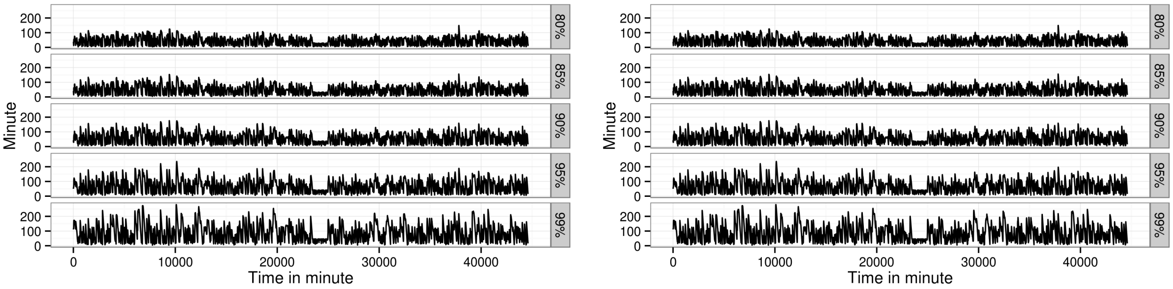


(a) Number of victims per hour in D_1: $|V(H;i,i+1)|$ (b) Number of victims per hour in D_2: $|V_1'(H;i,i+1)|$

Figure 4.2: Time series of the number of distinct victims per hour, namely $|V(H;i,i+1)|$ for $0 \leq i \leq 743$ corresponding to D_1 and $|V'(H;i,i+1)|$ for $0 \leq i \leq 743$ corresponding to D_2, where (for example) $|V(H;0,1)| > |V'(H;0,1)|$ and $|V(H;0,733)| > |V'(H;0,733)|$.

Figures 4.2a-4.2b present the times series of $|V(H;i,i+1)|$ in D_1 and the time series of $|V'(H;i,i+1)|$ in D_2, respectively. We make the following observations. First, there is a significant volatility at the 632nd hour, during which the number of distinct victims is as low as $4,377,079 \approx 2^{22}$. A careful examination shows that the total number of distinct attackers during the 632nd hour is very small, which might be the cause.

Second, most blackhole IP addresses are attacked within a single hour. For example, 15,998,907 (or $\tau = 96\%$ of $|V(H; 1, 733)|$ victims) blackhole IP addresses are attacked at least once during the first hour. Third, no victims other than $V(H; 0, 703)$ are attacked during the time interval $[704, 744)$ at the same time resolution.



(a) D_1: Time series of sweep-time at 1-minute resolution.

(b) D_2: Time series of sweep-time at 1-minute resolution.

Figure 4.3: Time series plots of sweep-time (y-axis) with respect to $\tau \in \{80\%, 85\%, 90\%, 95\%, 99\%\}$, where x-axis represents the starting observation time that is sampled at every 10 minute. In other words, the plotted points are the sample $(0, I_0), (10, I_{10}), (20, I_{20}), \ldots$ rather than $(0, I_0), (1, I_1), \ldots$.

Since we have observed that the sweep-time is often not exactly 1 or 2 hours, we use finer-grained time resolution, namely per-minute (rather than per-hour) to measure the sweep-time. Figure 4.3 plots the time series of sweep-time $I_0, I_{10}, I_{20}, \ldots$ with respect to per-minute time resolution (we only consider this sample of $I_0, I_1, I_2, \ldots$ because the latter is too time-consuming). We want to know the distributions of the sweep-time. Our statistical tests show that the sweep-time exhibits power-law distributions. Specifically, Table 4.1 summarizes the power-law test statistics of the sweep-time. We observe the following. First, for both D_1 and D_2, all the α values are very large. For threshold $\tau = 80\%$ in D_1, x_{min} is 78 minutes and the number of power-law sweep-times is 475 (10.6% out of 4,462) meaning that all the 89.4% non power-law sweep-times in the range between 0 and 78 minutes. Also, as the threshold τ increases, the x_{min} values also increase as expected. However, the number of power-law sweep-times decreases as τ values increase, which means that power-law distribution only fits smaller portion of the data as τ increases. Second, for the same threshold, D_1 and D_2 have similar x_{min} values as well as similar numbers of power-law sweep-times, which means that the "noise" traffic in D_1 does not affect the power-law property of D_1. For example, for $\tau = 80\%$, D_2 has x_{min} values of 82 minutes which is very close to 78 minutes in D_1. Moreover, D_2 has 391 (8.7% out of 4,462) power-law sweep-times which is close to 475 (10.6% out of 4,462) power-law sweep-times in D_1.

Table 4.1: Power-law test statistics of the sweep-time with respect to threshold $\tau \in \{80\%, 85\%, 90\%, 95\%, 99\%\}$, where α is the fitted power-law exponent ($\alpha \in (1, 2)$ means the mean and variance values does not exist; $\alpha \in (2, 3)$ means the mean value exist but the variance value does not exist), $x_{\min}$ is the cut-off parameter (i.e., we only consider the sweep-times that are greater than or equal to $x_{\min}$ units of time with respect to time resolution $r \in \{H, m\}$), KS is the Kolmogorov-Smirnov statistic for comparing fitted distribution with the input, $\# \geq x_{\min}$ represents the number of sweep-times that are used for fitting the distribution, which we refer as power-law sweep-times.

τ	α	$x_{\min}$	KS	p-value	$\# \geq x_{\min}$
Dataset D_1 with time resolution 1-minute					
80%	7.89	78	.05	.14	475
85%	8.46	94	.04	.52	385
90%	8.89	118	.06	.42	244
95%	9.52	148	.05	.68	193
99%	13.67	215	.04	.98	131
Dataset D_2 with time-resolution 1-minute					
80%	8.46	82	.05	.19	391
85%	8.37	95	.04	.36	379
90%	9.24	120	.05	.39	237
95%	12.82	170	.04	.99	72
99%	15.23	224	.04	.99	94

4.4 Dominance of the Number of Attackers from a Single Country

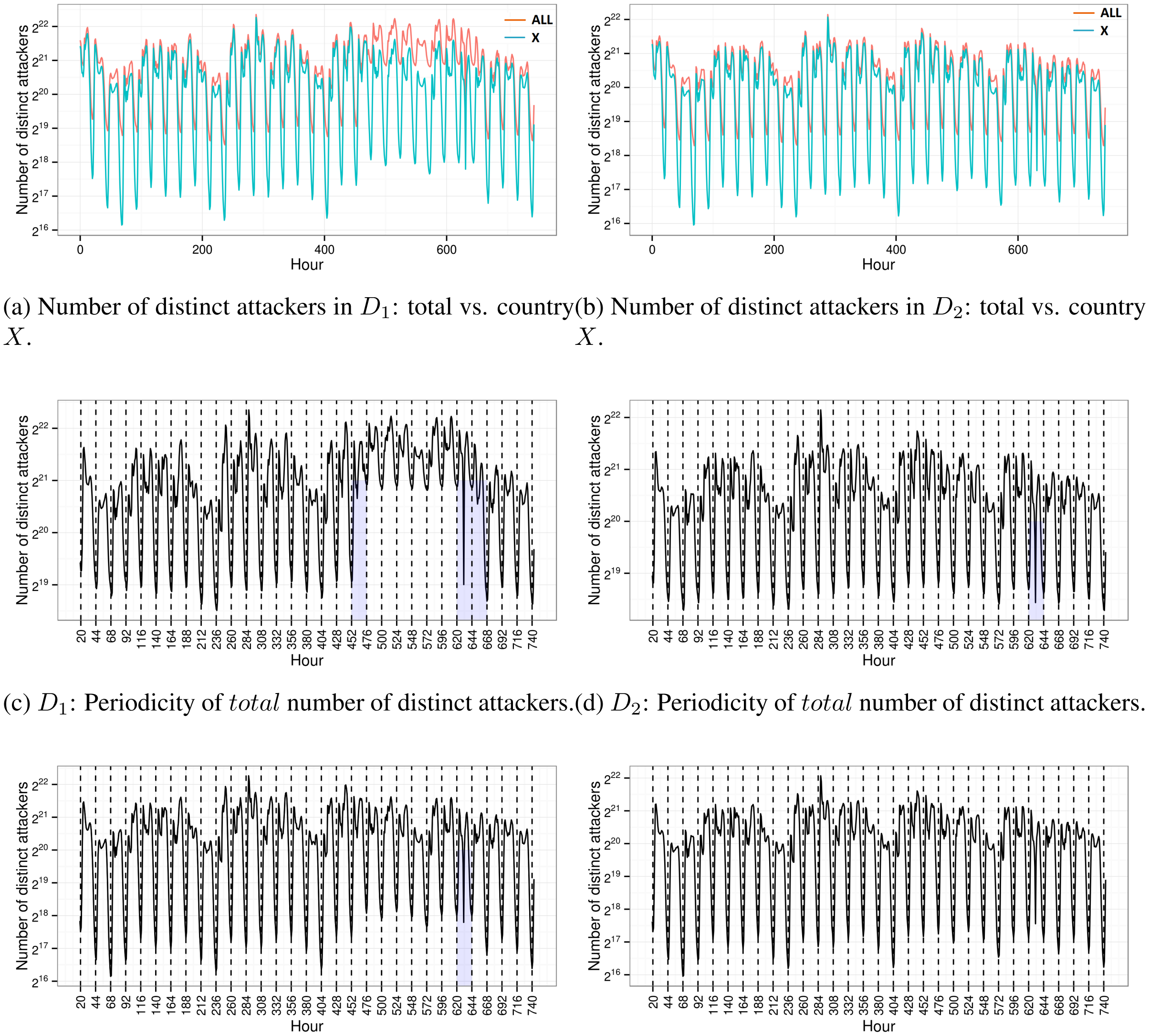
For each attacker IP address, we can use the WHOIS service to retrieve its country code. For an attacker IP address whose country code cannot retrieved from the WHOIS service, we use `NULL` instead. The term “total number of attackers" refers to all attackers, no matter whether their country code can be retrieved or not. This allows us to study the time series of the total number distinct attackers and the time series of the number of distinct attackers from individual countries. We report the following interesting phenomenon: The total number of distinct attackers observed by the blackhole during the month and the number of distinct attackers from a country (anonymized as country X) are surprisingly similar to each other.

4.4.1 The Informal Similarity Between Two Time Series

We report that for D_1, the top two countries, called X and Y, contribute 29.83% and 28.45% of the total number of 403,779,397 distinct attackers observed by the blackhole during the month, respectively. That is, countries X and Y together contribute almost 58% of all distinct attackers.

Specifically, Figure 4.4a compares the times series of the total number of distinct attackers observed by the blackhole and the time series of the number of distinct attackers from country X (per hour) in D_1. For the time series of the total number of distinct attackers, we observe that the number of distinct attackers fluctuates between 5,354,919 ($\approx 2^{22.35}$) and 373,183 ($\approx 2^{18.5}$) blackhole IP addresses. In particular, for time interval $[455, 630)$, namely during the period of 176 hours between the 455th hour (on March 19, 2013) and the 630th hour (on March 27, 2013), the number of distinct attackers (per hour) is greater than 1,651,184 ($\approx 2^{20.66}$) and can be up to 4,925,667 ($\approx 2^{22.23}$) attackers. More importantly, it seems that the total number of distinct attackers is largely determined by the number of distinct attackers from country X. This suggests us to plot Figures 4.4c and 4.4e.

Figure 4.4c and Figure 4.4e plot the time series starting from 20th hour which is the very first wave base and the vertical dashed lines are drawn every 24 hours. We observe that the wave bases are surprisingly periodic with period of 24 hours in both Figure 4.4c and Figure 4.4e. In other words, the wave bases appear at hours $20, 44, ..., 740$, which make 30 wave base ranges. Except for Figure 4.4e, the number of distinct attackers at the 631th hour is the smallest in the wave base range of [620, 644], which is marked in the blue rectangle. There are three such exceptions in Figure 4.4c. The number of distinct attackers at the 461th, the 631th and the 653th hours are among the smallest in the corresponding wave bases range covering them. After looking into the time zone of country X, we notice that the wave bases corresponding to the hour between 12:00 noon and 1 pm local time, meaning that least number of attackers are observed during that hour.



(a) Number of distinct attackers in D_1: total vs. country X.

(b) Number of distinct attackers in D_2: total vs. country X.

(c) D_1: Periodicity of $total$ number of distinct attackers.

(d) D_2: Periodicity of $total$ number of distinct attackers.

(e) D_1: Periodicity of number of distinct attackers from country X.

(f) D_2: Periodicity of number of distinct attackers from country X.

Figure 4.4: Time series of the total number of distinct attackers (per hour) and time series of the number of distinct attackers (per hour) from country X.

Analogous to Figures 4.4a, 4.4c and 4.4e that correspond to D_1, Figures 4.4b, 4.4d and 4.4f respectively correspond to D_2. We can see from Figure 4.4b that even after eliminating the rarely seen attackers from D_1, the total number of distinct attackers is still dominated by the number of distinct attackers from country X. Also, Figures 4.4d and 4.4f show the same 24-hour periodicity exhibited by the total number of distinct attackers and by the number of distinct attackers from country X. While Figures 4.4d has one exception at wave base range of $[620, 644]$, Figure 4.4f strongly suggests 24 hours periodicity because there is no exception like the ones shown in Figure 4.4e for D_1.

It is also interesting to zoom into each wave base range and see the patterns. We find that each wave base range demonstrates an "M" shape pattern as illustrated in Figure 4.5. Suppose we have a wave base range $[t_i, t_{i+24}]$ with two wave bases at t_i (i.e., 12:00 noon in country X) and t_{i+24}, we call the very first spike t_a and the very last change point t_b. We find that for each wave base range, $t_a - t_i \approx 5$ hours and $t_{i+24} - t_b \approx 10$ hours.

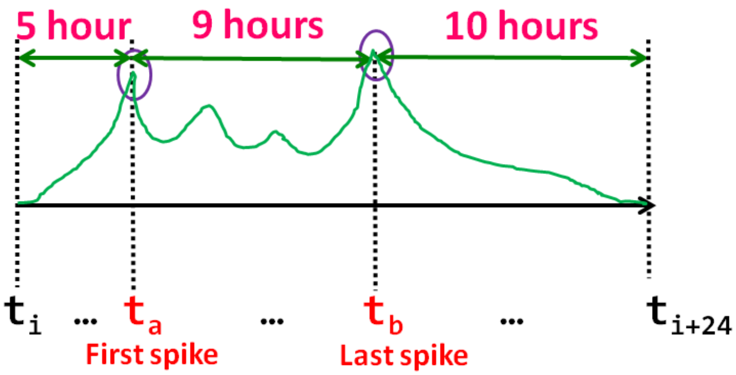


Figure 4.5: The "M" shape pattern within each periodic wave base range.

4.4.2 Formal Statistical Similarity Analysis Between the Two Time Series

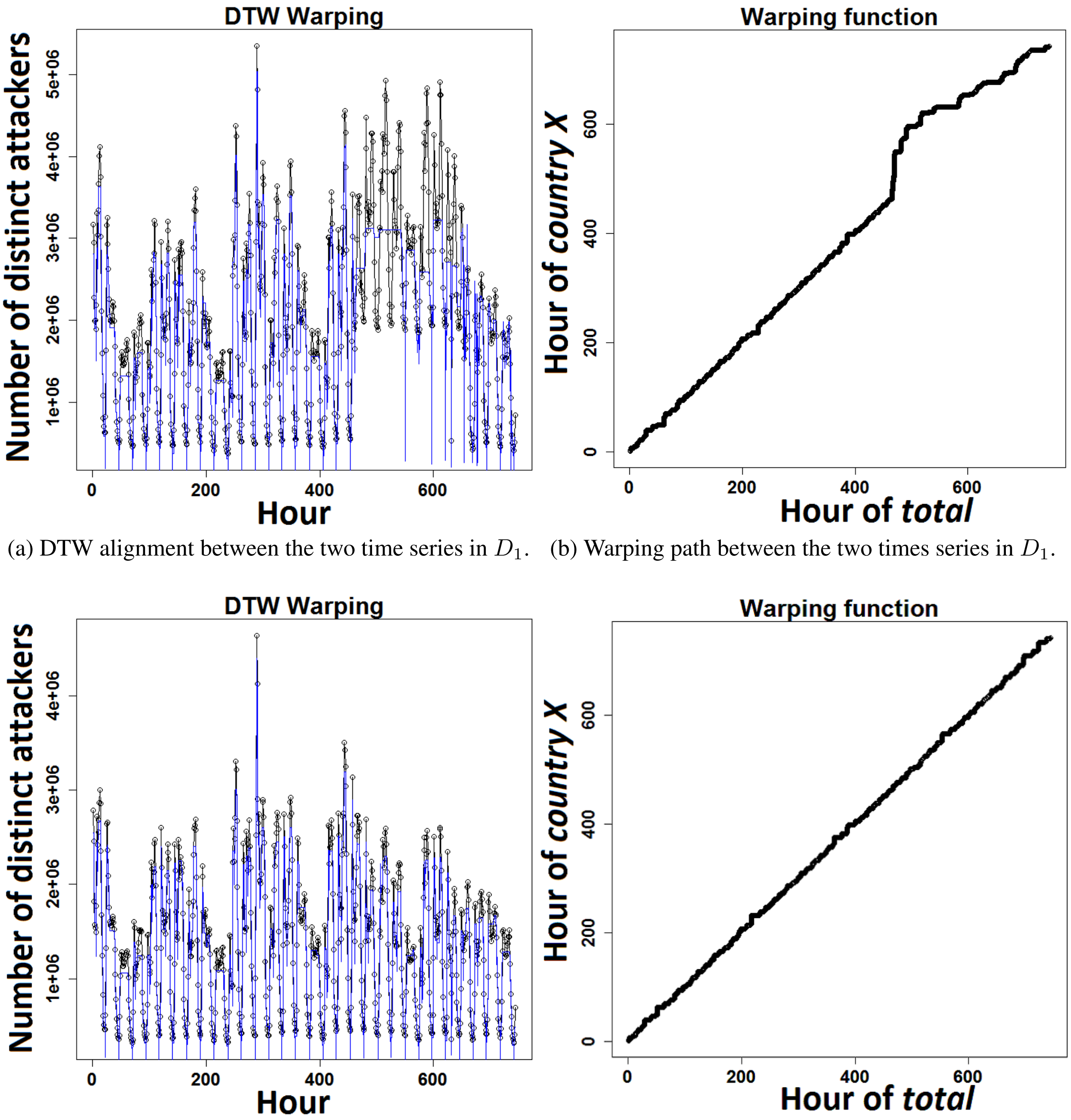
In the above we have observed the interesting phenomenon that (i) the time series of the number of distinct attackers from country X resembles (ii) the time series of the total number of distinct attackers. Now we use statistical methods to quantify the resemblance (similarity) between these two time series.

Similarity based on the distance between the two time series. We use the popular dynamic time warping (DTW) technique to quantify the optimal alignment between the two time series. DTW is often used to determine time series similarity, and find the corresponding regions between two time series. For the two time series corresponding to D_1, the alignment is displayed in Figure 4.6a. Each blue segment line connects a point in one time series to its corresponding point in the other time series. If the two time series are identical, all the connection lines would be straight and vertical because warping is not needed. As shown in Figure 4.6a, the two time series are very similar expect for the small region from the 470th to the 570th hours. Figure 4.6b shows the one-on-one mapping between the two times series. The ideal mapping is a straight line which can be described using linear regression function $y = x$. It can be seen that from the 470th to the 570th hours there is discrepancy between two time series.

For the two times series corresponding to D_2, we can see from Figures 4.6c and 4.6d that the two time series matches even better. This suggests the strong dominance of the number of distinct attackers from country X over the total number of distinct attackers that are observed by the blackhole. This also suggests that there are a significant number of *rarely seen* attackers that indeed might be caused by misconfigurations.

Similarity based on the fitted models. We use the multiplicative seasonal ARIMA model to fit the two time series corresponding to datasets D_1 and D_2. The model has nonseasonal orders (p, d, q), and seasonal orders (P, D, Q), and seasonal period s. The correlogram clearly indicates that there exists a very strong correlation at lag 24 in both data sets, which suggests that the seasonal model with $s = 24$ should be used. For model selection, the parameter sets are:

- $(p, d, q) \in [0, 5] \times \{0, 1\} \times [0, 5]$;
- $(P, D, Q) \in [0, 5] \times \{0, 1\} \times [0, 5]$.



(a) DTW alignment between the two time series in D_1.

(b) Warping path between the two times series in D_1.

(c) DTW alignment between the two time series in D_2.

(d) Warping path between the two time series in D_2.

Figure 4.6: DTW statistics between the times series of the total number of distinct attackers and the time series of the number of distinct attackers for country X.

According to the AIC criterion, both D_1 and D_2 and both time series prefer to model ARIMA$(1,0,1)\times(2,1,3)_{24}$, namely

$$W_t = \phi_1 W_{t-1} + e_t + \theta_1 e_{t-1} + \Phi_1 W_{t-24} + \Phi_2 W_{t-48} + \Theta_1 e_{t-24} + \Theta_2 e_{t-48} + \Theta_3 e_{t-96},$$

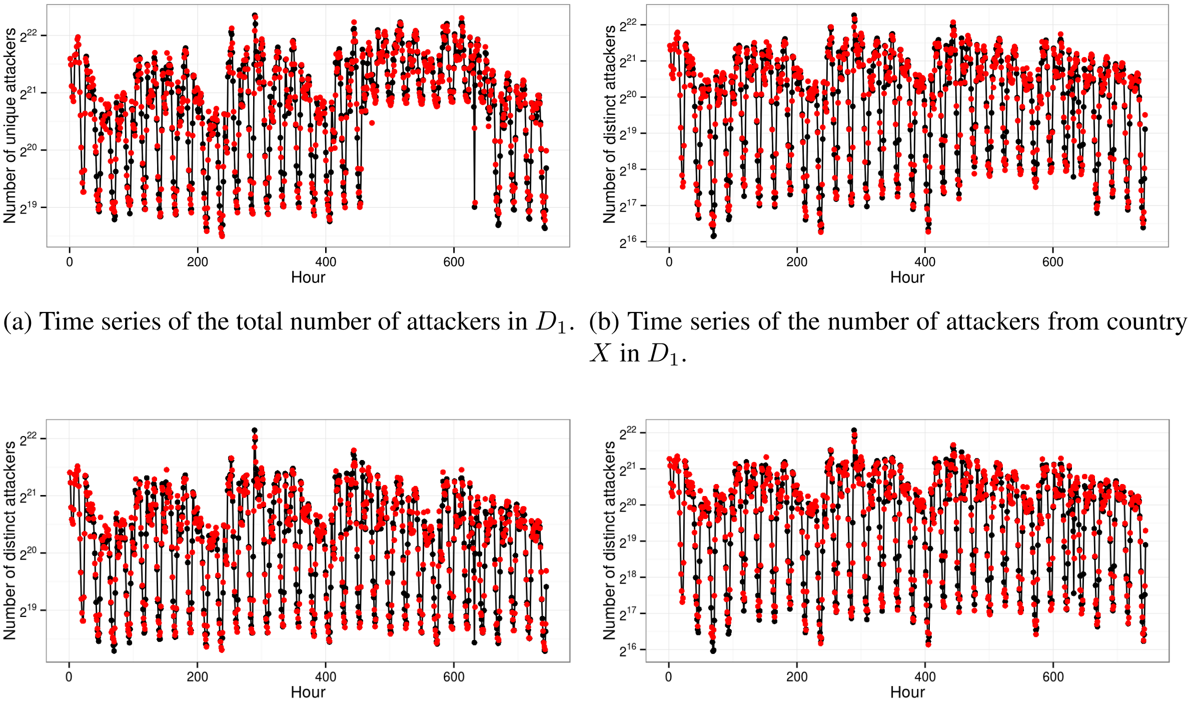
where $W_t = |A(r;t,t+1)| - |A(r;t-24,t-23)|$.

Table 4.2: Coefficients and standard deviation for fitted models with respect to the number of distinct attackers from country X and the total number of distinct attackers observed by the blackhole during the month.

	ϕ_1	θ_1	Φ_1	Φ_2	Θ_1	Θ_2	Θ_3
Time series of the number of distinct attackers from country X in D_1							
Coefficients	0.82	.39	1.22	-.99	-2.19	2.16	-.91
Standard deviation	.02	.03	.01	.007	.06	.13	.078
Time series of the total number of distinct attackers in D_1							
Coefficients	.91	.38	1.22	-.98	-2.15	2.11	-.86
Standard deviation	.017	.03	.02	.01	.08	.18	.10
Time series of the number of distinct attackers from country X in D_2							
Coefficients	.79	.4	1.21	-.99	-2.19	2.16	-.9
Standard deviation	.02	.04	.02	.007	.06	.13	.07
Time series of the total number of distinct attackers in D_2							
Coefficients	.79	.4	1.21	-.99	-2.18	2.16	-.9
Standard deviation	.02	.04	.02	.008	.06	.12	.07

Table 4.2 summarizes the fitting results. We make the following observations. Corresponding to D_1, the two fitted models are similar to each other in terms of the coefficients. Corresponding to D_2, the two fitted models are almost identical to each other.

Figures 4.7a and 4.7b show the fitting results of the two times series in D_1. Figures 4.7c and 4.7d show the fitting results of the two time series in D_2. We observe that both time series in D_1 and D_2 are fitted well. For D_1, the PMAD (i.e., fitting error) values for fitting the times series of the total number of distinct attackers and the number of distinct attackers from country X are 0.08 and 0.06, respectively. For D_1, the PMAD (i.e., fitting error) values for fitting the times series of the total number of distinct attackers and the number of distinct attackers from country X are 0.08 and 0.07, respectively.



(a) Time series of the total number of attackers in D_1.

(b) Time series of the number of attackers from country X in D_1.

(c) Time series of the total number of attackers in D_2.

(d) Time series of the number of attackers from country X in D_2.

Figure 4.7: Model fitting of the time series of the total number of distinct attackers and the time series of the number of distinct attackers from country X, where black-colored dots represent observed values and red-colored dots represent fitted values.

In summary, Table 4.2 and Figure lead us to draw the conclusion that the two time series, especially after filtering the traffic that is likely caused by misconfigurations, exhibit very similar fittability.

4.5 Limitations of the Study

The present study has several limitations. First, both D_1 and D_2 may contain misconfiguration-cause, non-malicious traffic. Due to the lack of interactions between blackhole IP addresses and remote computers (an inherent limitation of blackhole), it is hard to know the ground truth. Therefore, better filtering methods are needed so as to make the data approximate the ground truth as closely as possible.

Second, it is possible that some attackers are aware of the blackhole and therefore can instruct their attacks to bypass it. As a consequence, the data collected by blackhole may not faithfully reflect the cybersecurity posture because the data is not a spatial "uniform" sample of the attack traffic in Internet. Nevertheless, our methodology would be equally applicable to analyze more representative data, when it becomes available.

Third, the data collected by blackhole does not contain rich information that would allow us to conduct deeper analysis, such as analyzing the global characteristics of specific attacks. Still, our methodology is equally useful for analyzing data with richer information (when available).

4.6 Conclusion

We have studied the cybersecurity posture based on the data collected by CAIDA's blackhole during the month of March 2013. We have analyzed both the "as is" data (D_1) and the data obtained after heuristically filtering some rarely seen attackers (D_2). We have defined the notion of *sweep-time* and found that the sweep-time follows the power-law distribution. We have found that the *total* number of distinct attackers that are observed by the blackhole is largely determined by the number of distinct attackers from a certain country code.

We expect to publish more detailed analysis and (hypothetical) explanations of the newly identified phenomena elsewhere shortly.

Chapter 5: CONCLUSION

5.1 Summary

In this dissertation, we propose a systematic statistical framework for analyzing cyber attacks. Empowered by this new concept, we present results in the following three frontiers. First, we show that cyber attacks can exhibit the Long-Range Dependence (LRD) phenomenon, which is for the first time found to be relevant in the cyber security domain. We demonstrate how to exploit LRD to achieve gray-box (rather than black-box) prediction. Second, we show that cyber attacks can exhibit the Extreme Value (EV) phenomenon. We characterize the EV phenomenon and show how to exploit for even better prediction of extreme events. Third, we characterize spatial and temporal properties that are exhibited by blackhole-captured cyber attacks. These statistical characteristics are useful not only from a practice perspective (e.g., guiding proactive allocation of resources in anticipating the incoming attacks), but also from a theoretical perspective (e.g., guiding the development of theoretical cyber security models that can accommodate the desired statistical properties).

5.2 Future Work

Dependence is perhaps inherent to nature and perhaps to cyberspace as well. In theoretic cyber security models, dependence is often assumed so as to simplify the analysis. However, it may be possible that we cannot afford to assume away the dependence in question. In order to understand and characterize the significance of dependence in real cyber attacks, the present chapter studies the dependence between attack processes at multiple resolutions. Specifically, it plans to study the following:

- We showed that a network-level attack process is composed (or superposition) of computer-level attack processes. We want to answer the following questions: Are the computer-level attack processes dependent upon each other? How strong is the dependence? Intuitively, certain kinds of attacks would cause strong dependence (e.g., the attacker is attempting to attack a chunk of consecutive IP addresses).

- We showed that a computer-level attack process is composed of port-level attack processes. Are the port-level attack processes dependent upon each other? How strong is the dependence? Intuitively, strong dependence would be exhibited when the attacker attempts to attack the same computer's ports one after another.

- Extremal dependence. Dependence between extremal attack processes are very important. For example, for the port-level attacks, does the computer receive a large number of attacks from different ports simultaneously? If it does, how to measure this dependence?

Chapter 6: APPENDIX

6.1 Review of Some Statistical Techniques

6.1.1 Methods for Estimating Hurst Parameter

We used six popular methods (cf. [15] for details) for estimating the Hurst parameter, which is a well-accepted practice [58, 64].

1) RS method: For a time series $\{X_t, t \geq 1\}$, with partial sum $Y_t = \sum_{i=1}^{t} X_i$ and sample variance $S_t^2 = \frac{1}{t}\sum_{i=1}^{t} X_i^2 - \left(\frac{1}{t}\right)^2 Y_t^2$, the R/S statistic is defined as

$$\frac{R}{S}(n) = \frac{1}{S_n}\left[\max_{0\leq t\leq n}\left(Y_t - \frac{t}{n}Y_n\right) - \min_{0\leq t\leq n}\left(Y_t - \frac{t}{n}Y_n\right)\right].$$

For LRD series, we have

$$\mathsf{E}\left[\frac{R}{S}(n)\right] \sim C_H n^H, \;\; n \to \infty$$

where C_H is a positive, finite constant independent of n.

2) AGV (aggregated variance) method: Divide time series $\{X_t, t \geq 1\}$ into blocks of size m. The block average is

$$X^{(m)}(k) = \frac{1}{m}\sum_{i=(k-1)m+1}^{km} X_i, \quad k = 1, 2 \ldots.$$

Take the sample variance of $X^{(m)}(k)$ within each block, which is an estimator of $\mathrm{Var}(X^{(m)})$. For LRD series, we have $\beta = 2H - 2$ and

$$\mathrm{Var}\left(X^{(m)}\right) \sim cm^{-\beta}, \quad m \to \infty,$$

where c is a finite positive constant independent of m.

3) Peng method: The series is broken up into blocks of size m. Compute partial sums $Y(i)$, $i = 1, 2\ldots, m$ within blocks. Fit a least-square line to the $Y(i)$'s and compute the sample variance of

the residuals. This procedure is repeated for each of the blocks, and the resulting sample variances are averaged. The resulting number is proportional to m^{2H} for LRD series.

4) Per (Periodogram) method: One first calculates

$$I(\lambda) = \frac{1}{2\pi N} \left| \sum_{j=1}^{N} X_j e^{ij\lambda} \right|,$$

where λ is the frequency, N is the number of terms in the series, and X_j is the data. A LRD series should have a periodogram proportional to λ^{1-2H} for $\lambda \approx 0$. A regression of the logarithm of the periodogram on the logarithm of the frequency gives coefficient $1 - 2H$.

5) Box (Boxed Periodogram) method: This method was developed to deal with the problem of having most of the points, which are used to estimate H, on the right-hand side of the graph.

6) Wave (Wavelet) method: Wavelets can be thought of as akin to Fourier series but using waveforms other than sine waves. The estimator used here fits a straight line to a frequency spectrum derived using wavelets [5].

6.1.2 Heavy-tail Distributions

A random variable X is said to belong to the Maximum Domain of Attraction (MDA) of the extreme value distribution H_ξ if there exists constants $c_n \in \mathbb{R}_+$, $d_n \in \mathbb{R}$ such that its distribution function F that satisfies

$$\lim_{n\to\infty} nF(c_n x + d_n) = H_\xi(x).$$

In statistics, X is said to follow a heavy-tailed distribution if $F \in \mathrm{MDA}(H_\xi)$. There are many methods for estimating parameter α [27,59]. A widely-used method is called Point Over Threshold (POT). Let $X_1, \ldots, X_n$ be independent and identically distributed random variables from $F \in \mathrm{MDA}(H_\xi)$, then we may choose a high threshold u such that

$$\lim_{u\to x_F} \sup_{0<x<x_F-u} |\bar{F}_u(x) - \bar{G}_{\xi,\beta(\mu)}(x)| = 0,$$

where x_F is the right end poind point of X, and

$$F_u(x) = P(X - u \leq x | X > u), \quad x \geq 0,$$

and $\bar{G}_{\xi,\beta(\mu)} = 1 - G_{\xi,\beta(\mu)}$ is the survival function of generalized Pareto distribution (GPD)

$$\bar{G}_{\xi,\beta(\mu)}(x) = \begin{cases} \left(1 + \xi \dfrac{x}{\beta}\right)^{-1/\xi}, & \xi \neq 0 \\ \exp\{-x/\beta\}, & \xi = 0 \end{cases}$$

where $x \in \mathbb{R}^+$ if $\xi \in \mathbb{R}^+$, and $x \in [0, -\beta/\xi]$ if $\xi \in \mathbb{R}^-$. The POT method states that if $X_1, \ldots, X_n$ are heavy-tailed data, then $[X_i - u | X_i > u]$ follows a generalized Pareto distribution.

6.1.3 Goodness-of-fit Test Statistics

We use three popular goodness-of-fit test statistics: Kolmogorov-Smirnov (KS), Cramér-von Mises (CM), and Anderson-Darling (AD). Let $X_1, \ldots, X_n$ be independent and identical random variables with distribution F. The empirical distribution F_n is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(X_i \leq x),$$

where $\mathrm{I}(X_i \leq x)$ is the indicator function:

$$\mathrm{I}(X_i \leq x) = \begin{cases} 1, & X_i \leq x, \\ 0, & o/w. \end{cases}$$

The KS test statistic is defined as

$$\mathrm{KS} = \sqrt{n} \sup_x \left| F_n(x) - F(x) \right|.$$

The CM test statistic is defined as

$$\mathrm{CM} = n \int (F_n(x) - F(x))^2 dF(x).$$

The AD test statistic is defined as

$$\mathrm{AD} = n \int (F_n(x) - F(x))^2 w(x) dF(x),$$

where $w(x) = [F(x)(1 - F(x))]^{-1}$.

BIBLIOGRAPHY

[1] $http : //amunhoney.sourceforge.net/$.

[2] $http : //dionaea.carnivore.it/$.

[3] $https : //alliance.mwcollect.org/$.

[4] $http : //www.snort.org/$.

[5] P. Abry and D. Veitch. Wavelet analysis of long-range-dependent traffic. *IEEE Transactions on Information Theory*, 44(1):2–15, 1998.

[6] Mark Allman, Vern Paxson, and Jeff Terrell. A brief history of scanning. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07, pages 77–82, New York, NY, USA, 2007. ACM.

[7] S. Almotairi, A. Clark, G. Mohay, and J. Zimmermann. Characterization of attackers' activities in honeypot traffic using principal component analysis. In *Proceedings of the 2008 IFIP International Conference on Network and Parallel Computing*, pages 147–154, 2008.

[8] S. Almotairi, A. Clark, G. Mohay, and J. Zimmermann. A technique for detecting new attacks in low-interaction honeypot traffic. In *Proc. International Conference on Internet Monitoring and Protection*, pages 7–13, 2009.

[9] Saleh I. Almotairi, Andrew J. Clark, Marc Dacier, Corrado Leita, George M. Mohay, Van Hau Pham, Olivier Thonnard, and Jacob Zimmermann. Extracting inter-arrival time based behaviour from honeypot traffic using cliques. In *5th Australian Digital Forensics Conference*, pages 79–87, 2007.

[10] Saleh I. Almotairi, Andrew J. Clark, George M. Mohay, and Jacob Zimmermann. Characterization of attackers' activities in honeypot traffic using principal component analysis. In *Proc. IFIP International Conference on Network and Parallel Computing*, pages 147–154, 2008.

[11] Paul Baecher, Markus Koetter, Maximillian Dornseif, and Felix Freiling. The nepenthes platform: An efficient approach to collect malware. In *In Proceedings of the 9 th International Symposium on Recent Advances in Intrusion Detection (RAID)*, pages 165–184, 2006.

[12] M. Bailey, E. Cooke, F. Jahanian, A. Myrick, and S. Sinha. Practical darknet measurement. In *Information Sciences and Systems, 2006 40th Annual Conference on*, pages 1496–1501, March 2006.

[13] M. Bailey, E. Cooke, F. Jahanian, and D. Watson. The blaster worm: Then and now. *Security Privacy, IEEE*, 3(4):26–31, July 2005.

[14] Michael Bailey, Evan Cooke, Farnam Jahanian, Jose Nazario, David Watson, et al. The internet motion sensor-a distributed blackhole monitoring system. In *NDSS*, 2005.

[15] J. Beran. *Statistics for Long-Memory Processes*. Chapman and Hall, 1994.

[16] Tim Bollerslev, Jeffrey Russell, and Mark W Watson. *Volatility and Time Series Econometrics: Essays in Honor of Robert Engle*. Oxford University Press, 2010.

[17] Nevil Brownlee. One-way traffic monitoring with iatmon. In *Proceedings of the 13th International Conference on Passive and Active Measurement*, PAM'12, pages 179–188, Berlin, Heidelberg, 2012. Springer-Verlag.

[18] M. Chandra, N. D. Singpurwalla, and M. A. Stephens. Kolmogorov statistics for tests of fit for the extreme value and weibull distributions. *J. Amer. Statist. Assoc.*, 74:729–735, 1981.

[19] V. Choulakian and M.A. Stephens. Goodness-of-fit tests for the generalized pareto distribution. *Technometrics*, 43:478–484, 2001.

[20] Andrew Clark, Marc Dacier, George Mohay, Fabien Pouget, and Jakub Zimmermann. Internet attack knowledge discovery via clusters and cliques of attack traces. *Journal of Information Assurance and Security*, 1(1):21–32, 2006.

[21] Gregory Conti and Kulsoom Abdullah. Passive visual fingerprinting of network attack tools. In *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, pages 45–54, 2004.

[22] Evan Cooke, Michael Bailey, Z Morley Mao, David Watson, Farnam Jahanian, and Danny McPherson. Toward understanding distributed blackhole placement. In *Proceedings of the 2004 ACM workshop on Rapid malcode*, pages 54–64. ACM, 2004.

[23] Jonathan D. Cryer and Kung-Sik Chan. *Time Series Analysis With Applications in R*. Springer, New York, 2008.

[24] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes, Volume 1 (2nd ed.)*. Springer, 2002.

[25] Falko Dressler, Wolfgang Jaegers, and Reinhard German. Flow-based worm detection using correlated honeypot logs. *Communication in Distributed Systems (KiVS), 2007 ITG-GI Conference*, pages 1–6, 2007.

[26] Thomas Dubendorfer and Bernhard Plattner. Host behaviour based early detection of worm outbreaks in internet backbones. In *Proc. IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise*, pages 166–171, 2005.

[27] P. Embrechts, C. Kluppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin, 1997.

[28] Yan Gao, Zhichun Li, and Yan Chen. A dos resilient flow-level intrusion detection approach for high-speed networks. In *Proc. IEEE International Conference on Distributed Computing Systems (ICDCS'06)*, pages 39–, 2006.

[29] A. Ghourabi, T. Abbes, and A. Bouhoula. Data analyzer based on data mining for honeypot router. In *Computer Systems and Applications (AICCSA), 2010 IEEE/ACS International Conference on*, pages 1–6, 2010.

[30] Eduard Glatz and Xenofontas Dimitropoulos. Classifying internet one-way traffic. In *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, IMC '12, pages 37–50, New York, NY, USA, 2012. ACM.

[31] Eduard Glatz and Xenofontas A. Dimitropoulos. Classifying internet one-way traffic. In *Internet Measurement Conference*, pages 37–50, 2012.

[32] Shorak G.R. and Wellner J.A. *Empirical Processes with Applications to Statistics*. Springer, 1986.

[33] Alefiya Hussain, John Heidemann, and Christos Papadopoulos. A framework for classifying denial of service attacks. In *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, SIGCOMM '03, pages 99–110, New York, NY, USA, 2003. ACM.

[34] Myung-Sup Kim, Hun-Jeong Kang, Seong-Cheol Hong, Seung-Hwa Chung, and James W. Hong. A flow-based method for abnormal network traffic detection. In *NOMS (1)'04*, pages 599–612, 2004.

[35] F. Lau, S.H. Rubin, M.H. Smith, and L. Trajkovic. Distributed denial of service attacks. In *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, volume 3, pages 2275–2280 vol.3, 2000.

[36] DongJin Lee and Nevil Brownlee. Passive measurement of one-way and two-way flow lifetimes. *SIGCOMM Comput. Commun. Rev.*, 37(3):17–28, July 2007.

[37] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Trans. Netw.*, 2(1):1–15, 1994.

[38] Will E. Leland and Daniel V. Wilson. High time-resolution measurement and analysis of lan traffic: Implications for lan interconnection. In *INFOCOM*, pages 1360–1366, 1991.

[39] Z. Li, A. Goyal, Y. Chen, and V. Paxson. Towards situational awareness of large-scale botnet probing events. *Information Forensics and Security, IEEE Transactions on*, 6(1):175–188, march 2011.

[40] Zhichun Li, A. Goyal, Yan Chen, and A. Kuzmanovic. Measurement and diagnosis of address misconfigured p2p traffic. *Network, IEEE*, 25(3):22–28, May 2011.

[41] Carl Livadas, Robert Walsh, David Lapsley, and W. Timothy Strayer. Using machine learning techniques to identify botnet traffic. In *Proc. IEEE LCN Workshop on Network Security (WoNS'2006)*, pages 967–974, 2006.

[42] Matthew V. Mahoney and Philip K. Chan. Learning nonstationary models of normal network traffic for detecting novel attacks. In *Proc. 8th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'02)*, pages 376–385, 2002.

[43] Alexander J McNeil and Rüdiger Frey. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of empirical finance*, 7(3):271–300, 2000.

[44] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques, and tools*. Princeton university press, 2010.

[45] T. Mikosch and C. Starica. Nonstationarities in financial time series, the long-range dependence, and the igarch effects. *The Review of Economics and Statistics*, 86(1):378–390, February 2004.

[46] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver. Inside the slammer worm. *IEEE Security and Privacy*, 1(4):33–39, Aug 2003.

[47] D. Moore, C. Shannon, D. Brown, G. Voelker, and S. Savage. Inferring internet denial-of-service activity. *ACM Trans. Comput. Syst.*, 24(2):115–139, May 2006.

[48] D. Moore, C. Shannon, and J. Brown. Code-Red: a case study on the spread and victims of an Internet worm. In *Internet Measurement Workshop (IMW) 2002*, pages 273–284, Marseille, France, Nov 2002. ACM SIGCOMM/USENIX Internet Measurement Workshop.

[49] David Moore, Colleen Shannon, Geoffrey M Voelker, and Stefan Savage. *Network telescopes: Technical report*. Department of Computer Science and Engineering, University of California, San Diego, 2004.

[50] Ruoming Pang, Vinod Yegneswaran, Paul Barford, Vern Paxson, and Larry Peterson. Characteristics of internet background radiation. In *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, IMC '04, pages 27–40, New York, NY, USA, 2004. ACM.

[51] Ruoming Pang, Vinod Yegneswaran, Paul Barford, Vern Paxson, and Larry L. Peterson. Characteristics of internet background radiation. In *Internet Measurement Conference*, pages 27–40, 2004.

[52] B. Peter and D. Richard. *Introduction to Time Series and Forecasting*. Springer, 2002.

[53] Van Hau Pham. *Honeypot traces forensics by means of attack event identification*. PhD thesis, Thesis, 09 2009.

[54] Fabien Pouget and Marc Dacier. Honeypot-based forensics. In *AusCERT2004, AusCERT Asia Pacific Information technology Security Conference*, 05 2004.

[55] Niels Provos. A virtual honeypot framework. In *Proceedings of the 13th conference on USENIX Security Symposium*, pages 1–1, 2004.

[56] Zhongjun Qu. A test against spurious long memory. Boston University - Department of Economics - Working Papers Series WP2010-051, Boston University - Department of Economics, 2010.

[57] D'Agostino R.B. and Stephens M.A. *Tests Based on EDF Statistics*. Springer, 1986.

[58] W. Rea, M. Reale, and J. Brown. Estimators for long range dependence: An empirical study. *arXiv: 0901.0762v1*, 2009.

[59] S. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, 2007.

[60] S. Ross. *Stochastic Processes*. Wiley and Sons, 1996.

[61] Gennady Samorodnitsky. Long range dependence. *Foundations and Trends in Stochastic Systems*, 1(3):163–257, 2006.

[62] Xiaofeng Shao. A simple test of changes in mean in the possible presence of long-range dependence. *Journal of Time Series Analysis*, 32(6):598–606, November 2011.

[63] W. Timothy Strayer, David Lapsely, Robert Walsh, and Carl Livadas. Botnet detection based on network behavior. volume 36, pages 1–24. 2008.

[64] M. S. Taqqu, V. T. Teverovsky, and W. Willinger. Estimators for long range dependence: An empirical study. *Fractals*, 3(4):785–798, 1995.

[65] The CAIDA UCSD Network Telescope. `http://http://www.caida.org/`.

[66] The CAIDA UCSD Network Telescope. `http://www.caida.org/tools/measurement/corsaro/docs/plugins.html`.

[67] Olivier Thonnard and Marc Dacier. A framework for attack patterns' discovery in honeynet data. *Digital Investigation*, 5:S128–S139, 2008.

[68] Joanne Treurniet. A network activity classification schema and its application to scan detection. *IEEE/ACM Trans. Netw.*, 19(5):1396–1404, 2011.

[69] Tsay. *Analysis of Financial Time Series*. Wiley, 2010.

[70] N. Weiler. Honeypots for distributed denial-of-service attacks. In *Enabling Technologies: Infrastructure for Collaborative Enterprises, 2002. WET ICE 2002. Proceedings. Eleventh IEEE International Workshops on*, pages 109–114, 2002.

[71] W. Willinger, M. S. Taqqu, W. E. Leland, and V. Wilson. Self-similarity in high-speed packet traffic: analysis and modeling of ethernet traffic measurements. *Statistical Sci.*, 10:67–85, 1995.

[72] Walter Willinger, Murad S. Taqqu, Robert Sherman, and Daniel V. Wilson. Self-similarity through high-variability: statistical analysis of ethernet lan traffic at the source level. *IEEE/ACM Trans. Netw.*, 5(1):71–86, 1997.

[73] Eric Wustrow, Manish Karir, Michael Bailey, Farnam Jahanian, and Geoff Huston. Internet background radiation revisited. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, IMC'10, pages 62–74, New York, NY, USA, 2010. ACM.

[74] Eric Wustrow, Manish Karir, Michael Bailey, Farnam Jahanian, and Geoff Huston. Internet background radiation revisited. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, pages 62–74, New York, NY, USA, 2010. ACM.

[75] Vinod Yegneswaran, Paul Barford, and Dave Plonka. On the design and use of internet sinks for network abuse monitoring. In *Recent Advances in Intrusion Detection*, pages 146–165. Springer, 2004.

[76] Vinod Yegneswaran, Paul Barford, and Johannes Ullrich. Internet intrusions: Global characteristics and prevalence. In *Proceedings of the 2003 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '03, pages 138–147, New York, NY, USA, 2003. ACM.

[77] Zhenxin Zhan, Maochao Xu, and Shouhuai Xu. Characterizing honeypot-captured cyber attacks: Statistical framework and case study. *IEEE Transactions on Information Forensics and Security*, 8(11):1775–1789, 2013.

VITA

Zhenxin Zhan was born in Hubei province, China. He received his B.S. degree from Huazhong University of Science and Technology in 2006, and M.SC. degree from Huazhong University of Science and Technology in 2008. He started his Ph.D. study at University of Texas at San Antonio from September 2008.