© 2014 Roseline Estelle Sindolmane Kone

MONITORING UNKNOWN SOURCE IP ADDRESSES AND PACKET SIZES TO DETECT DDOS ATTACKS

BY

ROSELINE ESTELLE SINDOLMANE KONE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Industrial Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Professor Richard Sowers, Chair
Associate Professor Ali Abbas
Assistant Professor Negar Kiyavash
Professor Renming Song

ABSTRACT

This thesis presents three procedures to detect Distributed Denial of Service (DDoS) attacks. DDoS attacks are known as one of the most expensive and destructive Internet threats. Assuming network traffic is a marked Poisson process, two parametric detection models are developed. The arrival of packet flows is modeled as Poisson process with cluster sizes that follows a mixture of discrete and heavy tailed distributions. Both detection systems monitor the percentage of unknown source IP addresses. The first detection model is formulated as a fixed sample size binary hypothesis testing. The decision making is based on the Neyman-Pearson criteria. The second parametric model is a sequential probability ratio test where the sample size is a random variable. Acceptance and rejection boundaries are deduced based on Wald's Fundamental Identity. Given that parametric distributions may fail to capture the complex and dynamic nature of the Internet, a third non-parametric detection model is proposed. In addition to the percentage of unknown source IP addresses, a second test statistic is introduced. The latter represents the mean to standard deviation ratio of data packet sizes. The Neyman-Pearson threshold is estimated from the empirical distribution functions of both random variables.

A toute ma famille, a ma mere et mes soeurs en particulier.

ACKNOWLEDGMENTS

To my advisor Dr Sowers: I would like to thank you very much for giving me the opportunity to work with you and for teaching me what to do. More importantly, thank you for being patient as I was learning.

To my committee Dr Abbas, Dr Kiyavash, Dr Song: Thank you for agreeing to serve on my committee. I truly appreciate your time and consideration. Also, thank you for sharing your thoughts, they were more than valuable and helped me improve my work, for that I am grateful.

To Dr Moulin: Je ne pouvais pas finir sans vous remercier sincerement pour toutes nos conversations et surtout de m'avoir authorisee a assister a votre cour. Vous avez toujours ete disponible et je vous en suis tres reconnaissante.

To family and friends: Thank you for your continuous support. Merci d'avoir toujours ete la pour moi.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER 1 INTRODUCTION 1
1.1 Basic Definitions 2
1.2 Distributed Denial of Service (DDoS) 3
1.3 DDoS Defense Systems 5
1.4 Research Overview 6

CHAPTER 2 NETWORK TRAFFIC MODELING AND METHODOLOGIES 10
2.1 Network Traffic Model 10
2.2 Poisson Processes 15
2.3 Characteristic Functions 17
2.4 Saddlepoint Approximation 23
2.5 Fast Fourier Transform (FFT) 25
2.6 Detection Theory 26

CHAPTER 3 FIXED SAMPLE DETECTION MODEL 30
3.1 Distinct Source IP Statistic 30
3.2 Marked Poisson with Pareto Cluster Size 34
3.3 Marked Poisson with Mixture Cluster Size 54
3.4 Detection and Large Deviations for Heavy-Tailed Distributions 60
3.5 Summary 63

CHAPTER 4 SEQUENTIAL DETECTION MODEL 64
4.1 Poisson Processes and Uniformly Most Powerful Tests 64
4.2 Overview Sequential Hypothesis Testing 66
4.3 Detection Model Based on Distinct Source IP Addresses 73
4.4 Time-Discrete Sequential Test for Homogeneous Poisson with Cluster Size 78
4.5 Sequential HT for Homogeneous Poisson with Cluster Size 85
4.6 Summary 91

CHAPTER 5 EMPIRICAL DDOS DETECTION MODEL 92
5.1 Data Packet Sampling Procedure and Traffic Statistics 92
5.2 Supervised Learning Detection Model 98
5.3 Simulation Model and Computational Results 109
5.4 Summary 116

CHAPTER 6 CONCLUSION 118
6.1 Summary 118
6.2 Possible Research Extensions 119

APPENDIX 120

REFERENCES 122

LIST OF TABLES

1.1 Simulation Parameters 8

5.1 Percentiles Observation Window Size 95
5.2 N Distribution Statistics 96
5.3 S Distribution Statistics 96
5.4 Training Data with DDoS Data Packets 110
5.5 Probability of Event $\{N = 1, S = 1\}$ 111
5.6 P-Values 111
5.7 Testing Data with DDoS 112
5.8 FP-FN Rates with $H = 1$ 113
5.9 FP-FN Rates with $H = 3$ 113
5.10 γ_i's Expectated Values 113
5.11 Comparing with IP Filtering [101] 115
5.12 Comparing with IP Filtering [29] 116

LIST OF FIGURES

1.1 Packet Header 2
1.2 Protocol Layer 3
1.3 A Simple DDoS Attack Model 4
1.4 Observed Attack and Legitimate Flow 7

2.1 Histogram Small Flow Size- Auckland and W1M-MIT Datasets 11
2.2 QQ-Plot Inter-arrival times vs Exponential-Auckland 14
2.3 Packet Flow Count per 30 mns 15
2.4 Power Law Fitting to Cluster Size - Auckland 16
2.5 QQ-Plot Pareto vs Auckland and W1M-MIT Datasets 16

3.1 Convergence Sum Paretos to Stable 43
3.2 FFT Approximation of the Stable Pdf with $\alpha = 1.3$ 44
3.3 Saddlepoint Approximation for Stable-1 46
3.4 Pareto Pdf Simulation vs Saddlepoint Approximation - Alpha 3.45 49

5.1 Sampling Size (m) Distribution for W1-M 95
5.2 N-S Probability Distribution 97
5.3 Weighted Packet Size Statistic 102
5.4 Probability Grid with Acceptance and Rejection Regions 110
5.5 Pmf Distinct Unknown Source IP Addresses 114

CHAPTER 1

INTRODUCTION

In the past decade, several renowned e-commerce websites including Yahoo, eBay, E*Trade, and Microsoft have been the subjects of Denial of Service (DoS) attacks. DoS attacks are designed to prevent legitimate users from accessing the targeted network. They can be classified into two main categories. *Logic* attacks significantly impair target servers by exploiting software weaknesses, e.g., sending malformed data packets while *flooding* attacks seize all the target network resources by sending a significant amount of fake requests to occupy the network [28, 86]. While it is possible to design more sophisticated softwares to counter logic attacks, the second type of attacks is much harder to detect as attack data packets do not have specific characteristics that distinguish them form legitimate data packets. When the attack is led by a so called "master" and coordinated across many hijacked systems (zombies), it is called a Distributed DoS (DDoS). A DDoS attack can involve dozens, hundreds or as many as 60,000 zombies [112, 124], which significantly amplifies the attack. DDoS attacks are thus among the most destructive and expensive Internet security incidents. DoS attacks are also becoming more and more elaborate and prevalent; in 2001 alone, 12000 DDoSs attacks were reported over a 3-weeks period [18, 86].

Existing DDoS detection mechanisms can be classified as anomaly or signature based. The latter compares network traffic features to known attack patterns (signatures), which are pre-recorded in a database. Anomaly-based systems, which are the most widely used [20], raise an alarm when the observed traffic features deviate from what is labeled as legitimate activities. While signature-based systems cannot detect unknown attacks and consequently result in a high False Clear (FC) rate, an anomaly-based detection system may provide detection results with a high False Alarm (FA) rate [58]. A FA occurs when a detection system raises an alarm when a legitimate data packet is sent to the network. A FC indicates that a detection system has allowed an attack data packet to proceed without raising an alarm. In network security, both FA and FC are undesirable; FAs result in financial losses while FCs can be fatal to the network and result in even greater financial losses [7]. The ideal detection mechanism minimizes both the FA and FC rates. However, these are contradictory objectives. Designing an efficient DDoS detection system thus requires a trade-off. Using an anomaly-based approach, this thesis presents several novel techniques to detect flooding DDoS attacks. The next section introduce basic concepts that are used throughout the thesis.

1.1 Basic Definitions

A *data packet* is a formatted unit of data that is transported through a network. Beside the actual information that is to be sent, supplemental information such as the source and destination addresses must be encoded with the data before it is transmitted over the Internet. Figure 1.1 illustrates such a packet *header*. A *protocol* defines a set of rules for transmitting data between two

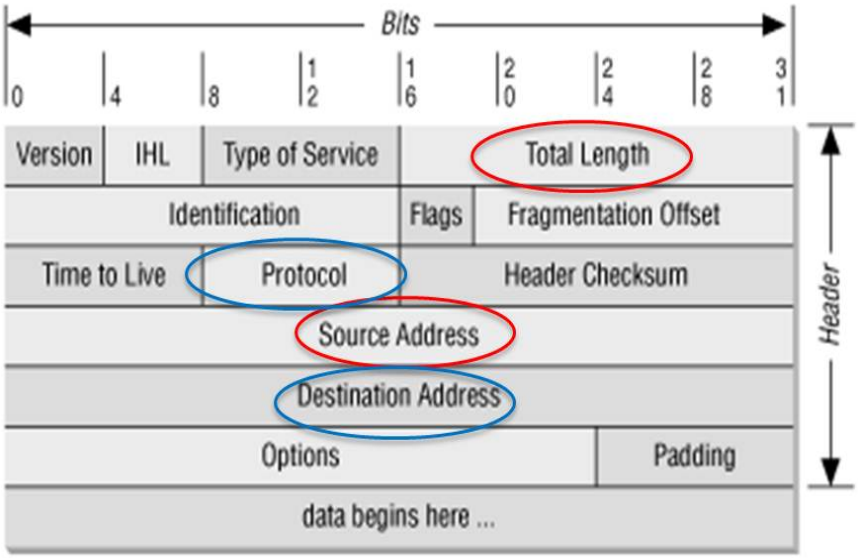


Figure 1.1: Packet Header

end-points (within or between computers). It specifies, among other parameters, the data format, address format, and routing rules. Communications through the Internet uses the Internet Protocol Suite (IP), which is a set of communication protocols. Several layers of protocol are necessary for transmitting data over the Internet. Figure 1.2 shows examples of protocols by layer type. The application layer is a set of methods defined for process-to-process communications across the Internet. A well known example is the Hypertext Transfer Protocol (HTTP). The Internet layer, e.g., the Internet Control Message Protocol (ICMP), constitutes the link between independent networks. The transport layer provides communication services between hosts and delivers data to the appropriate application process. A major transportation protocol is the Transmission Control Protocol (TCP). Together with the IP protocol, the pair is referred to as TCP/IP. Another well known transportation protocol is the User Datagram Protocol (UDP). The network interface layer, also known as the data link layer, constitutes the physical interface between the host system and the network hardware. It describes how data packets should be formatted in order to be appropriately routed. A *router* is an intermediary system that forwards data on behalf of the sender. On the Internet, networks are linked through routers, which are indispensable if the sender and receiver are not directly connected. Any device connected to the Internet is assigned a numerical label

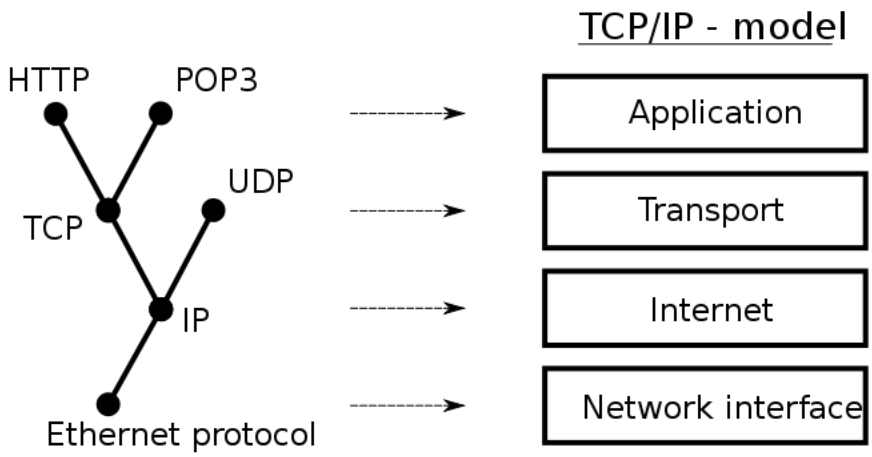


Figure 1.2: Protocol Layer

called the Internet Protocol (IP) address. The *source IP address* designates the address of the device where the data has originated. Similarly, the *destination IP address* represents the digital address of the receiver. *IP spoofing* is a technique used by attackers to conceal their true identity. It consists of altering data packet headers so that the sender appears to be a legitimate host. The next section describes DDoS attacks more in detail and presents an overview of various defense mechanisms.

1.2 Distributed Denial of Service (DDoS)

The objective of a DDoS attack is to prevent legitimate users from accessing an online service. During the attack, key resources such as network bandwidth, memory, and Central Processing Unit (CPU) are overloaded in a manner that degrades or entirely interrupts the intended service [86]. Throughout the thesis, *target network* or *victim* refers to the network under DDoS attack.

1.2.1 DDoS Attack Structure

As illustrated in Figure 1.3, a DDoS attack comprises three main elements: the master attacker, the zombie hosts and the victim.

Prior to launching the attack, the master attacker explores the Internet, searching for computers with low security levels. The technique is known as *scanning*. After gaining access to a machine, the latter is subdued thanks to a computer program called *bot*. The *botnet*, i.e., the set of zombies, may contain tens of thousands of computers [104, 112]. At the time of the attack, zombies are

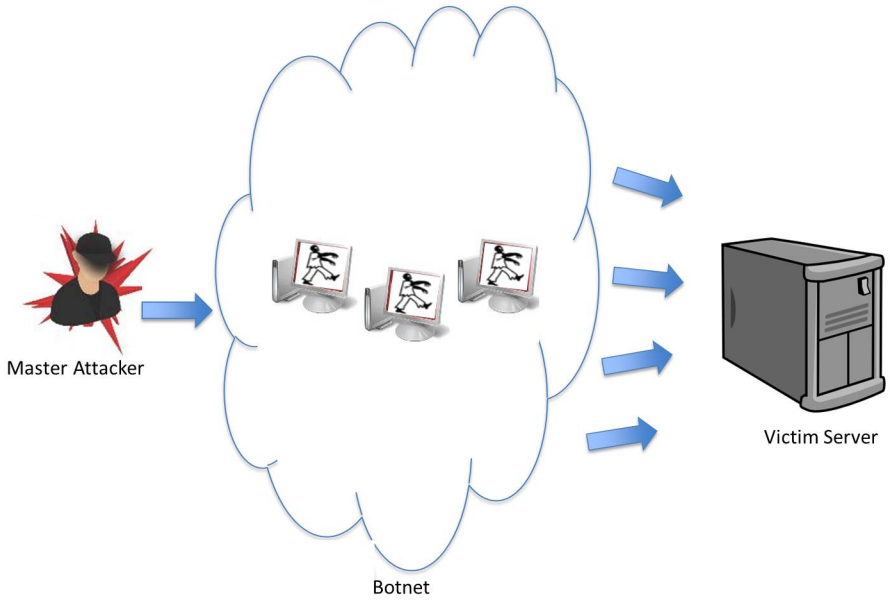


Figure 1.3: A Simple DDoS Attack Model

"ordered" to send data packets to the victim, all at once, flooding key resources and causing the target system to crash. A DDoS attack not only causes damages to the target network but also to the intermediary computers that have been compromised.

Multiple factors facilitate DDoS attacks. The first is the limitedness of Internet resources; each host can only handle so many requests. Furthermore, attackers take advantage of the structure of the Internet. That is, the Internet is designed to efficiently move data packets from one source to a destination. Neither users nor data packets are authenticated and ordinary routers are programmed to process any incoming data packet. The lack of authentication allows network intruders to use fake source IP addresses. Attackers also rely on zombies to hide their true identity. A third key factor is the global connectivity of the Internet. Indeed, large quantities of zombies can be recruited and DDoS attacks can be launched from anywhere in the world. Such a broad infrastructure would require an equally broad global security system. However, each local network usually has its own set of policies, which may be an obstacle to building such a system. Additionally, there may be privacy issues that prevent information sharing [82].

1.2.2 Attack Classification

DDoS attacks can be classified into four main categories. These include *flood* attacks, e.g., UDP and ICMP floods, *amplification* attacks, e.g., smurf and fragile attacks, *protocol exploit* attacks, and *malformed packet* attacks. A detailed overview can be found in [28]. The thesis focuses on

the first type, which is the most commonly observed DDoS attack type [102]. Contrary to logic attacks, flooding attacks usually need not exploit vulnerabilities at the victim; lauching the attack with a large quantity of zombies is enough to guarantee its success. The next section presents an overview of DDoS defense systems.

1.3 DDoS Defense Systems

Preventing and detecting DDoS flooding attacks are very arduous tasks. The greatest challenge is finding the most descriptive set of Internet traffic features to effectively distinguish legitimate from malicious activities. In general, flooding attack packets do not present any characteristic that could be directly used for their detection and filtering. Furthermore, the dynamic and complex nature of the Internet makes it difficult to model and study network traffic. Additional obstacles include the lack of a centralized control system for Internet and the insufficiency of detailed attack information [28, 91, 133].

DDoS defense systems can be classified as preventive or reactive. Detailed reviews of defense models appear in [28, 82, 102]. The defense layer can be achieved by increasing security, e.g., installing secure protocols and strong authentication systems. Attack detection consists of designing efficient models that can assert whether an attack is taking place or not, by observing traffic statistics. In this case, detection delay is key.

Moreover, a DDos defense mechanism can be autonomous or deployed as part of a distributed system [21, 81]. The latter type can be difficult to implement since independent networks have different control and security policies. Autonomous systems can be set up at three main points in a given network: the victim-end, the source-end or at an intermediate point in the network. The main advantage of victim-end defense mechanisms is that network traffic can be observed locally and modeled so that anomalies can be detected. However, once at the victim-end, the attack traffic has reached its full potential and defense systems can also be overwhelmed and fail to perform their designated task. This outcome can be prevented by choosing an intermediate network point away from the target, the main disadvantage being that incoming attack traffic cannot be fully observed. Since flooding attacks are more likely to be detected near the target, the present thesis focuses on developing victim-end detection models.

Furthermore, DDoS defense systems are signature or anomaly based [18]. The second type consists of monitoring Internet traffic statistics and measuring deviations from what is defined as "normal" behavior. The choice of the statistic has a significant impact on the detection performance. Examples of traffic metrics include *packet loss* (i.e., number of data packets or bytes lost due to the attack), *traffic throughput* (i.e., number of bytes transferred per unit time from the source to the destination), *response delay* (i.e., time interval between when a request is made and the time a complete response is received from the destination), *transaction duration* (i.e., time required for

the exchange of meaningful set of messages between a source and a destination), and *resources allocation* (i.e., proportion of a critical resource (e.g. bandwidth) allocated to legitimate and attack traffic) [83]. Additional statistical methods applied to DDoS detection are presented in [35]. Chen and Hwang [21] propose an DDoS detection model that monitors traffic volume at various routers in the network. The main drawback with this technique is that at earlier stages of an attack, traffic volume changes are hard to detect since traffic fluctuations are very subtle and similar to normal traffic. While normal traffic presents a large range of packet sizes, attack data are usually of the same size. Therefore, Du and Abe [29] propose to analyze the entropy of packet size distribution, which is more effective than observing traffic volume since short-term attacks do not necessarily alter traffic volume. This technique performs well when all attack packets have the same size. However, attackers may generate data similar to the victim's legitimate traffic [35, 91], i.e., with variable packet sizes, in which case the entropy method is no longer effective. Peng et al. [101] apply a sequential change point detection method where they evaluate the increase in the percentage of unique new source IP addresses compared to a threshold. Additional sequential change-point detection models include [116, 123]. An extensive review is presented in [18].

Our detection models evaluates the percentage of new source IP addresses in a given observation window, but unlike [101], each individual source IP address is accounted for, regardless of duplicates. The next section gives an overview of the research and presents the detection problem.

1.4 Research Overview

This section states the research problem and objectives. It also emphasizes the main contributions of the thesis and finally gives an outline of the report.

1.4.1 Problem Statement

Consider a network of computers, which is subjected to a flooding DDoS attack. Define the *control router* as the router through which incoming Internet traffic gain access to the target network. Assume attack and legitimate packet flows arrive independently at the control router. As Figure 1.4 illustrates, only one flow of data packets comprising both type of traffic is observed at the victim-end. Using an anomaly-based approach, the objective of the research is to develop efficient and easy-to-implement techniques to determine when a DDoS attack is underway. Efficiency is measured in terms of detection delay, False Alarm (FA), and False Clear (FC) rates.

Definition 1. *A false alarm occurs when the detection system signals an attack when in fact only legitimate traffic is observed.*

Definition 2. *A false clear occurs when an attack is taking or has taken place without raising an alarm from the detection system.*

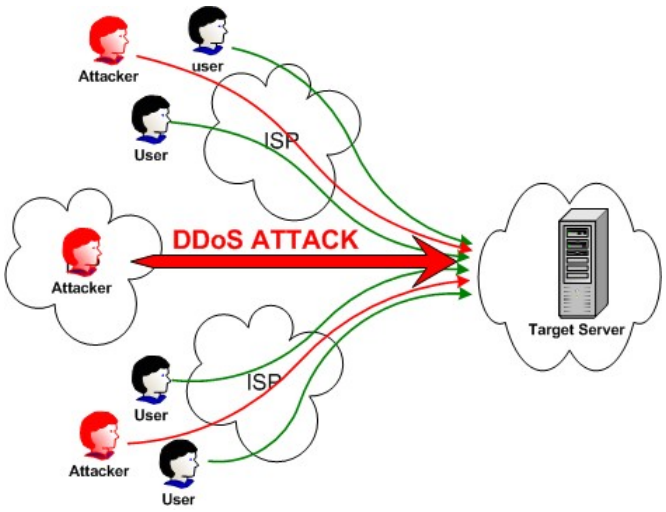


Figure 1.4: Observed Attack and Legitimate Flow

As both FA and FC are detrimental to the victim, the detection mechanisms must be as accurate as possible so that the FA and FC rates are minimized. Moreover, detection must occur as soon as possible. The earlier an attack is detected, the more effective the response can be.

The most common detection models are non-parametric. Traffic statistics are directly collected from simulated or real network traces; thresholds and other model parameters are computed numerically. Given that Internet traffic is very dynamic and complex, non parametric models can be very efficient. On the other hand, fitting an adequate theoretical model to network traffic allow applying detection theory results that are proven to be very efficient. In this case, thresholds can be determined analytically. Parametric models tend to be more elaborate as they require a combination of estimation and detection theories. We present both parametric and non-parametric detection models.

The main statistic of interest in this thesis is the percentage of new source IP addresses. The method proposed by Peng et al. [101] can be circumvented if the attacker use a small number of spoofed source IP addresses. In other words, the percentage of *distinct* IP addresses will not increase enough to warrant an alarm if attack data packets appear to come from only few unknown sources, since legitimate data packets can also come from unknown sources. We propose a slightly different approach and account for every unknown IP individually, regardless of duplicates. Therefore our statistic is independent of the number of spoofed IP addresses.

As mentioned earlier, designing a DDoS detection system that minimizes both the FA and FC rates is very challenging since these are contradictory objectives. First, DDoS attack data packets, especially in the flooding attack case, do not present specific characteristics that clearly distinguish them from legitimate traffic data. Second, legitimate traffic continues to arrive at the control router even as the attack occurs, which renders the detection problem even more difficult. For instance,

even though attack data packets are more likely to originate from an unknown source, legitimate traffic also generate data packets with unknown source IP addresses. Although the rate of legitimate data packet with unknown IP is usually smaller compared to what attack traffic produces, the DDoS detection problem does not become any simpler. Another issue is that attackers constantly modify their tools [82]. In addition to simulated data, real network traces are collected to evaluate our detection models. The next section present the two types of traces we used.

1.4.2 Network Traces

To evaluate our detection models, we use real Internet traffic as our legitimate background traffic. Attack data is simulated and combined with legitimate packets to represent a network under DDoS attack. The first dataset is the 1999 DARPA dataset obtained from MIT Lincoln Laboratory [72]. The dataset was filtered to keep inbound traffic only. We use the first week (W1) of data for training and week three (W3) data for testing; week two was ignored because it contains attack packets other than DDoS packets. The second set of traces originates from the University of Waikato, which is a much larger network than the Lincoln laboratory network. Multiple traces are available, we selected Auckland IV, which was the training dataset used in [55]. We focus on inbound traffic as well. For both the MIT and Waikato networks, data packets sent through the TCP protocol represent the largest percentage of data packets. For instance MIT's W1-M contains 623857 incoming TCP data packets out of a total of 651218 data packets, which makes 95.8% of the traffic. In general TCP traffic represents more than 80% of traffic [55]. Therefore, we filter and focus on TCP data packets. The relevant input parameters to our detection models include packet arrival time, source/destination IP addresses, packet size, protocol and ports addresses.

DDoS attack data are simulated and combined with legitimate data packets according to packet arrival time. Monte Carlo simulations (rather than using real DDoS traces) allow us to control attack parameters including the rate, duration and packet size. Table 1.1 shows different DDoS parameters that were used in the simulation. Base on the literature [15, 18, 29, 83, 84, 86], three levels were defined for each parameter.

Table 1.1: Simulation Parameters

Parameters	Low	Medium	High
Rate λ (pps)	10	160	500
Duration d (s)	600	1800	3600
Packet Size (bytes)	40	576	1500

1.4.3 Main Contributions

The main contributions of the thesis can be summarized as follows. While the vast majority of DDoS detection models are non-parametric, we introduce two efficient parametric detection systems. We apply results from detection theory to prove the effectiveness of the proposed methods. Furthermore, the parametric models are not only applicable to network traffic but other areas such as modeling stock returns in finance and total claim in insurance.

The second major contributions is that our detection models have been developed to handle stealthy and realistic attacks. For instance, we suppose that attackers can apply the source IP address spoofing technique such that attack data packets appear to originate from what the target network considers as trustworthy sources. Our models also integrate the possibility that attack data packets arrive at a dynamic rate.

Another contribution is that we develop traffic statistics that are more effective. For example, we monitor the percentage of unknown source IP addresses, regardless of duplicates. Unlike Peng et al. [101], our test statistic does not depend on the number of spoofed IP addresses used by the attacker. More detailed contributions are summarized at the end of each chapter. Next, we give a brief outline of the thesis.

1.4.4 Thesis Outline

The detection models presented in this thesis can be classified as parametric and non-parametric. Each chapter is dedicated to a specific model. The thesis is organized as follows. Chapter 2 gives an overview of some mathematical tools that are applied throughout the thesis. Chapter 3 and 4 both present parametric detection models, which are formulated as binary hypothesis detection problems. The former describes a fixed sample detection system while the latter addresses sequential detection. An empirical detection model, i.e., non-parametric is presented in Chapter 5. Finally, Chapter 6 provides concluding remarks and discusses potential extensions of the work presented here.

CHAPTER 2

NETWORK TRAFFIC MODELING AND METHODOLOGIES

This chapter presents the marked Poisson process used to model network traffic. It also introduces various mathematical tools that are applied throughout the thesis.

2.1 Network Traffic Model

Understanding and modeling the behavior of network traffic is essential for developing efficient security systems. Several network traffic models have been proposed, see [8, 129] for detailed reviews. Wilson [129] emphasizes that packet length and packet inter-arrival distributions are the two main characteristics of traffic models. The others, including routes, distribution of destinations are of less importance. Various models such as Poisson [76], compound Poisson [60], Markov-modulated Poisson [53], and Packet Trains [60] have been developed for network traffic. In this study, we focus on packet inter-arrival distributions and model the arrival of packet flows. A packet flow is a set of data packets exchanged between two end points [55]. More specifically,

Definition 3. *A flow is a collection of data packets which share the same source/destination IP addresses and source/destination ports. Furthermore, the interarrival time between any two data packets cannot exceed a certain time period, here fixed as 64s [55].*

Definition 4. *The cluster size of a given packet flow denote the total number of data packets in that flow.*

Assuming that packet flows are statistically independent, Hohn et al. [55] introduces the Poisson cluster process as a novel approach to modeling network traffic. Faÿ et al. [34] provides an extensive analysis of Poisson cluster models. In this thesis, we model legitimate network traffic as a Marked Poisson process. Let T_j denote the random arrival time of the first data packet in each flow j. The arrival of packet flows can be either a homogeneous or inhomogeneous Poisson process. In the absence of a closed form for the intensity function, the latter can be approximated as a piecewise homogeneous Poisson process. We present detection models for both instances of packet flows. First, we address the modeling of flow size, which can be measured in number of packets or bytes.

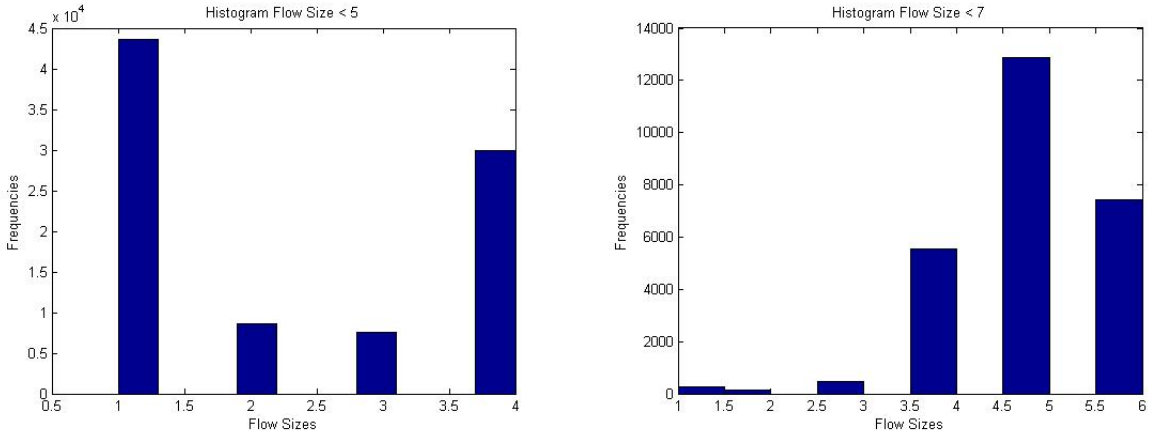


Figure 2.1: Histogram Small Flow Size- Auckland and W1M-MIT Datasets

2.1.1 Modeling Cluster Size

Define K_j as the random cluster size of flow j. Previous studies [99, 128] propose that network traffic be modeled with heavy-tailed distributions that exhibit power law behavior. The two properties are defined below [4].

Definition 5. *For all $\theta > 0$, the distribution function of a random variable X is said to be heavy-tailed if*

$$\lim_{x\to\infty} e^{\theta x}\mathbb{P}\{X > x\} = \infty. \tag{2.1.1}$$

Definition 6. *The distribution function of a random variable X has a power law behavior if*

$$\mathbb{P}\{X = x\} = L(x)x^{-\alpha}, \tag{2.1.2}$$

where $\alpha > 0$, and $L(x)$ is a slowly varying function, i.e., $\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1$, for a constant $t > 0$.

The population of network flow sizes is non-homogeneous, suggesting mixture distributions for cluster sizes. Power law tail behavior usually occurs far in the distribution tail. Small flow sizes must then be modeled separately. González-Arévalo and Roy [48] propose to model small clusters as a discrete uniform random variable. However, Figure 2.1 invalidates the uniform distribution assumption. Instead, we define a discrete distribution on the set $\{1, 2, \ldots, \lfloor c_1 \rfloor - 1\}$, $c_1 > 0$ and $0 < q_1 < 1$. Note that the first packet of each flow counts as part of the cluster size.

Let $\{K_j\}$, $K_j \geq 1\ \forall j$, be iid mixture random variables such that

$$K_j = \begin{cases} D_j & \text{with probability } q \\ \lfloor V_i \rfloor & \text{with probability } q_1 \\ \lfloor W_j \rfloor & \text{with probability } q_2 = 1 - q - q_1, \end{cases} \tag{2.1.3}$$

where (V_j) are iid Pareto random variables with complementary cdf

$$\mathbb{P}\{V_j > x\} = \begin{cases} \left(\frac{c_1}{x}\right)^{\alpha_1} & \text{for } x \geq c_1 \\ 0 & \text{for } x < c_1, \end{cases} \tag{2.1.4}$$

where $0 < \alpha_1$. $\{W_j\}$ are defined likewise, except $\alpha_2 > 0$ and $c_2 > 0$. $\{D_j\}$ are iid discrete random variables with the domain $\{1, 2, \ldots, \lfloor c_1 \rfloor - 1\}$ with probabilities $p_1, p_2, \ldots, p_{c_1-1}$.

The marked Poisson process, denoted (T_j, K_j), can be applied to model not only legitimate traffic but also the aggregated attack traffic sent from the zombie machines in the botnet. Suppose there are Z zombies. Independently of the total number of data packets in each flow, assume each zombie sends attack packet flows at the constant rate of $\tilde{\lambda}$ packets per second (pps). Let p be the probability of that one attack data packet arrives in the interval $[0, t]$.

Lemma 1. *For Z large and p small, the aggregated packet flows at the control router follows a Poisson process with rate $Z\lambda t$.*

Proof. The proof follows [29]. Define $N_z(t)$ as the number of packet flows that arrive from zombie z in the time interval $[0, t]$, $t < 1/\tilde{\lambda}$. P_k is the probability that k attack packet flows arrive in the time interval $[0, t]$, $t < 1/\tilde{\lambda}$ from Z zombies.
$p = \mathbb{P}\{N_z(t) = 1\} = \tilde{\lambda}\, t$ and $q = P\{N_z(t) = 0\} = 1 - p = 1 - \tilde{\lambda}\, t$. P_k is a binomial random variable such that

$$P_k = \mathbb{P}\left\{\sum_{i=1}^{Z} N_z(t) = k\right\} = \binom{Z}{k} p^k q^{Z-k}$$

Recall that a DDoS attack may involve dozens, hundreds or as many as 60,000 zombies [112, 124]. So assume Z is large enough, then the binomial distribution can be approximated by a Poisson distribution such that

$$P_k \approx \frac{(Z\,\tilde{\lambda}\, t)^k}{k!}\, e^{-Z\,\tilde{\lambda}\, t}$$

The approximation holds since

$$\mathbb{E}\left[e^{i\theta \sum_{i=1}^{Z} N_z(t)}\right] = (1 - p + pe^{i\theta})^Z = (1 + p(e^{i\theta} - 1))^Z$$

For p small and Z sufficiently large, $(1 + p(e^{i\theta} - 1))^Z \to e^{Zp(e^{i\theta}-1)}$. The mean of the Poisson distribution is approximated as Zp. □

The next section shows how the marked Poisson process is extended to model and monitor the percentage of unknown source IP addresses for each observation window.

2.1.2 Modeling Source IP Addresses

Assume a library of *known* source IP addresses, denoted SIL, is built and updated regularly. The library contains addresses with which the target network communicates the most frequently. Additional criteria to build an effective SIL are presented in [102]. A data packet is labeled unknown if its source IP address is not recorded in SIL. Attack data packets usually originate from sources that are unknown to the target network. However, legitimate traffic can also have unknown source IP addresses, e.g., a user is visiting a website for the first time.

Let ε and ς be the percentage of legitimate and attack packet flows with unknown source IP addresses, respectively. Peng et al. [102] estimate that $\varepsilon = 10 - 20\%$. Assume $\varsigma > \varepsilon$. Define B_j as a Bernoulli random variable with parameter ς or ε. The complete network traffic model is then the marked Poisson process (T_j, K_j, B_j), where B_j, K_j and T_j are mutually independent. Note that modeling network traffic at the flow level with no cluster size implies that we monitor the percentage of *distinct* unknown source IP addresses. This follows since by Definition 3, all data packets of a given flow share the same source IP address. On the other hand, when we take into account cluster size, every unknown source IP address is counted as such, regardless of duplicates. Our detection models are based on the second approach, given that the former statistic can be misleading if the attacker uses a small number of spoofed IP addresses.

The data fitting section, presented next, describes the procedure to estimate distribution parameters and provides distribution fitting results. Distribution parameters are estimated only for legitimate traffic given that attack traffic is Monte-Carlo simulated with user-input parameters.

2.1.3 Numerical Results for Data Fitting

To identify power-law behavior in empirical data, graphical models such as Q-Q and log-log plots can be used. Q-Q plots compare the empirical quantiles of the log-transformed data to the quantiles of the fitted distribution. Log-log graphs use logarithmic scales to compare the estimator of a number to its true value. If the dataset has a power law distribution, the log-log plot shows a straight line.

Hohn et al. [55] performed extensive wavelet analysis to validate the Poisson cluster model. In this section, we present several illustrations. The first step is to test the Poisson process assumption, for which we fit an exponential distribution to the flow inter-arrival times. Figure 2.2 illustrates the Q-Q plot of packet flows inter-arrival times against the exponential distribution with parameter $\mu = 0.0957$ for the Auckland IV dataset. Moreover, figure 2.3 depicts the number of flows for every 30 minutes time intervals for the W1M-MIT dataset. Therefore, the marked Poisson model should be considered inhomogeneous. The inhomogeneous intensity function can be approximated by a piecewise constant function. The estimator for ζ, the rate of legitimate packet flows arrival process, corresponds to the sample mean of flow inter-arrival times.

The next graphs illustrate that the distribution of flow cluster sizes behaves as a power law, i.e.

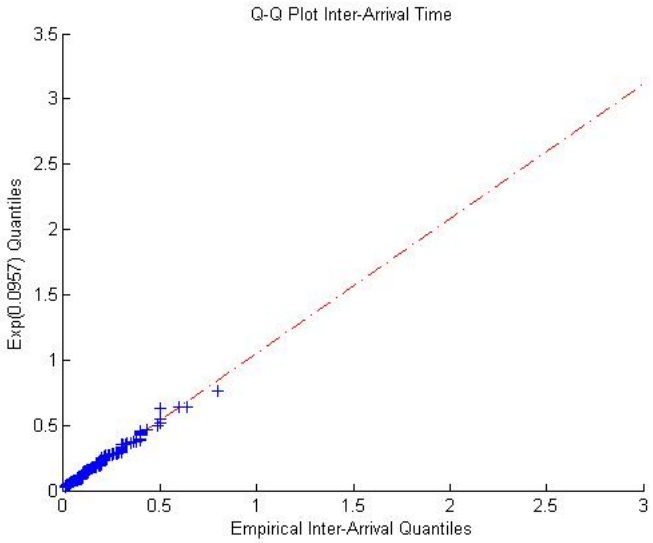


Figure 2.2: QQ-Plot Inter-arrival times vs Exponential-Auckland

$\bar{F}(x) = 1 - \mathbb{P}\{X \leq x\} \sim x^{-\alpha}$. Since heavy tailed distributions exhibit more variability on the tails [38, 14], plots should then be interpreted with care. Using Kolmogorov-Smirvov tests, Clauset et al. [22] describe a procedure to estimate the lower bound on power law behavior. The tail index parameter α is estimated using a Maximum Likelihood Estimation (MLE). Weron [127] emphasize that heavy tailness may induce bias in the tail index estimation . For a vector of size n,

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right]$$

Alternatively, α can be measured as the slope of the log-log plot $\frac{d \log \bar{F}(x)}{d \log x} \sim -\alpha$. Figure 2.4 illustrates log-log plots of two power laws with minimum 5, and 34 fitted to the Auckland dataset. The corresponding tail indices are $\alpha = 1.98, 2.34$, respectively.

The figures above show that power laws are a good fit for flow cluster sizes. Moreover, they demonstrate that one power law does not fit the entire range of cluster size, therefore mixture distributions are more appropriate. The Pareto distribution is one of the most common examples of power-law distributions. Consider the Auckland and W1M-MIT datasets. We illustrate two examples of Pareto distribution fitting when lower bounds are respectively set to 34 and 120. Figure 2.5 illustrates the Q-Q plots of empirical against the Pareto distribution quantiles. The tail indices are $\alpha = 1.434$ and $\alpha = 2.185$, respectively

Although these two examples imply that $1 < \alpha < 2$ or $\alpha > 2$, previous studies of web traffic [3, 24, 100, 107] demonstrate power laws with heavier tails, i.e. $0 < \alpha < 1$. We consider all three cases in our detection models.

Based on the network model described, we develop theoretical detection models that will be

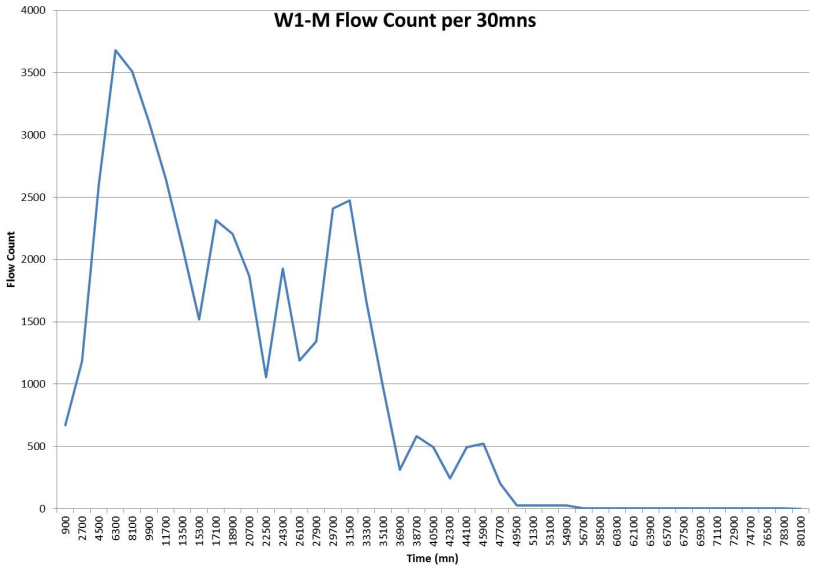


Figure 2.3: Packet Flow Count per 30 mns

validated by simulation. The following section provides some background on the Pareto distribution, and the Poisson process. Additional mathematical tools used in the thesis are introduced as well.

2.2 Poisson Processes

The Poisson process is a well-known stochastic counting process used to model the times at which certain events occur. Let $\mathcal{N}(t)$ denote the number of packet flow arrivals in the time interval $(0, t]$. If packet flows arrive at a constant rate $\lambda > 0$, $\{\mathcal{N}(t), t \geq 0\}$ is an an homogeneous Poisson process and the inter-arrival time between events are iid exponential random variables with probability density function (pdf) $f(x) = \lambda e^{-\lambda t}$, $x \geq 0$. Poisson processes have stationary and independent increment properties.

$$\mathbb{P}\{\mathcal{N}(t) = n\} = \frac{(\lambda t)^n e^{-\lambda t}}{n!} \tag{2.2.1}$$

More generally, the rate of the Poisson process may vary with time; the resulting process is called an inhomogeneous Poisson process. The corresponding intensity function is defined as

$$\lambda(t) = \int_0^t \lambda(v) dv \tag{2.2.2}$$

Poisson processes have multiple desirable properties including the thinning property, which Lemma 2 illustrates. Define $\{U(t), t \geq 0\}$ and $\{A(t), t \geq 0\}$ as the counting functions associated with the legitimate and attack packet flow with unknown source IP address, respectively.

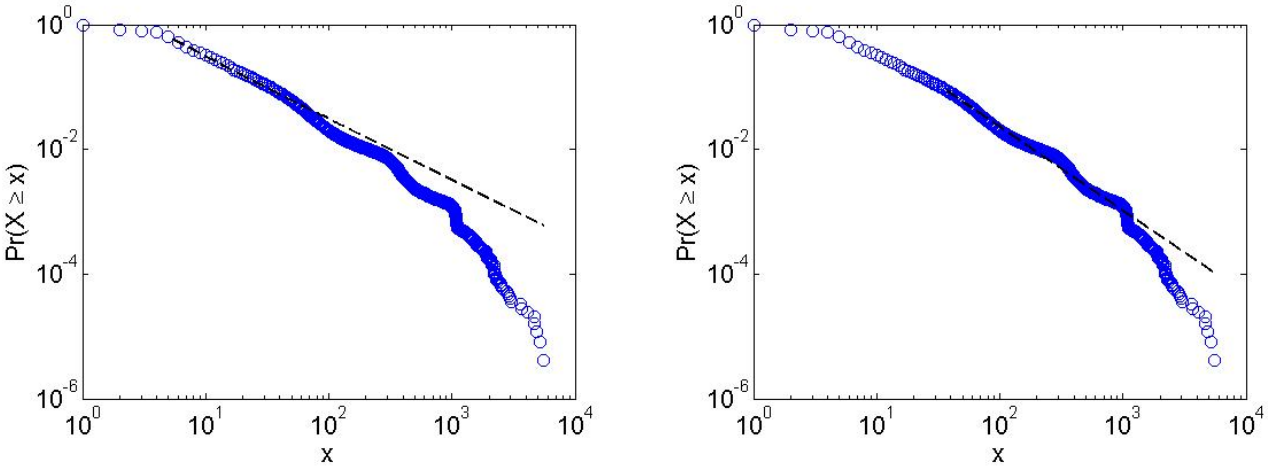


Figure 2.4: Power Law Fitting to Cluster Size - Auckland

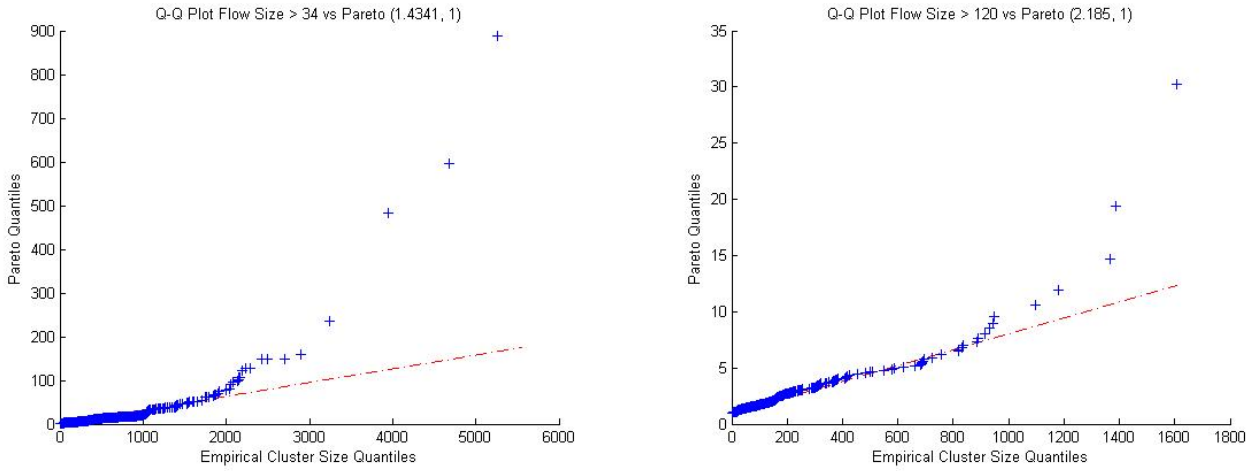


Figure 2.5: QQ-Plot Pareto vs Auckland and W1M-MIT Datasets

Lemma 2. *Independently of the cluster size K_j, $\{U(t), t \geq 0\}$ is a Poisson process with rate $\zeta\varepsilon$. Similarly, $\{A(t), t \geq 0\}$ has rate $\lambda\varsigma$ [109].*

Proof. We show the proof for $U(t)$, the thinning property similarly applies to the counting function $A(t)$.
Independently of cluster sizes $\{K_j\}$, the process of legitimate packet flow, call it $\mathcal{N}(t)$, is the Marked Poisson process (T_j, B_j) where the Poisson process has rate ζ and $B_j \sim \mathcal{B}(\varepsilon)$ and $\mathcal{B}$ stands for the Bernoulli distribution. The process $U(t)$ has independent and stationary increments and $U(0) = 0$ [109]. Moreover, given that $U(t)$ is the counting process of legitimate packet flow with unknown source IP addresses, then $U(t) = \sum_{j=1}^{\mathcal{N}(t)} B_j$.

Consider an arbitrary time interval of length T. Conditional on $\mathcal{N}(T) = n$, $U(T) \sim \mathcal{B}(n, \varepsilon)$.

$$\begin{aligned}
\mathbb{E}\left[e^{i\theta U(T)}\right] = \mathbb{E}\left[e^{i\theta \sum_{j=1}^{\mathcal{N}(t)} B_j}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{i\theta B_j}\right]^{\mathcal{N}(t)}\right] \\
&= \sum_{n=0}^{\infty} \left(\mathbb{E}\left[e^{i\theta B_j}\right]\right)^n \frac{(\zeta T)^n}{n!} e^{-\zeta T} \\
&= e^{-\zeta T} \sum_{n=0}^{\infty} \frac{\left(\left[1-\varepsilon+\varepsilon e^{i\theta}\right]\zeta T\right)^n}{n!} \\
&= e^{-\zeta T} \exp\left\{\left[1-\varepsilon+\varepsilon e^{i\theta}\right]\zeta T\right\} = \exp\left\{\zeta T\varepsilon\left(e^{i\theta}-1\right)\right\}
\end{aligned}$$

□

The network traffic model proposed in Section 2.1 is a marked Poisson with a Pareto and Bernoulli marks. Given that the Bernoulli mark space is denumerable, the thinning property of Poisson processes can be applied, as in Lemma 2, to split the packet flow traffic into two independent processes. The two processes have unknown and known source IP addresses, respectively. Next, we provide some background on probability theory. In subsequent sections, X represents a continuous random variable with pdf $f_X(x)$ and cdf $F_X(x)$.

2.3 Characteristic Functions

Define the characteristic function of X as

$$\psi_X(\theta) = \mathbb{E}\left[e^{i\theta X}\right] = \int_{-\infty}^{\infty} e^{i\theta x} f_X(x) dx \tag{2.3.1}$$

$\psi_X(\theta)$ is a complex-valued function that uniquely defines the cdf $F_X(x)$. Some properties of characteristic functions include the following ([30])

1. $\psi(\theta)$ exists for any distribution
2. $\psi(0) = 1$
3. If $Y = aX + b$ then $\psi_Y(\theta) = e^{ia\theta}\psi_X(b\theta)$
4. If X_1 and X_2 are independent, then $\psi_{X_1+X_2}(\theta) = \psi_{X_1}(\theta)\psi_{X_2}(\theta)$

The moment generating function (mgf) of a random variable X is computed as $\mathcal{M}_X(\theta) = \mathbb{E}\left[e^{\theta X}\right] = \psi_X(-i\theta)$. As opposed to the characteristic function, the mgf does not exist for certain distributions, e.g. the Cauchy distribution.

Given the characteristic function $\psi(\theta)$, the pdf f, when it exists, is computed as the inverse Fourier transform of $\psi(\theta)$ [30].

Theorem 1. *If the characteristic function is absolutely integrable then pdf is given by the inverse Fourier transform of $\psi_X(\theta)$, i.e.,*

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\theta x}\psi_X(\theta)d\theta \tag{2.3.2}$$

Proof. See [30]. □

Equation (2.3.2) is very useful for random variables without a closed-form pdf. Gil-Pelaez [45] applied the inversion theorem to derive the cdf of a random variable

$$F_{\widetilde{Y}}(y) = \frac{1}{2} + \frac{1}{2\pi}\int_0^{\infty} \frac{e^{i\theta x}\psi_{\widetilde{Y}}(-\theta) - e^{-i\theta x}\psi_{\widetilde{Y}}(\theta)}{i\theta}d\theta\ . \tag{2.3.3}$$

Waller et al. [122] provides a review of numerical methods to invert characteristic functions. Tauberian theorems shows that the behavior of $f_X(x)$ for $x \to \infty$ is mainly determined by the behavior of $\psi_X(\theta)$ near zero [36].

2.3.1 Pareto Distribution

This section introduces the Pareto distribution and proves that it possesses the required properties to be adequate models for network traffic. These properties include heavy-tailness and power law behavior, which are described in Definitions 5 and 6, respectively. The Pareto distribution is one of the most well-known power law functions. It commonly appears in economic studies where it is applied to describe the distribution of income [77]. A nonnegative random variable X has a Pareto

(type I) distribution, denoted $X \sim \text{Par}(\alpha, c)$, if for some $\alpha > 0$ and c is a positive constant,

$$f_X(x) = \frac{\alpha c^\alpha}{x^{\alpha+1}} \mathbb{1}\{x \geq c\} \ . \tag{2.3.4}$$

α and c are the shape and scale parameters, respectively. X has a power law probability tail behavior since $\bar{F}_X(x) = \mathbb{P}(X > x) = Cx^{-\alpha}$ [62]. As α decreases, the tail thickness increases. Lemma 3 gives the Laplace transform of a Pareto random variable [89].

Lemma 3. *Let X be a Pareto random variable with pdf as described in (2.3.4). The Laplace transform of X is $\mathcal{M}_X(-\theta)$, where $\mathcal{M}_X(\theta)$ denotes the mgf of X.*

$$\mathcal{M}_X(-\theta) = \alpha(c\,\theta)^\alpha \Gamma(-\alpha, c\,\theta),\ \theta > 0 \tag{2.3.5}$$

Proof.

$$\begin{aligned} \mathcal{M}_X(-\theta) &= \mathbb{E}[\exp(-\theta X)] \\ &= \int_c^\infty e^{-\theta x} \frac{\alpha c^\alpha}{x^{\alpha+1}} dx = \alpha c^\alpha \int_c^\infty e^{-\theta x} x^{-\alpha-1} dx \end{aligned}$$

Perform a change of variables i.e. set $t = \theta x$ and define the "upper" incomplete gamma function $\Gamma(a, z)$ as

$$\Gamma(a, z) = \int_z^\infty t^{a-1} e^{-t}\,, \tag{2.3.6}$$

to obtain (2.3.5). □

The Laplace transform (2.3.5) can be also expressed in terms of Whittaker functions [89]. The incomplete Gamma is more convenient for our applications. Note that heavy tailness implies that the mgf, $\mathcal{M}(\theta)$, is infinite for all $\theta > 0$, which is the case for the Pareto distribution.

If $0 < \alpha < 2$, the Pareto random variable has an infinite variance. For $\alpha \leq 1$, the mean is also infinite, else

$$\mathbb{E}[X] = \frac{\alpha c}{\alpha - 1} \tag{2.3.7}$$

In general, moments of order k exists only if $\alpha \geq k$ [110]. This follows since the k^{th} moment is computed as

$$\begin{aligned} \mu_k &= \mathbb{E}[x^k] = \int_{-\infty}^\infty x^k f(x) dx = \int_c^\infty x^k \frac{\alpha c^\alpha}{x^{\alpha+1}} dx \\ &= \alpha c^\alpha \int_c^\infty x^{k-\alpha-1} dx = \alpha c^{k-1} \int_1^\infty y^{k-\alpha-1} dx \\ &= \alpha c^{k-1} \left[\frac{y^{k-\alpha}}{k-\alpha}\right]_1^\infty \end{aligned}$$

The last two equalities follow from applying the change of variable $y = x/c$. If $k < \alpha$, the limit at ∞ in (2.3.8) is zero, else it diverges.

Unless otherwise specified, subsequent sections assume that the minimum value of the Pareto distribution is $c = 1$. Lemma 4 shows that any Pareto random variable can be normalized to satisfy this condition.

Lemma 4. *If $X \sim Par(\alpha, c)$ then $\frac{X}{c} \sim Par(\alpha, 1)$.*

Proof. The characteristic function of a Pareto random variable can be deduced from Lemma 3 as

$$\psi_X(\theta) = \alpha(-i\,c\,\theta)^{\alpha}\Gamma(-\alpha, -ic\,\theta). \tag{2.3.8}$$

Then, the characteristic function of a Pareto random variable with scale parameter $c = 1$ is

$$\begin{aligned} \psi_{X/c}(\theta) &= \mathbb{E}\left[e^{i\theta\frac{X}{c}}\right] = \psi_X\left(\frac{\theta}{c}\right) \\ &= \alpha(-i\,\theta)^{\alpha}\Gamma(-\alpha, -i\,\theta) \end{aligned}$$

□

For detection purpose, we are interested in the sum of iid Pareto random variables. In case the tail index satisfies $0 < \alpha < 2$, i.e. the variance is infinite, the sum can be approximated by a stable distribution, which is described next.

2.3.2 Stable Distributions

Definition 7. *A random variable X is said to have a stable distribution if there exists a sequence of positive numbers $\{b_n\}$ and real numbers $\{a_n\}$ such that*

$$\frac{Y_1 + Y_2 + \ldots + Y_n}{b_n} + a_n \stackrel{d}{\Rightarrow} X\ , \tag{2.3.9}$$

where Y_i's are iid random variables [110].

Definition 7 indicates that the distribution of normalized sums of iid random variables can be approximated by stable distributions [110]. An equivalent definition also due to [110] illustrates the "stability" property as

Definition 8. *Let $n \geq 2$. A random variable X is said to have a stable distribution if there exists a positive number $\{b_n\}$ and a real numbers $\{a_n\}$ such that*

$$X_1 + X_2 + \ldots + X_n \stackrel{d}{\Rightarrow} b_n X + a_n\ , \tag{2.3.10}$$

where X_i's are iid copies of X [110].

In other words, the family of stable distribution is closed under convolution. Feller [36] proves that for $0 < \alpha \leq 2$, $b_n = n^{1/\alpha}$. The univariate stable distribution $\mathcal{S}_\alpha(\gamma, \beta, \delta)$ is characterized by four parameters [110]:

- $\alpha \in (0, 2]$: index of stability
- $\gamma > 0$: scale parameter
- $\beta \in [-1, 1]$: skewness parameter
- $\delta \in \mathbb{R}$: location parameter

The normal distribution is a stable distribution with $\alpha = 2$, in which case β is irrelevant. It is the only stable distribution that has a finite second moment. The Levy and Cauchy distribution have parameters $\alpha = 1/2$, $\beta = 1$ and $\alpha = 1$, $\beta = 0$, respectively. Only these three instances of stable distribution have an explicit closed-form pdf. Other stable random variables are represented by their characteristic function

$$\psi_S(\theta) = \begin{cases} \exp\left\{-|\theta|^\alpha \left[1 + i\beta(\operatorname{sign}\theta)(\tan\frac{\pi\alpha}{2})\left(|\theta|^{1-\alpha} - 1\right)\right]\right\} & \text{if } \alpha \neq 1 \\ \exp\left\{-|\theta|\left[1 + i\beta(\operatorname{sign}\theta)(\frac{2}{\pi})\ln t\right]\right\} & \text{if } \alpha = 1 \end{cases} \tag{2.3.11}$$

Their pdf can be calculated as the inverse Fourier transform of $\psi_S(\theta)$ (Theorem 1).

The central limit theorem can be extended to random variables with infinite moments, where the stable distribution represents the approximating distribution. The tails of stable distributions decay like power law function, i.e., Pareto-like behavior [38], therefore they make good candidates to model random variables when heavy tails are observed. The smaller the α (tail index), the slower the decrease, and the heavier the tail [39]. Theorem 2 shows that the sum of n symmetric Pareto random variables can be approximated by a symmetric stable distribution ($\beta = 0$, $\delta = 0$). A similar version of Theorem 2 is known as the Generalized Central Limit Theorem (GCLT) [119].

Theorem 2. *[67] Let X_1, X_2, ..., X_n be iid with a probability distribution which behaves asymptotically as*

$$f_X(x) \propto \frac{C}{|x|^{1+\alpha}} , \tag{2.3.12}$$

where C is a constant and $0 < \alpha < 2$. Let $S_n = X_1 + X_2 + \ldots X_n$ then $S_n \sim \mathcal{S}_\alpha(\gamma, 0, 0)$

The central limit theorem with the normal distribution does not apply when the random variables X_i are heavy-tailed with an infinite variance. Proving Theorem 2 requires Lemma 5, which provides the small θ expansion of the characteristic function of X_i's.

Lemma 5. *For any i, the small θ expansion of $\psi_{X_i}(\theta)$ is*

$$\psi_{X_i}(\theta) \approx 1 - \widetilde{C}|\theta|^\alpha , \tag{2.3.13}$$

with

$$\widetilde{C} = \frac{\pi \alpha c^{\alpha}}{\Gamma(1+\alpha)\sin(\pi\alpha/2)} \tag{2.3.14}$$

Proof. By definition $\psi_{X_i}(\theta)$ is the Fourier transform of the pdf $f_{X_i}(x)$, i.e.,

$$\begin{aligned} \psi_{X_i}(\theta) &= \int_{-\infty}^{\infty} e^{i\theta x} f_{X_i}(x) dx \\ &= 1 - (1 - \psi_{X_i}(\theta)) \\ &= 1 - \int_{-\infty}^{\infty} (1 - \cos\theta x) f_{X_i}(x) dx \end{aligned}$$

The last inequality holds since $\sin\theta x$ is an odd function, therefore its integral is 0 by symmetry. Also $f_{X_i}(x)$ is a pdf which integrates to 1. Next, perform a change of integral variable and set $y = \theta x$

$$\begin{aligned} \int_{-\infty}^{\infty} (1 - \cos\theta x) f_{X_i}(x) dx &= \frac{1}{\theta} \int_{-\infty}^{\infty} (1 - \cos y) f_{X_i}\left(\frac{y}{\theta}\right) dy \\ &= \frac{1}{\theta} \int_{-\infty}^{\infty} (1 - \cos y) \frac{C}{|y/\theta|^{1+\alpha}} dy \\ &= C|\theta|^{\alpha} \times 2 \int_{0}^{\infty} \frac{(1 - \cos y)}{y^{1+\alpha}} dy \end{aligned}$$

Using the property of the cosine function that $\cos(2x) = \cos^2(x) - \sin^2(x)$ and $\cos^2(x) + \sin^2(x) = 1$, write

$$1 - \cos y = 1 - \left[\cos^2\left(\frac{y}{2}\right) - \sin^2\left(\frac{y}{2}\right)\right] = 2\sin^2\left(\frac{y}{2}\right)$$

From [51]

$$\int_0^{\infty} \frac{\sin^2(\nu)}{\nu^{1+\alpha}} d\nu = -\frac{\Gamma(-\alpha)\cos(\frac{\alpha\pi}{2})}{2^{-\alpha+1}}$$

Let $\nu = \frac{y}{2}$, then

$$2\int_0^{\infty} \frac{(1 - \cos y)}{y^{1+\alpha}} dy = \frac{8}{2^{1+\alpha}} \int_0^{\infty} \frac{\sin^2(\nu)}{\nu^{1+\alpha}} d\nu = -2\Gamma(-\alpha)\cos(\frac{\alpha\pi}{2})$$

Applying Euler's reflexion formula, i.e., $\Gamma(1-z)\Gamma(z) = \frac{\pi}{\sin(\pi z)}$, rewrite $\Gamma(-\alpha) = -\frac{\pi}{\sin(\pi\alpha)} \frac{1}{\Gamma(1+\alpha)}$. Since $\sin(\pi\alpha) = 2\sin(\pi\alpha/2)\cos(\pi\alpha/2)$

$$2\int_0^{\infty} \frac{(1 - \cos y)}{y^{1+\alpha}} dy = \frac{\pi}{\Gamma(1+\alpha)\sin(\pi\alpha/2)} \ ,$$

and

$$\int_{-\infty}^{\infty}(1-\cos\theta x)f_K(x)dx = \widetilde{C}|\theta|^{\alpha}$$

Given $0<\alpha<2$, (2.3.13) follows. □

The proof of Theorem 2 follows since

$$\psi_{S_n}(\theta) = (1-\widetilde{C}|\theta|^{\alpha}+\ldots)^n \to \exp\left\{-n\widetilde{C}|\theta|^{\alpha}\right\} ,$$

which is the characteristic function of symmetric α-stable distributions [36]. If $\alpha>2$, (2.3.13) is not a characteristic function; a proof can be found in [67].

The next sections present some approximation techniques that can be used to compute the pdf of a random variable with known characteristic function.

2.4 Saddlepoint Approximation

Let X be a random variable. If its mgf $\mathcal{M}_X(\theta)$ is known, the pdf $f_X(x)$ can approximated by using the saddlepoint technique. Theorem 3 illustrates the concept, which was first introduced by Daniels in [25].

Theorem 3.

$$f_X(x) \simeq \frac{\mathcal{M}_X(\hat{s})e^{-x\hat{s}}}{\sqrt{-2\pi\mathcal{K}_X''(\hat{s})}} , \tag{2.4.1}$$

where $\hat{s}$ is the saddlepoint which satisfies $\mathcal{K}_X(\hat{s}) = \ln\mathcal{M}_X(\hat{s}) = x$.

Proof. Recall $\mathcal{M}_X(\theta)$ the moment generating function of a random variable X with pdf $f_X(x)$ is

$$\mathcal{M}_X(\theta) = \int_{-\infty}^{\infty} e^{\theta x} f_X(x)dx .$$

Given that $\psi_X(\theta) = \mathcal{M}_X(i\theta)$, the pdf is given as the inverse Fourier [30],

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\bar{\theta}x}\mathcal{M}_X(i\bar{\theta})d\bar{\theta} . \tag{2.4.2}$$

By a change of variable $\theta = i\bar{\theta}$, Daniels [25] rewrites the density above as

$$f_X(x) = \frac{1}{2\pi i}\int_{\tau-i\infty}^{\tau'+i\infty} e^{\mathcal{K}_X(\theta)-\theta x}d\theta , \tag{2.4.3}$$

where $\mathcal{K}_X(\theta)$ is the cumulant generating function of X and satisfies $\mathcal{K}_X(\theta) = \ln\mathcal{M}_X(\theta)$. By the Closed Curve Theorem, the choice of τ can be arbitrary since the integral (2.4.3) is the same over

all paths that are parallel to the imaginary axis in a neighborhood of zero where $\mathcal{M}_X(\theta)$ exists [49]. The pdf f_X can then be approximated by choosing the path of integration to pass through a saddlepoint $\hat{s}$ which satisfies $\mathcal{K}'_X(\hat{s}) = x$. In other words, choose τ in (2.4.3) to be $\tau = \hat{s}$. Note that $\hat{s}$ is a function of x. The integrand is negligible outside the immediate neighborhood of $\hat{s}$ [25]. Therefore it is sufficient to use the Taylor expansion of $\mathcal{K}_X(\theta) - \theta x$ around the saddlepoint $\hat{s}$, which is

$$\mathcal{K}_X(\theta) - \theta x \simeq \mathcal{K}_X(\hat{s}) - \hat{s}x + \frac{(\theta - \hat{s})^2}{2}\mathcal{K}''_X(\hat{s}) + \mathcal{O}(\theta^3) \tag{2.4.4}$$

The linear term in the expansion vanishes given that $\hat{s}$ is an extremum for $\mathcal{K}_X(\theta) - \theta x$. The next step consists of substituting the exponent (2.4.4) in (2.4.3) and integrating with respect to θ along the line parallel to the imaginary axis through the saddlepoint $\hat{s}$. Using the well-known results for Gaussian integrals, i.e., $\int_{-\infty}^{\infty} e^{-ax^2} = \sqrt{\frac{\pi}{a}}$ leads to the desired approximation [132]. □

If the saddlepoint method is used to approximate the pdf of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ then the approximation presents a relative error of order $\mathcal{O}(1/n)$, which can be reduced to $\mathcal{O}(n^{-3/2})$ by normalized the resulting pdf so that it integrates to one. It is a very accurate approximation which achieves high precision in the tails as well. For instance, the saddlepoint approximation is exact for the normal distribution.

$$\mathcal{M}_X(t) = \exp\{\frac{1}{2}t^2\} \tag{2.4.5}$$

Let the cumulant generating function be

$$\mathcal{K}_X(t) = \ln \mathcal{M}_X(t) = \frac{1}{2}t^2 \tag{2.4.6}$$

The first and second derivatives, $\mathcal{K}'_X(t)$ and $\mathcal{K}''_X(t)$ are respectively t and 1.
Solving $\mathcal{K}'_X(t) = x$ gives $t^* = x$.
The saddle point approximation is given by

$$f_X(x) \approx \frac{1}{\sqrt{2\pi\mathcal{K}''_X(t^*)}} \exp\{\mathcal{K}_X(t^*) - xt^*\} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{2.4.7}$$

By considering more terms in the Taylor expansion (2.4.4) a second order saddle can be defined as

$$f_X(x) \simeq \frac{\mathcal{M}_X(\hat{s})e^{-x\hat{s}}}{\sqrt{-2\pi\mathcal{K}''_X(\hat{s})}} \left[1 + \frac{1}{8}\frac{\mathcal{K}^{(4)}_X(\hat{s})}{\left(\mathcal{K}''_X(\hat{s})\right)^2} - \frac{5}{24}\frac{\left(\mathcal{K}^{(3)}_X(\hat{s})\right)^2}{\left(\mathcal{K}''_X(\hat{s})\right)^3}\right] \tag{2.4.8}$$

However, applying the higher order does not necessarily improve the approximation. In fact, the approximated pdf may "explode" , i.e. reach very large values [46].

Lugannani and Rice [75] applied the saddlepoint approximation technique to determine the tail distribution of the sum of n iid random variables. The complementary cdf, i.e. $\bar{F}_X(x) = \mathbb{P}\{\widetilde{Y} \geq x\}$,

is given as

$$\bar{F}_X(x) \simeq \begin{cases} 1 - \Phi(\hat{w}) + \phi(\hat{w}) \left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right) & x \neq \mathbb{E}[\widetilde{Y}] \\ \frac{1}{2} - \frac{\mathcal{K}'''_X(0)}{6\sqrt{2\pi}\left(\mathcal{K}''_X(0)\right)^{3/2}} & x = \mathbb{E}[\widetilde{Y}] \end{cases} \tag{2.4.9}$$

where Φ and ϕ are the cdf and pdf of the standard normal distribution, respectively. In (2.4.9)

$$\hat{w} = \text{sign}(\hat{s})\sqrt{2\hat{s}x - 2\mathcal{K}_{\widetilde{Y}}(\hat{s})} \tag{2.4.10}$$

and

$$\hat{u} = \begin{cases} \hat{s}\sqrt{\mathcal{K}''_{\widetilde{Y}}(\hat{s})} & \text{if } x \text{ is continuous} \\ \left(1 - e^{-\hat{s}}\right)\sqrt{\mathcal{K}''_{\widetilde{Y}}(\hat{s})} & \text{if } x \text{ is discrete} \end{cases} \tag{2.4.11}$$

Note that if $y = \mathbb{E}[\widetilde{Y}]$, then $\mathcal{K}'_{\widetilde{Y}}(0) = \mathbb{E}[\widetilde{Y}]$ and $\hat{s} = 0$ is the saddle point for $\mathbb{E}[\widetilde{Y}]$ [17]. In that case $\hat{w} = 0$ and $\bar{F}_X(x)$ must be approximated otherwise, hence the second formula. Wood et al. [131] extend this result to allow general base distributions other than the normal distribution. The saddlepoint approximation method does not always apply, especially when the pdf is heavy-tailed with infinite variance. An efficient alternative method is the Fast Fourier Transform, which is presented in the next section.

2.5 Fast Fourier Transform (FFT)

Consider that the pdf f_X of a random variable X has to be approximated on the interval $[a, b]$. X has characteristic function $\psi_X(\theta) = \mathbb{E}\left[e^{i\theta X}\right]$ and f_X is the inverse Fourier transform of $\psi_X(\theta)$ (see (2.3.2)). The pdf of any random variable with known characteristic function can be approximated using the FFT method. Given that all random variables possess a characteristic function, the method has a wide range of application. FFT method is particularly useful when the characteristic function has a very complex form such that analytical results are difficult to derive.

Hürlimann [57] presents an FFT method combined with quadrature approximations. The approach requires to partition $[a, b]$ into N equally spaced subintervals of length $h = (b - a)/N$. Let $c = \pi/h$ and $x_k = a + hk$ for $k = 0, 1, \ldots, N - 1$. For a random variable X, the characteristic function and the Fourier transform $\hat{f}$ satisfy $\psi_X(-2\pi\theta) = \hat{f}(\theta)$. Therefore, given N sufficiently large, c is also large and

$$\begin{aligned} f_X(x_k) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\theta x_k}\psi_X(\theta)d\theta \simeq \frac{1}{2\pi}\int_{-c}^{c} e^{-i\theta x_k}\psi_X(\theta)d\theta \\ &= \int_{-N/2(b-a)}^{N/2(b-a)} e^{-2\pi u x_k}\psi_X(2\pi u)du \end{aligned} \tag{2.5.1}$$

The density function f_X is obtained by numerically inverting the characteristic function ψ_X, which is approximated by the Discrete Fourier Transform (DFT). When N is defined as a power of two, then the FFT algorithm can be applied to compute the DFT. The detailed procedure is described in the appendix. The cumulative distribution function of X is computed recursively as

$$\begin{cases} F_X(x_k) = 0 & k = 0 \\ F_X(x_k) = F_X(x_{k-1}) + h f_X(x_{k-1}) & k = 1, 2, \ldots, N-1. \end{cases} \tag{2.5.2}$$

Note that interpolation can be used to obtain values between the points x_k. Numerical rounding errors can occur and the results may contain complex components, in which case, only the real components must be taken into consideration [32]. For a discussion on the accuracy of the FFT approximations based on the discrete total variation distance see [57]. To achieve better approximations, exponential tilting may be used prior to applying the FFT algorithm [32, 111]. Alternatively, the grid size N can be increased to the expense of higher computational costs. Another method is the "padding by zeros" technique which consists of setting elements of the vector to zero [111].

So far we have described different techniques to approximate pdfs. These tools are to used to solve our DDoS detection problem which can be formulated as a binary hypothesis problem. The next section gives an overview of detection theory.

2.6 Detection Theory

Detection theory, as the name indicates, comprises tools to detect an information of interest distorted by some random patterns called noise. Applications include target detection, signal processing, biomedecin, and communication [103]. Detection problems can usually be formulated using hypothesis testing, as shown below.

2.6.1 Hypothesis Testing

A binary hypothesis test, with state space $j \in \mathcal{X} = \{0, 1\}$, consists of evaluating observations Y_1, Y_2, ..., $Y_n \sim f(y|\theta)$ to determine which of the following hypothesis holds

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases} \tag{2.6.1}$$

H_0 is known as the *null* hypothesis and H_1 is the *alternative* hypothesis. If the set Θ contains a single element, then the corresponding hypothesis is qualified as *simple*, otherwise it is *composite*. In our DDoS detection case, H_0 would be the hypothesis that only legitimate packet flows are observed, while H_1 would be a combination of attack and legitimate packet flows. The definition

of H_1 takes into account the fact that legitimate users continue to access the victim network while the attack is taking place.

Let $\mathcal{Y}$ be the space of possible observations $\{Y_i\}$. The testing procedure involves partitioning the observation space $\mathcal{Y}$ into two mutually exclusive regions, the rejection $\mathcal{R}$ and acceptance $\mathcal{A}$ regions [88]. If an observation Y_i falls into $\mathcal{R}$, then Y_i is said to be a threat and an alarm is triggered, else Y_i is labeled as legitimate and no further action is taken. In binary hypothesis testing, two error probabilities can occur.

Definition 9. *The type I error probability, also known as the probability of false alarm, is defined as* $P_{FA} = \mathbb{P}\{\textit{Decide } H_1 | H_0 \textit{ is true}\}$.
The type II error probability, i.e. $P_{FC} = \mathbb{P}\{\textit{Decide } H_0 | H_1 \textit{ is true}\}$*, is the probability of false clear or miss.*

In testing two simple hypothesis, the Neyman-Pearson test can be applied to identify the regions $\mathcal{R}$ and $\mathcal{A}$. Let $f_{Y|H_0}(y)$ and $f_{Y|H_1}(y)$ be the conditional pdfs under H_0 and H_1, respectively. The likelihood ratio is defined as the ratio of the two conditional pdfs.

$$\mathbb{L}(y) = \frac{f_{Y|H_1}(y)}{f_{Y|H_0}(y)} \tag{2.6.2}$$

Definition 10. *Let* φ *be a fixed limit on the probability of false alarm. The Neyman-Pearson test rejects* H_0 *if* $\mathbb{L}(y) > \eta$*. The threshold* η *is chosen in order to minimize* φ *under the constraint that* ξ *does not exceed a predetermined limit.*

In general, practical application of hypothesis testing will require both error probabilities φ and ξ to be small, i.e., $\varphi, \xi \in [0.01, 0.05]$ [121]. However, a tradeoff must be made between the false alarm and false clear probabilities. The performance of a particular detector can be measured based on the values of these error probabilities. For hypothesis testing with no closed-form pdf or very complex pdfs, φ and ξ cannot be determined analytically. In that case, large deviation theory can be applied to computed upper bounds on these error probabilities. The next section defines such a procedure.

2.6.2 Large Deviations

Large deviation theory is concerned with the behavior of the probability $\mathbb{P}\{X > x\}$ when $x \to \infty$. These results apply to detection problems to measure the type I and type II error probabilities. When it is impossible to compute the exact probabilities, large deviation bounds provide upper limits on the probabilities of false alarms and false clear (see Definition 9). One of the most famous large deviation result is the Chernoff bound. The latter is an extension of the Markov's and Chebyshev's inequalities, which are introduced next.

For a given nonnegative random variable X and a constant $b > 0$, Markov's inequality is given as

$$\mathbb{P}\{Z > b\} \leq \frac{\mathbb{E}[Z]}{b} . \tag{2.6.3}$$

Given c, a positive constant, the Chebyshev's inequality follows by applying the Markov's Inequality with $b = c^2$ and $Z = (X - \mathbb{E}[X])^2/\mathrm{Var}(X)$

$$\mathbb{P}\left\{\frac{|X - \mathbb{E}[X]|}{\sqrt{\mathrm{Var}(X)}} > c\right\} \leq c^{-2} , \tag{2.6.4}$$

The Markov's and Chebyshev's inequalities are usually loose bounds, therefore a tighter exponential bound is introduced by Chernoff. The main difference between Markov's Inequality and Chernoff bounds is that the latter relies on the moment generating function of the random variable of interest, whereas the former uses only the mean of the distribution. Since the mgf encodes the entire probability distribution, Chernoff bounds are tighter (if the mgf is finite).

Let $Z = e^{sX}$ be a nonnegative random variable and $b = e^{sa}$, for $a \in \mathbb{R}$; the variable s is a free parameter that has to be optimized. Given $s > 0$, the function e^{sa} is increasing and the exponential bound is deduced from Markov's inequality as

$$\mathbb{P}\{X > a\} = \mathbb{P}\left\{e^{sX} > e^{sa}\right\} \leq \frac{\mathbb{E}\left[e^{sX}\right]}{e^{sa}} . \tag{2.6.5}$$

Recall the mgf of any random variable X is $\mathcal{M}_X(s) = \mathbb{E}\left[e^{sX}\right]$. As mentioned earlier, define the cgf of X as $\mathcal{K}_X(s) = \ln \mathcal{M}_X(s)$. Then the tightest exponential upper bound is achieved for $\hat{s} > 0$, where $\hat{s}$ maximizes the function $g_a(s) = sa - \mathcal{K}(s)$. Note that $g_a(0) = 0$ and $g_a'(0) = a - \mathbb{E}[X]$, which implies that $\hat{s} > 0$ for $a > \mathbb{E}[X]$ and $\hat{s} < 0$ when $a < \mathbb{E}[X]$. The large deviation function is defined as

$$g^*(a) = \sup_{s \in \mathbb{R}} \{sa - \mathcal{K}(s)\} . \tag{2.6.6}$$

The function $g_a(s)$ is concave since sa and $-\mathcal{K}_X(s)$ are both concave. Therefore finding the optimal $\hat{s}$ for (2.6.6) reduces to solving a concave optimization problem. If the resulting supremum is positive, i.e. $\hat{s} > 0$, then the upper bound is

$$\mathbb{P}\{X > a\} \leq e^{-g^*(a)} \tag{2.6.7}$$

Else if $\hat{s} < 0$ then $\mathbb{P}\{X < a\} \leq e^{-g^*(a)}$.

Cramer Theorem extends these results to deduce bounds on the tail probability of the sum of n iid random variables [88].

Theorem 4 (Cramer-Chernoff Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be iid with common mgf $\mathcal{M}_X(\theta) < \infty$*

for all θ in a neighborhood of $\theta = 0$. Define $I(a) = \sup_\theta \{\theta a - \ln \mathcal{M}_X(\theta)\}$. Then, for any $a > \mathbb{E}[X_i]$,

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i \geq an\right\} \leq e^{-nI(a)}. \tag{2.6.8}$$

The result can be stated as

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left\{\sum_{i=1}^{n} X_i \geq an\right\} = -I(a). \tag{2.6.9}$$

Equation (2.6.9) indicates that large deviation probabilities converge exponentially to zero.

Note that Cramer's bound (2.6.8) as given in Theorem 4 is relevant only if $I(a) > 0$. The latter condition is satisfied if $a > \mathbb{E}[X]$ and $\mathcal{M}_X(\theta) < \infty$ for some $\theta > 0$. In other words, the tail of X must decay exponentially. In some of our applications, the random variable of interest has a heavy tailed distribution, i.e., $\bar{F}(k) = 1 - F(k) = (ck)^\alpha$ with $\alpha > 0$ and $\mathcal{M}_K(\theta) = \mathbb{E}\left[e^{\theta Y}\right] = \infty$ for $\theta > 0$. Therefore, the standard Cramer-Chernoff bound is not applicable.

CHAPTER 3

FIXED SAMPLE DETECTION MODEL

In this chapter, the DDoS detection model is formulated as a fixed sample binary hypothesis testing problem where the threshold is selected according to the Neyman-Pearson criterion. Let Y be the percentage of new source IP address in a given observation window of length T. Two cases are addressed: Y counts the number of *distinct* unknown source IP addresses or the total number of unknown source IP addresses, regardless of duplicates. As described in Section 2.1, the arrival of data packets is modeled as a Marked Poisson process.

For detection purpose, we are interested in asymptotic tails and quantiles estimation. Let f_Y and F_Y denote the pdf and cdf of Y, respectively. Let $\varphi \in (0, 1)$ be the limit on the probability of false alarms (P_{FA}). Assume the Neyman-Pearson test, introduced in Definition 10, can be expressed in terms of the observation $Y = y$. Then η, the Neyman-Pearson threshold corresponds to the $(1-\varphi)$-quantile of the distribution function of Y, i.e. $F_Y^{-1}(1-\varphi)$. Note that as $\varphi \to 0$, the corresponding quantile $y \to \infty$ [13]. For large y, the behavior of the pdf $f_Y(y)$ is mostly determined by that of the characteristic function $\psi_Y(\theta)$ when θ is small [64, 67]. Therefore, subsequent sections put an emphasis on evaluating the small θ expansion of characteristic functions. The first section addresses the case of monitoring the percentage of distinct unknown source IP addresses.

3.1 Distinct Source IP Statistic

Suppose we are concerned with the percentage of *distinct* unknown source IP addresses [101]. Thus, cluster size of packet flows are irrelevant since all data packets in the same flow have the same source IP address. Therefore, the "reduced" marked Poisson model is (T_j, B_j) where B_j and T_j are independent. T_j denotes the random arrival times of packet flow j and B_j is the Bernoulli mark such that $B_j = 0$ if the source IP of flow j is recorded into SIL (known), else $B_j = 1$ (unknown). Suppose that a total of M packet flows are observed during the time interval of length T. The percentage of new IP addresses is

$$N = \frac{Y}{M} \tag{3.1.1}$$

where Y represents the number of packet flows with *distinct* unknown source IP addresses in the observation interval. Y is a Poisson process, which can be homogeneous or inhomogeneous. The next two sections present DDoS detection models for both cases.

3.1.1 Homogeneous Flow Arrival

Assume the counting process of legitimate and attack packet flows are homogeneous Poisson processes with rates ζ and λ, respectively. Applying the thinning property of Poisson processes (see Lemma 2) shows that the legitimate and attack packet flows with unknown source IP addresses are also homogeneous Poisson processes with rates $\zeta\varepsilon$ and $\lambda\varsigma$, respectively. Recall ε and ς are the fixed percentages of unknown source IP addresses for the legitimate and attack processes, respectively and $\varepsilon \simeq 10 - 20\%$ [102]. Detecting a DDoS attack can be formulated as the following binary hypothesis

$$\begin{cases} H_0 : Y \sim \text{Poisson with intensity } (\zeta\varepsilon)T \\ H_1 : Y \sim \text{Poisson with intensity } (\zeta\varepsilon + \lambda\varsigma)T \ . \end{cases} \tag{3.1.2}$$

The first hypothesis denotes the case where only legitimate traffic is observed. H_1 is the alternative hypothesis of having an attack underway. The attack and legitimate packet flows are assumed to be independent.

Let $L = \zeta\varepsilon + \lambda\varsigma$, the likelihood ratio corresponding to the binary hypothesis problem formulated in (3.1.2) is

$$\mathbb{L}(y) = \frac{\mathbb{P}\{Y = y|H_1\}}{\mathbb{P}\{Y = y|H_0\}} \ , \tag{3.1.3}$$

where

$$\mathbb{P}\{Y = y|H_1\} = e^{-LT}\frac{(LT)^y}{y!} \text{ and } \mathbb{P}\{Y = y|H_0\} = e^{-\zeta\varepsilon T}\frac{(\zeta\varepsilon T)^y}{y!} \ . \tag{3.1.4}$$

Thus the log-likelihood ratio is $\ln \mathbb{L}(y) = (-L + \zeta\varepsilon)T + y \ln\left(\frac{L}{\zeta\varepsilon}\right)$.

Define P_D and P_{FA} as the probabilities of detection (power test) and false alarm (FA), respectively. The Neyman-Pearson (NP) test [103] seeks a threshold η such that

$$P_{FA} = \mathbb{P}\{\mathbb{L}(y) > \eta|H_0\} = \varphi \tag{3.1.5}$$

The NP test is

$$y \underset{H_1}{\overset{H_0}{\lessgtr}} (\ln\eta + (L - \zeta\varepsilon)T) \times \frac{\ln\zeta\varepsilon}{\ln L} = \tilde{\eta} \ . \tag{3.1.6}$$

Given φ the fixed limit on P_{FA}, the NP threshold η must be selected such that

$$\mathbb{P}\{Y > \tilde{\eta}|H_0\} = 1 - \mathbb{P}\{Y \leq \tilde{\eta}|H_0\} = \varphi \tag{3.1.7}$$

Given H_0, Y is Poisson distributed with rate $\zeta\varepsilon T$. The Poisson cdf is not invertible, therefore the NP threshold, η, must be computed numerically. Alternatively, assume $\zeta\varepsilon T$, the mean of the Poisson distribution under H_0, is large. Then, the Poisson distribution can be approximated by a normal distribution with mean and variance both equal to $\zeta\varepsilon T$. Therefore,

$$\mathbb{P}\{Y > \tilde{\eta}|H_0\} = \varphi \Rightarrow \mathbb{P}\left\{\frac{Y - \zeta\varepsilon T}{\sqrt{\zeta\varepsilon T}} > \frac{\tilde{\eta} - \zeta\varepsilon T}{\sqrt{\zeta\varepsilon T}}\right\} = \varphi$$

And we reject H_0 if $Y \geq \zeta\varepsilon T + z_\varphi\sqrt{\zeta\varepsilon T}$, where z_φ is the normal quantile that corresponds to the limit on type I error probability (φ).

The Poisson process of packet flows may have a rate that varies with time. The case is illustrated in Figure 2.3 (see Section 2.1.3). The next section addresses this particular case.

3.1.2 Inhomogeneous Flow Arrival

Assume the counting process of legitimate and attack packet flows are inhomogeneous Poisson processes with intensity function $\zeta(t)$ and $\lambda(t)$, respectively. The thinning property of Poisson processes also applies when rates are functions of time. Let $\Lambda_0(t) = \zeta(t)\varepsilon$ and $\Lambda_1(t) = \zeta(t)\varepsilon + \lambda(t)\varsigma$, with $0 \leq t \leq T$, be the intensity functions under H_0 and H_1, respectively. When the Poisson intensities are known, the sample-function density can be used to formulate the hypothesis testing problem [88]. Let N be the number of events in the time interval of length T and denote $T_1 < T_2 < \ldots < T_N$ the random arrival times of packet flows. Then, the joint density of N and $\{T_i\}$ is

$$\mathbb{P}\{N = n, t_1, t_2, \ldots, t_n\} = \exp\left\{\int_0^T \ln\Lambda(t)dy(t) - \int_0^T \Lambda(t)dt\right\}, \tag{3.1.8}$$

where $\Lambda(t)$ is the intensity function of the Poisson process, $n = 0, 1, \ldots$ and $0 < t_1 < t_2 < \ldots < t_n < T$. The function $y(t) = \sum_{i=1}^{n} \mathbb{1}\{t \geq t_i\}$ corresponds to the counting function with n jumps occurring at times $t_1, t_2, \ldots, t_n$. The likelihood ratio can be written in terms of the sample-function

density as

$$\begin{aligned}
\mathbb{L}(Y) &= \frac{\mathbb{P}\{N=n, t_1, t_2, \ldots, t_n | H_1\}}{\mathbb{P}\{N=n, t_1, t_2, \ldots, t_n | H_0\}} \\
&= \exp\left\{-\int_0^T [\Lambda_1(t) - \Lambda_0(t)]\, dt\right\} \mathbb{1}_{\{N=0\}} \\
&+ \exp\left\{\sum_{i=1}^{N} \ln \frac{\Lambda_1(T_i)}{\Lambda_0(T_i)} - \int_0^T [\Lambda_1(t) - \Lambda_0(t)]\, dt\right\} \mathbb{1}_{\{N>0\}} \\
&= \exp\left\{\ln \frac{\Lambda_1(t)}{\Lambda_0(t)} dY(t) - \int_0^T [\Lambda_1(t) - \Lambda_0(t)]\, dt\right\}.
\end{aligned}$$

The Neyman-Pearson test can be written as

$$\sum_{i=1}^{N} \ln \frac{\Lambda_1(T_i)}{\Lambda_0(T_i)} \underset{H_1}{\overset{H_0}{\lessgtr}} \kappa = \ln \eta + \int_0^T [\Lambda_1(t) - \Lambda_0(t)]\, dt \tag{3.1.9}$$

Note that the intensity function of the inhomogeneous Poisson processes (packet flows process) can be approximated as piecewise constant function. So the procedure described for homogeneous Poisson processes in Section 3.1.1 can be applied as well. Detection error probabilities can be bounded using Chernoff inequalities.

3.1.3 Chernoff Bounds on Detection Error Probabilities

As mentioned in Section 2.6.2, error probabilities can be bounded using large deviation theory. Since Y represents the percentage of distinct new source IP addresses, Y has a finite variance, its mgf is well defined for $\theta > 0$ and Chernoff bounds can be computed. Analytical bounds on the probability errors can be derived as follows:

Proposition 1. *The probabilities of false alarm and false clear are bounded above as*

$$P_{FA} = \mathbb{P}\{Y > \tilde{\eta} | H_0\} \leq \frac{e^{-\zeta\varepsilon T}(e\zeta\varepsilon T)^{\tilde{\eta}}}{\tilde{\eta}^{\tilde{\eta}}}\ ,\ \tilde{\eta} > \zeta\varepsilon T \tag{3.1.10}$$

$$P_{FC} = \mathbb{P}\{Y < \tilde{\eta} | H_1\} \leq \frac{e^{-LT}(eLT)^{\tilde{\eta}}}{\tilde{\eta}^{\tilde{\eta}}}\ ,\ \tilde{\eta} > LT \tag{3.1.11}$$

Proof. Under H_0, Y follows a Poisson distribution with parameter $\zeta\varepsilon T$. Let $\mathcal{M}_Y(\theta)$ denote the moment generating function of Y.

$$\mathcal{M}_Y(s) = e^{\zeta\varepsilon T(e^s - 1)}$$

The exponential upper bound is given as [88]

$$\mathbb{P}\{Y > y\} \leq e^{-sy - \ln \mathcal{M}_Y(s)},\ s > 0, y \in \mathbb{R}.$$

Let $g_y(s) \triangleq sy - \zeta\varepsilon T\left(e^s - 1\right)$.
Solving for the optimal $\hat{s}$ requires finding the maximum of $g_y(s)$. The derivative is $g'_y(s) = y - \zeta\varepsilon e^s$. Set $g'_y(0)$ and deduce that $\hat{s} = \ln\left(\frac{y}{\zeta\varepsilon T}\right)$.
From (2.6.7), deduce that

$$\begin{aligned}\mathbb{P}\left\{Y > y\right\} &\leq \exp\left\{-\ln\left(\frac{y}{\zeta\varepsilon}\right)y + \zeta\varepsilon T\left(e^{\ln\left(\frac{y}{\zeta\varepsilon}\right)} - 1\right)\right\} \\ &= \frac{e^{-\zeta\varepsilon T}(e\zeta\varepsilon T)^y}{y^y}\end{aligned}$$

Equations (3.1.10) follow by the Neyman-Pearson test defined in (3.1.6). □

This section has presented a detection system which monitors the distinct number of unknown IP addresses. Subsequent sections in this chapter introduce a detection system that counts the total number of data packets with unknown source IP address, regardless of duplicates. The second statistic does not depend on the number of spoofed source IP addresses used by the attacker, which is an improvement over the former statistic.

3.2 Marked Poisson with Pareto Cluster Size

The DDoS detection model presented in this section monitors the percentage of unknown source IP addresses by counting individual data packets rather than flows. Network traffic is modeled as the marked Poisson process (T_j, K_j, B_j) where B_j, K_j and T_j are mutually independent. The thinning property of Poisson processes applied to the Bernoulli mark reduces the network model (T_j, K_j, B_j) to $\left(\widetilde{T}_j, K_j\right)$, where $\widetilde{T}_j$ now represents the random arrival times of packet flow j with unknown source IP address. As in Section 3.1.1, consider that the counting process of legitimate $(U(t))$ and attack $A(t)$ packet flows are homogeneous Poisson processes with rates $\zeta\varepsilon$ and $\lambda\varsigma$, respectively.

Assume flow cluster sizes, $\{K_j\}$, are iid Pareto random variables. The sum of n iid Pareto random is dominated by their maximum, i.e.,

$$\lim_{k\to\infty} \frac{\mathbb{P}\left\{K_1 + \ldots + K_n > k\right\}}{\mathbb{P}\left\{\max\left\{K_1 + \ldots + K_n\right\} > x\right\}} \approx 1 \tag{3.2.1}$$

By definition, the distribution of cluster sizes (see Section 2.1) has a heavy tail. Equation (3.2.1) indicates that the sum of cluster sizes will be dominated by large rather than small cluster sizes. Therefore, as a first step, we assume that cluster sizes are iid Pareto random variables. The case of mixture distribution will be addressed later.

The DDoS detection problem is formulated as a binary hypothesis problem (see 3.1.2). However, the total number of data packets with unknown source IP addresses, i.e. $Y = \sum K_j$, is now a compound Poisson random variable [114]. The objective of the detection model is to derive an optimal

likelihood-ratio test for differentiating two compound Poisson processes, given the percentage of unknown source IP addresses observed in a fixed time interval of length T. A decision, attack or legitimate, has to be made at the end of the observation period.

Let $K_j \sim \text{Par}(\alpha_0, c_0)$ and $K_j \sim \text{Par}(\alpha, c)$, given H_0 and H_1, respectively. With proper normalization, Lemma 4 shows that the scale parameters of the two Pareto distribution can be set to $c_0 = c = 1$. Also, set $\mathcal{N}(T) = U(T) + A(T)$, where $U(t)$ and $A(t)$ denote the legitimate and attack processes with unknown source IP addresses, respectively. The conditional pdfs defining the likelihood ratio in (4.1) become

$$\begin{aligned}
\mathbb{P}\{Y = y|H_1\} &= \mathbb{P}\left\{\sum_{j=1}^{\mathcal{N}(T)} K_j = y\right\},\ \mathcal{N}(T) \sim \text{Pois}(LT),\ K_j \sim \text{Par}(\alpha, 1) \\
\mathbb{P}\{Y = y|H_0\} &= \mathbb{P}\left\{\sum_{j=1}^{U(T)} K_j = y\right\},\ U(T) \sim \text{Pois}(\zeta\varepsilon T),\ K_j \sim \text{Par}(\alpha_0, 1)
\end{aligned}$$

The next step in the detection procedure is to evaluate the pdfs $f_Y(y)$ above and deduce the Neyman-Pearson threshold. The random variable Y is a random sum of iid Pareto random variables.

The limited literature expresses how challenging it is to compute the sum of n iid Pareto random variables. We present a brief overview of different proposed methods. Recall the pdf of a Pareto distribution is defined as

$$f(x) = \frac{\alpha c^\alpha}{x^{\alpha+1}}, \tag{3.2.2}$$

where $x \geq c > 0$.

Assuming that random variables $\{X_i\}$ are iid Pareto random variables with tail index and scale parameters $\alpha = c = 1$, Hagstroem [52] computes the closed form for the convolution of $n = 2$ and $n = 3$ Pareto random variables. Additional analytical expression of the sum of n iid Pareto random variables include [12] and [16]. The work of Blum [12] shows that

$$f_n(x) = \frac{-1}{\pi} \sum_{j=1}^{n} \binom{n}{j} (-\Gamma(1-\alpha))^j \sin(\pi\alpha j) \sum_{m=0}^{\infty} C_{n-j,m} \frac{\Gamma(m+\alpha j+1)}{x^{m+\alpha j+1}}\ ,$$

where $C_{k,m}$ is the mth coefficient in the series expansion of the kth power of the confluent hypergeometric function,

$$\sum_{m=0}^{\infty} C_{k,m} t^m = ({}_1F_1(-\alpha, 1-\alpha, t))^k.$$

Furthermore, Feller [36] proves that as $x \to \infty$

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i > x\right\} \sim n\left(1 + \frac{x}{c}\right)^{-\alpha} L(x)\ , \tag{3.2.3}$$

where $L(x)$ is a slowly varying function, which by definition means for a given constant $t > 0$, $\frac{L(tx)}{L(x)} \to 1$ as $x \to \infty$. Roehner and Winiwarter [108] also introduce the stable distribution as the asymptotic density of the normalized sum of n iid Pareto random variables as $n \to \infty$.

Each method presented so far considers that n, the number of summands, is deterministic. Assuming the number of summands is random, Ramsay [106] addresses the case of the compound Poisson-Pareto distribution. Let $Y = \sum_{i=1}^{N} X_i$, where $\{X_i\}$ are iid Pareto and $N \sim \text{Pois}(\lambda)$. The expression of the right tail distribution of Y is given as

$$\bar{F}_Y(t) = 1 - \mathbb{P}\{Y \leq x\} = \frac{e^{-\lambda}}{\pi} \int_0^\infty \frac{e^{-tz/\delta}}{z} e^{\lambda R(z,\alpha)} \sin(\lambda I(z,\alpha)) dz, \tag{3.2.4}$$

where

$$R(z,\gamma) = \begin{cases} -\gamma e^{-} z Ei_{\alpha+1}(z) & \text{if } \alpha = 1, 2, \ldots \\ M(1; 1-\alpha; -z) - I(z,\alpha)\cot(\pi\alpha) & \text{if } \alpha \neq 1, 2, \ldots \end{cases}$$

$$M(1; 1-\alpha; -z) = 1 + \sum_{r=0}^{\infty} \frac{z_r}{(\alpha-1)\ldots(\alpha-r)},$$

and $Ei_m(z)$ is defined as

$$Ei_m(z) = \frac{z^{m-1}}{(m-1)!} \left[\gamma + \ln z - \sum_{r=1}^{m-1} \frac{1}{r}\right] + \sum_{\substack{r=0 \\ r \neq m-1}}^{\infty} \frac{z^r}{(r-m+1)r!}, \tag{3.2.5}$$

where γ is the Euler constant.

The methods presented above are very involved and computationally arduous. Next, we present simpler and efficient methods for approximating the pdf of the compound Poisson- Pareto random variable Y. For detection purposes, we are particularly interested in large deviations. This allows some simplifications over the explicit formula in (3.2.4). Tail probabilities can be obtained by direct approximation or by numerical integration of the pdf. Recall the Poisson process has different intensity depending on the hypothesis H_0 or H_1. We present the approximation procedures for the latter, the former case is similar. Note that the approximation methods depend on the value of the Pareto tail index (α).

3.2.1 Effect of Randomness due to Poisson

Compound random variables do not have a closed-form pdf and should be evaluated through their characteristic functions. In this section, flow cluster sizes are assumed to be (pure) Pareto distributed, thus have characteristic function $\psi_K(\theta) = \alpha(-i\,c\,\theta)^\alpha \Gamma(-\alpha, -ic\theta)$. Lemma 6 determines the characteristic function of the corresponding compound random variable, $Y = \sum K_j$.

Lemma 6. *The characteristic function of Y is*

$$\psi_Y(\theta) = \exp\left\{\alpha LT(-i\,c\,\theta)^\alpha \Gamma(-\alpha, -ic\,\theta) - LT\right\} \tag{3.2.6}$$

Proof. Cluster size K_j's are iid then by conditioning on $\mathcal{N}(T)$,

$$\begin{aligned}
\psi_Y(\theta) = \mathbb{E}\left[e^{i\theta Y}\right] &= \sum_{n=0}^{\infty} \mathbb{E}\left[\exp\left\{i\theta \sum_{j=1}^{n} K_j\right\}\right] \mathbb{P}\left\{U(t) + A(t) = n\right\} \\
&= \sum_{n=0}^{\infty} \left(\mathbb{E}\left[\exp\left\{i\theta K_j\right\}\right]\right)^n \frac{e^{-LT}\,(LT)^n}{n!} \\
&= e^{-LT} \sum_{n=0}^{\infty} \left\{\psi_K(\theta)\right\}^n \frac{(LT)^n}{n!} \\
&= \exp\left\{LT\left[\psi_K(\theta) - 1\right]\right\}
\end{aligned}$$

and the result follows from replacing $\psi_K(\theta)$ by its value. □

The compound random variable $Y = \sum_{j=1}^{\mathcal{N}(T)} K_i$ has two sources of randomness, the random cluster sizes and the random number of summands. First, we study the effect of the randomness due to the Poisson component, particularly when the rate of the Poisson process is very large.

Recall under hypothesis $H_1, \mathcal{N}(T) \sim \text{Pois}(LT)$. For LT sufficiently large, the Poisson distribution can be approximated by a normal distribution with mean and variance both equal to LT. That is, $\mathcal{N}(T) \approx LT + \sqrt{LT}\,\mathfrak{N}$, where $\mathfrak{N}$ is a standard normal random variable. Evaluating the "effect of the Poisson" basically consists of comparing the tails of $\sum_{j=1}^{LT} K_i$ and $\sum_{j=1}^{LT+\sqrt{LT}\,\mathfrak{N}} K_i$, for $LT \to \infty$.

Theorem 5. *The effect due to the Poisson component is negligible. The asymptotic tails of $\sum_{j=1}^{LT} K_i$ and $\sum_{j=1}^{LT+\sqrt{LT}\,\mathfrak{N}} K_i$ have the same behavior up to $\frac{1}{(LT)^{1/\alpha}}$ when $LT \to \infty$.*

The proof of Theorem 5 requires the small θ expansion of the Pareto characteristic function, $\psi_K(\theta)$, given in Lemma 7 [108].

Lemma 7. *Define the normalized cluster size as*

$$\hat{K}_i = \frac{K_i - \mu_K}{(LT)^{1/\alpha}}\ . \tag{3.2.7}$$

Near zero, the characteristic function of the normalized cluster size $\hat{K}_j$ expands as

$$\begin{aligned}\psi_{\hat{K}}(\theta) = \mathbb{E}\left[\exp\left\{i\theta\hat{K}_j\right\}\right] &= \left[1 - \frac{i\theta\mu_K}{(LT)^{1/\alpha}} - \frac{1}{2}\left(\frac{\theta\,\mu_K}{(LT)^{1/\alpha}}\right)^2 + \ldots\right] \\ &\times \left[1 + \alpha\frac{(-i\theta)^\alpha}{LT}\Gamma(-\alpha) + \frac{\alpha}{\alpha-1}\frac{(i\theta)}{(LT)^{1/\alpha}} + \ldots\right]\end{aligned} \tag{3.2.8}$$

Proof. Characteristic functions have the property that $\psi_{aX+b}(\theta) = e^{i\theta b}\psi_X(a\theta)$. Therefore,

$$\psi_{\widetilde{Y}}(\theta) = \exp\left\{-i\theta\mu_K\left(LT\right)^{-\frac{1}{\alpha}}\right\}\psi_K\left(\frac{\theta}{(LT)^{1/\alpha}}\right) \tag{3.2.9}$$

The next step is to expand the incomplete Gamma and exponential function around zero. $\Gamma(a,z)$ admits two series expansion, which both converge unless a is a negative integer [97, 130].

$$\Gamma(a,z) = \Gamma(a) - \sum_{n=0}^{\infty}\frac{(-1)^n z^{n+a}}{(a+n)n!} \tag{3.2.10}$$

Further details about expansions of the incomplete Gamma function can be found in [2, 10, 11, 42, 43, 97, 117, 130]. Let $a = \alpha$ and $z = -i\theta$. For θ small,

$$\begin{aligned}\psi_K(\theta) &= \alpha(-i\theta)^\alpha\Gamma(-\alpha,-i\theta) \\ &\simeq \alpha(-i\theta)^\alpha\left[\Gamma(-\alpha) - \frac{(-i\theta)^{-\alpha}}{-\alpha}\left(1 - \frac{-\alpha(-i\theta)}{-\alpha+1} + \frac{-\alpha(-i\theta)^2}{2(-\alpha+2)} - \mathcal{O}((-i\theta)^3)\right)\right] \\ &\simeq 1 + \alpha(-i\theta)^\alpha\Gamma(-\alpha) - \frac{\alpha}{\alpha-1}(-i\theta) + \ldots\end{aligned}$$

Furthermore, the Taylor expansion of the exponential function is $e^x \approx 1 + x + \frac{1}{2}x^2 + \ldots$. Therefore,

$$\exp\left\{-i\theta\mu_K\left(LT\right)^{-\frac{1}{\alpha}}\right\} \approx 1 - \frac{i\theta\mu_K}{(LT)^{1/\alpha}} - \frac{1}{2}\left(\frac{\theta\,\mu_K}{(LT)^{1/\alpha}}\right)^2 + \ldots$$

Equation (3.2.8) then follows. □

Proof of Theorem 5. Consider the normalized number of data packets with unknown source IP address in (3.2.19).

$$\widetilde{Y} = \sum_{j=1}^{\mathcal{N}(T)}\hat{K}_i = \sum_{j=1}^{\mathcal{N}(T)}\frac{K_i - \mu_K}{(LT)^{1/\alpha}}\,.$$

Then

$$\begin{aligned}\mathbb{E}\left[e^{i\theta\widetilde{Y}}|\mathcal{N}(T)\right] &= \left(\mathbb{E}\left[\exp\left\{i\theta\hat{K}_j\right\}\right]\right)^{\mathcal{N}(T)} \\ &= \exp\left\{\mathcal{N}(T)\ln\mathbb{E}\left[\exp\left\{i\theta\hat{K}_j\right\}\right]\right\}\end{aligned} \tag{3.2.11}$$

The conditional characteristic function in (3.2.11) can be rewritten as

$$\mathbb{E}\left[e^{i\theta\widetilde{Y}}|\mathcal{N}(T)\right] \simeq \exp\left\{\left(LT+\sqrt{LT}\,\mathfrak{N}\right)\ln\mathbb{E}\left[e^{i\theta\hat{K}_j}\right]\right\}$$

The characteristic function of $\widetilde{Y}$ is

$$\psi_{\widetilde{Y}}(\theta)=\mathbb{E}\left[\mathbb{E}\left[e^{i\theta\widetilde{Y}}|\mathcal{N}(T)\right]\right] \simeq \exp\left\{LT\ln\mathbb{E}\left[e^{i\theta\hat{K}_j}\right]\right\}\times\mathbb{E}\left[\exp\left\{\sqrt{LT}\,\mathfrak{N}\ln\mathbb{E}\left[e^{i\theta\hat{K}_j}\right]\right\}\right]$$

For any θ, $\mathcal{M}_{\mathfrak{N}}(\theta)=\frac{1}{2}\theta^2$. Thus,

$$\psi_{\widetilde{Y}}(\theta) \simeq \exp\left\{LT\ln\mathbb{E}\left[e^{i\theta\hat{K}_j}\right]\right\}\times\exp\left\{\frac{LT}{2}\left(\ln\mathbb{E}\left[e^{i\theta\hat{K}_j}\right]\right)^2\right\}$$

Using the small θ expansion of the characteristic function in Lemma 7, we deduce that

$$\begin{aligned}\ln\psi_{\hat{K}}(\theta) &= \ln\left(1-\frac{i\theta\mu_K}{(LT)^{1/\alpha}}-\frac{1}{2}\left(\frac{\theta}{(LT)^{1/\alpha}}\right)^2+\ldots\right) \\ &+ \ln\left(1+\alpha\frac{(-i\theta)^\alpha}{LT}\Gamma(-\alpha)+\frac{\alpha}{\alpha-1}\frac{(i\theta)}{(LT)^{1/\alpha}}+\ldots\right) \\ &\approx -\frac{i\theta\mu_K}{(LT)^{1/\alpha}}-\frac{1}{2}\left(\frac{\theta}{(LT)^{1/\alpha}}\right)^2+\ldots+\alpha\frac{(-i\theta)^\alpha}{LT}\Gamma(-\alpha)+\frac{\alpha}{\alpha-1}\frac{(i\theta)}{(LT)^{1/\alpha}}+\ldots\end{aligned} \tag{3.2.12}$$

The last expression follows from letting $LT\to\infty$ and applying the Taylor expansion of the natural

logarithm function, i.e., $\ln(1+x) \simeq x$, as $x \to 0$.

$$\begin{aligned}
\psi_{\widetilde{Y}}(\theta) &\simeq \exp\left\{\left[\alpha\frac{(-i\theta)^\alpha}{LT}\Gamma(-\alpha) - \frac{1}{2}\left(\frac{\theta}{(LT)^{1/\alpha}}\right)^2 + \ldots\right] LT\right\} \\
&\times \exp\left\{\frac{LT}{2}\left(\alpha\frac{(-i\theta)^\alpha}{LT}\Gamma(-\alpha) - \frac{1}{2}\left(\frac{\theta}{(LT)^{1/\alpha}}\right)^2 + \ldots\right)^2\right\} \\
&\simeq \exp\left\{\alpha(-i\theta)^\alpha\Gamma(-\alpha) - \frac{1}{2}\frac{\theta^2}{(LT)^{2/\alpha-1}} + \ldots\right\} \\
&\times \exp\left\{\left(\alpha\frac{(-i\theta)^\alpha}{\sqrt{2}\,(LT)^{1/2}}\Gamma(-\alpha) - \frac{1}{2\sqrt{2}}\frac{\theta^2}{(LT)^{2/\alpha-1/2}} + \ldots\right)^2\right\} \\
&\simeq \exp\left\{\alpha(-i\theta)^\alpha\Gamma(-\alpha) - \frac{1}{2}\frac{\theta^2}{(LT)^{2/\alpha-1}}\right\} \qquad (3.2.13)
\end{aligned}$$

□

Note that, f_Y, the pdf of Y can be approximated by applying the inverse Fourier transform method to the characteristic function in (3.2.13). The latter is equivalent to the characteristic function of a stable distribution. Theorem 2 shows that the limiting distribution of the sum of n iid symmetric Pareto random variable is a stable distribution. The next section extend results from [108, 110] to show that the sum of a random number of iid (non-symmetric) Pareto random variables can be approximated with a stable distribution as well.

3.2.2 Stable Distribution as a Limiting Distribution

Let $\mathcal{N}(T)$ be a Poisson random variable, the next two theorems show that the asymptotic distribution of the normalized sum of $\mathcal{N}(T)$ iid Pareto with tail index $0 < \alpha < 2$ is a stable distribution. The cases $0 < \alpha < 1$ and $0 < \alpha < 2$ must be treated separately.

If $0 < \alpha < 1$, the first moment is infinite. In that case, define the normalized cluster size as

$$\widetilde{K}_j = \frac{K_j}{(LT)^{1/\alpha}}\,, \qquad (3.2.14)$$

where K_j is a Pareto random variable, $K_j \sim \text{Par}(\alpha, 1)$. If $LT \to \infty$, the distribution of the normalized number of unknown source IP addresses, i.e. the compound random variable $\widetilde{Y} = \sum_{j=1}^{\mathcal{N}(T)} \widetilde{K}_j$, can be approximated by a stable distribution.

Theorem 6. *If* $0 < \alpha < 1$ *and* $LT \to \infty$ *then*

$$\widetilde{Y} \sim \mathcal{S}_\alpha \left(\left[\Gamma(1-\alpha) \cos\left(\frac{\pi\alpha}{2}\right) \right]^{1/\alpha}, 1, 0 \right) \tag{3.2.15}$$

and

$$\psi_{\widetilde{Y}}(\theta) = \exp\left\{ -|\theta|^\alpha \Gamma(1-\alpha) \cos\left(\frac{\pi\alpha}{2}\right) \left[1 - i \ (sign\,\theta) \tan\left(\frac{\pi\alpha}{2}\right) \right] \right\} \tag{3.2.16}$$

Proof. Lemma 4 can be adapted to show that $\widetilde{K}_j$ is also Pareto distributed as $\widetilde{K}_j \sim \text{Par}\left(\alpha, \frac{1}{(LT)^{1/\alpha}}\right)$. Furthermore, Lemma 6 shows that $\psi_Y(\theta) = \exp\left\{ LT \left[\psi_K\left(\frac{\theta}{(LT)^{1/\alpha}}\right) - 1 \right] \right\}$. Then

$$\psi_{\widetilde{Y}}(\theta) = \exp\left\{ LT \left[\int_{\frac{1}{(LT)^{1/\alpha}}}^{\infty} e^{i\theta x} \frac{\alpha \left(\frac{1}{(LT)^{1/\alpha}}\right)^\alpha}{x^{1+\alpha}} \, dx - 1 \right] \right\}$$

$$\psi_{\widetilde{Y}}(\theta) = \exp\left\{ \int_{\frac{1}{(LT)^{1/\alpha}}}^{\infty} \frac{\alpha}{x^{1+\alpha}} \left(e^{i\theta x} - 1 \right) \, dx \right\}$$

As $LT \to \infty$, $\frac{1}{(LT)^{1/\alpha}} \to 0$ and

$$\lim_{LT\to\infty} \psi_{\widetilde{Y}}(\theta) \quad = \quad \exp\left\{ \alpha \int_0^{\infty} \frac{\left(e^{i\theta x} - 1\right)}{x^{1+\alpha}} \, dx \right\}$$

Feller [36] proves that for $0 < \alpha < 1$ the integral converges and

$$\int_0^{\infty} \frac{\left(e^{i\theta x} - 1\right)}{x^{1+\alpha}} \, dx = -\frac{1}{\alpha} \Gamma(1-\alpha) \theta^\alpha e^{-i\alpha\pi/2}$$

Therefore,

$$\begin{aligned} \lim_{LT\to\infty} \psi_{\widetilde{Y}}(\theta) &= \exp\left\{ -|\theta|^\alpha \Gamma(1-\alpha) \left[\cos\left(\frac{\pi\alpha}{2}\right) - i \ (\text{sign}\,\theta) \sin\left(\frac{\pi\alpha}{2}\right) \right] \right\} \\ &= \exp\left\{ -|\theta|^\alpha \Gamma(1-\alpha) \cos\left(\frac{\pi\alpha}{2}\right) \left[1 - i \ (\text{sign}\,\theta) \tan\left(\frac{\pi\alpha}{2}\right) \right] \right\} , \end{aligned}$$

which indicates

$$\widetilde{Y} \sim \mathcal{S}_\alpha \left(\left[\Gamma(1-\alpha) \cos\left(\frac{\pi\alpha}{2}\right) \right]^{1/\alpha}, 1, 0 \right) \tag{3.2.17}$$

□

Let $K_j \sim \text{Par}(\alpha, 1)$. If $1 < \alpha < 2$, for any flow j the expected value of the cluster size is

$$\mu_K = \mathbb{E}[K_j] = \frac{\alpha}{\alpha - 1} \,. \tag{3.2.18}$$

Define the normalized number of data packets with unknown source IP address as

$$\widetilde{Y} = \frac{1}{(LT)^{1/\alpha}} \left(\sum_{j=1}^{\mathcal{N}(T)} K_j - LT\mu_K \right) \,. \tag{3.2.19}$$

Given H_1, the expected number of packet flows with unknown source IP addresses in the time inverval of length T is $\mathbb{E}[\mathcal{N}(T) = U(T) + A(T)] = LT$. Then (3.2.19) follows by Wald's identity, i.e., $\mathbb{E}[Y] = \mathbb{E}[\mathcal{N}(T)]\mathbb{E}[K_j] = LT\mathbb{E}[K_j]$. Furthermore, Lemma 7 shows that if $\theta \to 0$, then

$$\psi_K(\theta) = 1 + \alpha(-i\theta)^\alpha \Gamma(-\alpha) - \frac{\alpha}{\alpha - 1}(-i\theta) + \dots \tag{3.2.20}$$

Using (3.2.20), Theorem 7 gives the limiting distribution of $\widetilde{Y}$.

Theorem 7. *Consider the normalized number of data packets with unknown source IP address in* (3.2.19). *If* $LT \to \infty$ *then*

$$\psi_{\widetilde{Y}}(\theta) = \exp\left\{\alpha(-i\theta)^\alpha \Gamma(-\alpha)\right\}. \tag{3.2.21}$$

Formula (3.2.21) *corresponds to the characteristic function of a stable distribution with tail index* $1 < \alpha < 2$, *i.e.,*

$$\psi_{\widetilde{Y}}(\theta) \to \exp\left\{-|\theta|^\alpha \left[1 + i\beta(sign\,\theta)(\tan\frac{\pi\alpha}{2})\left(|\theta|^{1-\alpha} - 1\right)\right]\right\}, \tag{3.2.22}$$

with parameters $\gamma^\alpha = \alpha\Gamma(-\alpha)\cos\frac{\alpha\pi}{2}$, $\beta = 1$, *and* $\delta = 0$.

Proof. Recall from the proof of Lemma 7,

$$\psi_{\widetilde{Y}}(\theta) = \exp\left\{-i\theta\mu_K (LT)^{1-\frac{1}{\alpha}}\right\} \exp\left\{LT\left[\psi_K\left(\frac{\theta}{(LT)^{1/\alpha}}\right) - 1\right]\right\}$$

If $LT \to \infty$ then $\frac{\theta}{(LT)^{1/\alpha}} \to 0$ and we can replace ψ_K by its small θ expansion given in (3.2.20). The approximated characteristic function of $\widetilde{Y}$ can be written as

$$\begin{aligned}
\psi_{\widetilde{Y}}(\theta) &= \exp\left\{-i\theta\mu_K (LT)^{1-\frac{1}{\alpha}}\right\} \exp\left\{LT\left[1 + \alpha\frac{(-i\theta)^\alpha}{LT}\Gamma(-\alpha) + \frac{\alpha}{\alpha-1}\frac{(i\theta)}{(LT)^{\frac{1}{\alpha}}} + \dots - 1\right]\right\} \\
&= \exp\left\{-i\theta\mu_K (LT)^{1-\frac{1}{\alpha}}\right\} \exp\left\{\alpha(-i\theta)^\alpha\Gamma(-\alpha) + \frac{\alpha}{\alpha-1}(i\theta)(LT)^{1-\frac{1}{\alpha}} + \dots\right\},
\end{aligned}$$

where μ_K is given by (4.4.1). Then (3.2.21) follows. Feller [36] shows that $(-i\theta)^\alpha = \theta^\alpha e^{-i\frac{\alpha\pi}{2}}$, which

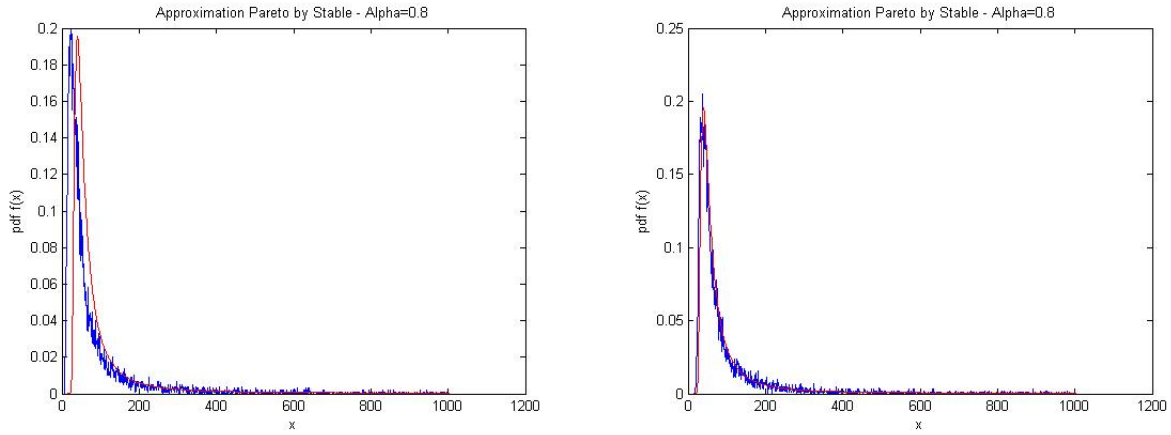


Figure 3.1: Convergence Sum Paretos to Stable

leads to (3.2.22). □

Remark 1. *With the CLT, the convergence of the sum of n iid random variables to the Gaussian distribution is generally achieved for* $n > 30$. *However with heavy tails the convergence is much slower, in some case, n must be greater than* 10000 *[134].*

Figure 3.1 illustrates Remark 1 with Poisson rate $LT = 30, 20000$, respectively. As LT increases, the stable distribution becomes a better approximation.

If the Pareto tail index satisfies $\alpha > 2$, the standard CLT with the normal distribution applies. Note that the normal distribution itself is a stable distribution with $\alpha = 2$. The next section presents the procedure to approximate the distribution function of the stable distribution and the compound Pareto when the tail index $\alpha > 2$.

3.2.3 Density Estimation and Saddlepoint Approximations

Stable distributions with tail index satisfying $0 < \alpha < 2$ do not have a closed-form density function. The only exceptions are the Levy, Cauchy and Normal distributions, which have tail index $\alpha = \frac{1}{2}$, $\alpha = 1$ and $\alpha = 2$, respectively. Approximation methods to estimate the pdf of stable distributions include numerical integration [93, 94], Fast Fourier transform (FFT) [85], and Laguerre quadrature [9]. The pdf can be also numerically approximated from the characteristic function using the FFT approach described in Section 2.5 [57]. This technique is developed based on characteristic functions and does not make any assumption about the packet flow rate or the length of the observation window. Figure 3.2 illustrates the stable pdf approximated by FFT.

As an alternative to these techniques, we propose the method of saddlepoint approximation to compute $f_{\widetilde{Y}}(y)$. Daniels [25] introduces the saddlepoint method to approximate the distribution

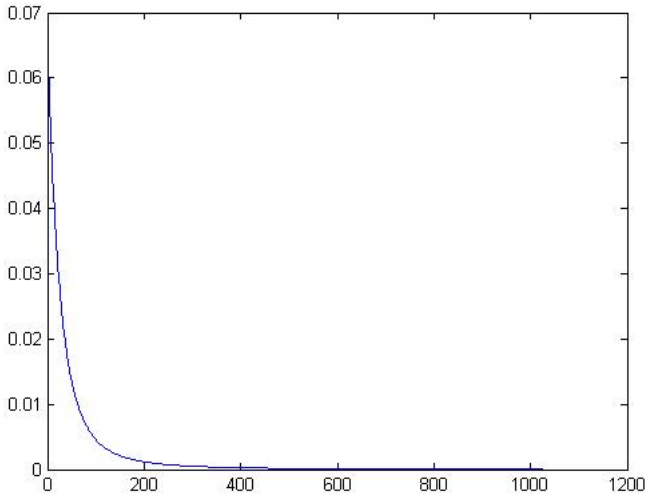


Figure 3.2: FFT Approximation of the Stable Pdf with $\alpha = 1.3$

of the sample mean of n iid random variables and proves it to be very effective even for small n. The technique can be adapted to deduce the distribution of one random variable [49] in this case the stable distribution. Lemma 8 provides the Laplace transform $\mathcal{M}_{\widetilde{Y}}(-\theta)$, where $\widetilde{Y}$ represents the normalized number of data packets with unknown source IP addresses [110].

Lemma 8. *The Laplace transform of* $\widetilde{Y}$ *is*

$$\mathcal{M}_{\widetilde{Y}}(-\theta) = \exp\left\{-\theta^{\alpha}\Gamma(1-\alpha)\right\} \ , \theta > 0\,. \tag{3.2.23}$$

Proof. Using results from the proof of Theorem 6,

$$\mathbb{E}\left[\exp(-\theta\widetilde{Y})\right] = \exp\left\{\int_{\frac{1}{(LT)^{1/\alpha}}}^{\infty} \frac{\alpha}{x^{1+\alpha}}\left(e^{-\theta x}-1\right)\ dx\right\} \tag{3.2.24}$$

As $LT \to \infty$, $\frac{1}{(LT)^{1/\alpha}} \to 0$ and

$$\begin{aligned}\mathbb{E}\left[\exp(-\theta\widetilde{Y})\right] &= \exp\left\{\int_{0}^{\infty} \frac{\alpha}{x^{1+\alpha}}\left(e^{-\theta x}-1\right)\ dx\right\} \\ &= \exp\left\{\theta^{\alpha}\int_{0}^{\infty} \alpha x^{-(1+\alpha)}\left(e^{-\theta x}-1\right)\ dx\right\}\end{aligned}$$

The last equality follows by performing the change of variable $t = \theta x$.

Integrating once by parts yields

$$\mathbb{E}\left[\exp(-\theta\widetilde{Y})\right] = \exp\left\{-\theta^{\alpha}\int_0^{\infty}\alpha x^{-\alpha}e^{-\theta x}\,dx\right\}$$

□

Proposition 2. *The pdf $f_{\widetilde{Y}}(y)$ can be approximated via saddlepoint approximation as follows*

$$f_{\widetilde{Y}}(y) \approx \frac{\exp\left\{-\Gamma(1-\alpha)(-\hat{s})^{\alpha} - y\hat{s}\right\}}{\sqrt{2\pi\alpha(\alpha-1)(-\hat{s})^{\alpha-2}\Gamma(1-\alpha)}}\,, \tag{3.2.25}$$

where

$$\hat{s} = \left(\frac{y}{\alpha\Gamma(1-\alpha)(-1)^{\alpha-1}}\right)^{\frac{1}{\alpha-1}}\,. \tag{3.2.26}$$

Proof. Using results from [25] and the inverse Fourier transform theorem [30], write the pdf $f_{\widetilde{Y}}(y)$ as

$$\begin{aligned} f_{\widetilde{Y}}(y) &= \frac{1}{2\pi i}\int_{\tau-i\infty}^{\tau'+i\infty} e^{-\theta x}\mathcal{M}_{\widetilde{Y}}(i\theta)d\theta \\ &= \frac{1}{2\pi i}\int_{\tau-i\infty}^{\tau'+i\infty} e^{\mathcal{K}_{\widetilde{Y}}(\theta)-\theta y}d\theta, \end{aligned}$$

where $\mathcal{K}_{\widetilde{Y}}(\theta) = \ln\mathcal{M}_{\widetilde{Y}}(\theta) = -(-\theta)^{\alpha}\Gamma(1-\alpha)$ be the cgf of $\widetilde{Y}$. From Lemma 8, deduce that $\mathcal{M}_{\widetilde{Y}}(\theta) = \exp\left\{-(-\theta)^{\alpha}\Gamma(1-\alpha)\right\}$. Given that $0<\alpha<1$, all moments of $\widetilde{Y}$ are infinite and the mgf is defined only for $\theta<0$.

Applying Taylor expansion around the saddlepoint $\hat{s}$ of the function $\mathcal{K}_{\widetilde{Y}}(\theta)-\theta y$ shows that

$$\mathcal{K}_{\widetilde{Y}}(\theta) - \theta y \simeq \mathcal{K}_{\widetilde{Y}}(\hat{s}) - \hat{s}y + \frac{(\theta-\hat{s})^2}{2}\mathcal{K}_{\widetilde{Y}}(\hat{s}) + \mathcal{O}(\theta^3) \tag{3.2.27}$$

The saddlepoint equation to solve is

$$\mathcal{K}'_{\widetilde{Y}}(\hat{s}) = y \triangleq \alpha\Gamma(1-\alpha)(-\hat{s})^{\alpha-1} = y$$

This yields the saddlepoint (3.2.26).

Next, replace the exponent with its Taylor expansion in the integral above. Then integrate with respect to θ.

The second derivative of the cgf is $\mathcal{K}''_{\widetilde{Y}}(\theta) = -\alpha(\alpha-1)(-\theta)^{\alpha-2}\Gamma(1-\alpha)$. Therefore, the first order

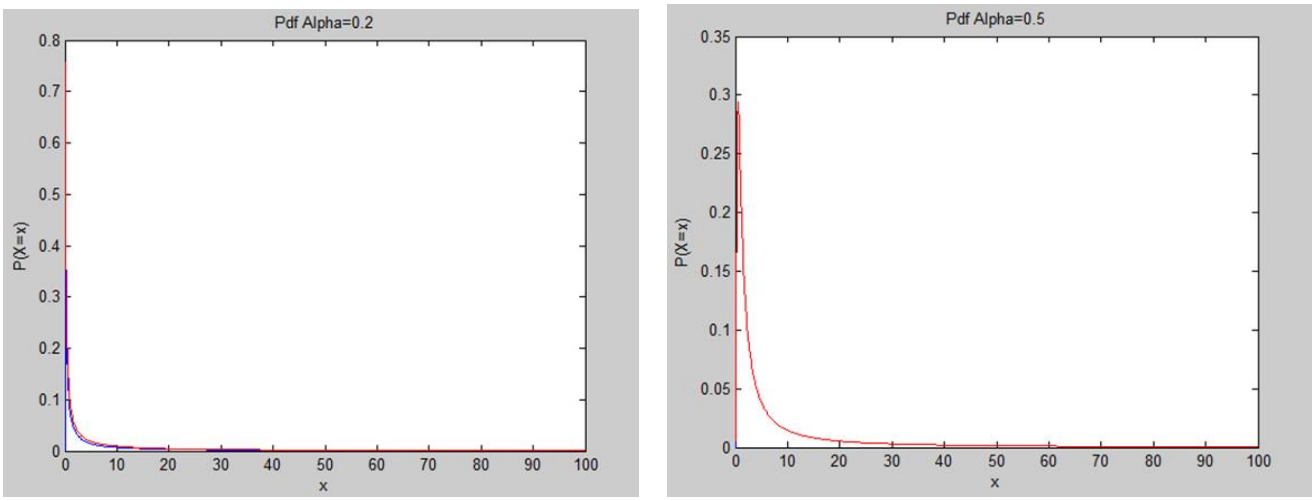


Figure 3.3: Saddlepoint Approximation for Stable-1

saddlepoint approximation is

$$f_{\widetilde{Y}}(y) \approx \frac{1}{\sqrt{2\pi \mathcal{K}''_{\widetilde{Y}}(\hat{s})}} \exp\left\{\mathcal{K}_{\widetilde{Y}}(\hat{s}) - y\hat{s}\right\} , \tag{3.2.28}$$

which gives the desired result (3.2.25). □

Remark 2. *The saddlepoint (3.2.26) is always negative and therefore well defined on the domain of the mgf. The same approximation method applies for the case of the stable distribution with tail index* $1 < \alpha < 2$ *if* $y < \mathbb{E}\left[\widetilde{Y}\right]$.

The graphs 3.3 illustrate the accuracy of the saddlepoint approximation by comparing the approximated pdf to the simulated stable pdf with the same parameters. Note that the saddlepoint approximation is virtually the same as the exact stable density when $\alpha = 0.5$. Computational difficulties may arise as $\alpha \to 1$. The reason being that the Gamma function $\Gamma(z)$ has a simple pole at $z = 0$. As $\alpha \to 1$, $\Gamma(1-\alpha) \to \infty$ and the saddlepoint tail may decrease faster than the real tail.

In this section, the saddlepoint approximation is applied to estimate distribution of the compound Pareto when the tail index is greater than 2. Section 2.4 introduces the saddlepoint approximation method as a mean to evaluate the sum of n iid random variables. Although the number of summands is random in our case, the saddlepoint method still applies [41, 61, 66]. When the tail index of cluster size satisfies $\alpha > 2$, then the variance exists and is finite and the saddlepoint approximation is well defined.

Proposition 3 shows the procedure to approximate the normalized distribution $\mathbb{P}\{Y = y|H_1\}$; steps to determine $\mathbb{P}\{Y = y|H_0\}$ are similar. First, recall $L = \zeta\varepsilon + \lambda\varsigma$ is the intensity of the Poisson process and denotes the rate of packet flow with unknown source IP addresses under H_1. Further-

more, the characteristic function of a Pareto random variable is $\psi_K(\theta) = \alpha(-i\,c\,\theta)^\alpha \Gamma(-\alpha, -ic\theta)$. Thus,

$$\psi_Y(\theta) = \exp\left\{LT\left[\alpha(-i\theta)^\alpha \Gamma\left(-\alpha, -i\theta\right) - 1\right]\right\} \tag{3.2.29}$$

Proposition 3. *Let* $\bar{Y} = \frac{1}{LT}\sum_{j=1}^{\mathcal{N}(T)} K_j$, *where* $K_j \sim Par(\alpha, 1)$. *Then,*

$$f_{\bar{Y}}(y) \approx \sqrt{\frac{LT}{-2\pi\psi_K''(-i\hat{s})}}\exp\left\{LT\left[\psi_K(-i\hat{s}) - y\hat{s} - 1\right]\right\}\ , \tag{3.2.30}$$

where the saddlepoint $\hat{s}$ *satisfies the equation*

$$\psi_K'(-i\hat{s}) = iy \triangleq -\alpha^2(-\hat{s})^{\alpha-1}\Gamma\left(-\alpha, -\hat{s}\right) - \alpha\frac{e^{\hat{s}}}{(\hat{s})} = y\ , \tag{3.2.31}$$

and

$$\psi_K''(-i\hat{s}) = -\alpha^2(\alpha-1)(-\hat{s})^{\alpha-2}\Gamma\left(-\alpha, -\hat{s}\right) - (\alpha-\alpha^2)\frac{e^{\hat{s}}}{\hat{s}^2} + \alpha\frac{e^{\hat{s}}}{\hat{s}}\ . \tag{3.2.32}$$

Proof. Using results from [25] and the inverse Fourier transform theorem [30], write the pdf $f_{\bar{Y}}(y)$ as

$$\begin{aligned} f_{\bar{Y}}(y) &= \frac{1}{2\pi i}\int_{\tau-i\infty}^{\tau'+i\infty} e^{-\theta x}\mathcal{M}_{\bar{Y}}(i\theta)d\theta \\ &= \frac{1}{2\pi i}\int_{\tau-i\infty}^{\tau'+i\infty} e^{-\theta x}\exp\left\{LT\left[\psi_K\left(\frac{\theta}{LT}\right) - 1\right]\right\}d\theta \\ &= \frac{LT}{2\pi i}\int_{\tau-i\infty}^{\tau'+i\infty} e^{LT[\psi_K(t)-1-ty]}dt \end{aligned}$$

where the last equality follows by letting $t = \frac{\theta}{LT}$. The Taylor expansion of the function $\psi_K(t)-1-ty$ around the saddlepoint $\hat{s}$ is

$$\psi_K(t) - 1 - ty \simeq \psi_K(\hat{s}) - \hat{s}y - 1 + \frac{(t-\hat{s})^2}{2}\psi_K''(\hat{s}) + \mathcal{O}(t^3) \tag{3.2.33}$$

The linear term in the expansion vanishes given that $\hat{s}$ is an extremum for the function $\psi_K(t)-1-ty$. Substitute (3.2.33) in the last integral to obtain

$$f_{\bar{Y}}(y) = \frac{LT}{2\pi i}e^{LT[\psi_K(t)-1-ty]}\int_{\tau-i\infty}^{\tau'+i\infty} e^{\frac{LT}{2}\psi_K''(t)(t-\hat{s})^2}dt$$

Choosing $\tau = \hat{s}$, letting $x = t - \hat{s}$, and using $\int_{-\infty}^{\infty} e^{-ax^2} = \sqrt{\frac{\pi}{a}}$ leads to (3.2.30). The saddlepoint

equation is derived from the Pareto characteristic function.

$$\psi'_K(\theta) = \alpha^2(-i)(-i\theta)^{\alpha-1}\Gamma\left(-\alpha, -i\theta\right) + \alpha(-i\theta)^{\alpha} i \frac{e^{i\theta}}{(-i\theta)^{1+\alpha}}$$

Use the property that $\psi(-i\theta) = \mathcal{M}(\theta)$, write

$$\psi'_K(-i\hat{s}) = \alpha^2(-i)(-\hat{s})^{\alpha-1}\Gamma\left(-\alpha, -\hat{s}\right) - i\alpha\frac{e^{\hat{s}}}{\hat{s}} \ ,$$

where $\hat{s}$ is the saddlepoint we seek to satisfy the equation

$$\psi'_K(-i\hat{s}) = iy \tag{3.2.34}$$

The result follows. □

A closed form solution to the saddlepoint equation (3.2.34) can not be determined. When the saddlepoint equation is intractable as it is with the Pareto distribution, Easton and Ronchetti [31] recommends applying a Taylor expansion to the cdf around at 0

$$\widetilde{\mathcal{K}}_X(\theta) = \mu\theta + \frac{1}{2}\sigma^2\theta^2 + \frac{1}{6}\kappa_3\theta^3 + \frac{1}{24}\kappa_4\theta^4 \ .$$

using the first four cumulants. The resulting saddlepoint equation, $\widetilde{\mathcal{K}}'_X(\theta) = y$, is a polynomial equation of degree four which must be solve numerically. Wang [125] improves Ronchetti's method to prevent the saddlepoint equation from generating multiple roots. The modification is as follows

$$\widetilde{\mathcal{K}}_X(s) = \mu s + \frac{1}{2}\sigma^2 s^2 + \left(\frac{1}{6}\kappa_3 s^3 + \frac{1}{24}\kappa_4 s^4\right) g_b(s) \ ,$$

where $g_b(s) = \exp\left\{-\kappa_2 b^2 s^2/2\right\}$ with $b > 0$.
Lieberman [71] expands $\mathcal{K}_X(\theta)$ around $\theta = 0$ in a Taylor series and deduces the third order approximated saddlepoint.

$$\hat{s}_3 = \frac{y-\mu}{\sigma^2} - \frac{\kappa_3}{2\sigma^2}\left(\frac{y-\mu}{\sigma^2}\right)^2 + \left(\frac{\kappa_3^2}{2\sigma^4} - \frac{\kappa_4}{6\sigma^2}\right)\left(\frac{y-\mu}{\sigma^2}\right)^3 \ .$$

[47] proves that Lieberman's expansion does not always yield accurate results when y increases and proposes instead.

$$\hat{s}_3 = \hat{s}_3 + \frac{y-\mathcal{K}'(\hat{s}_3)}{\mathcal{K}''(\hat{s}_3)} - \frac{\mathcal{K}'''(\hat{s}_3)}{2\mathcal{K}''(\hat{s}_3)}\left(\frac{y-\mathcal{K}'(\hat{s}_3)}{\mathcal{K}''(\hat{s}_3)}\right)^2 + \left(\frac{\left(\mathcal{K}^{(3)}(\hat{s}_3)\right)^2}{2\left(\mathcal{K}''(\hat{s}_3)\right)^2} - \frac{\mathcal{K}^{(4)}(\hat{s}_3)}{6\mathcal{K}''(\hat{s}_3)}\right)\left(\frac{y-\mathcal{K}'(\hat{s}_3)}{\mathcal{K}''(\hat{s}_3)}\right)^3 \ .$$

Ohman-Strickland and Casella [96] compares the performances of these approximations in terms

of relative errors. As shown above, they involve the first four cumulants of the distribution under consideration. Applying those in the Pareto case requires that the tail index satisfies $\alpha > 4$. We choose to compute the saddlepoint numerically using a root finder. The cgf must be convex so that there is a unique saddlepoint. It also implies that the saddlepoint is real. [25]. Figure 3.4 illustrates the compound Poisson-Pareto pdf approximated by the normalized saddlepoint method when the tail index $\alpha = 3.45$.

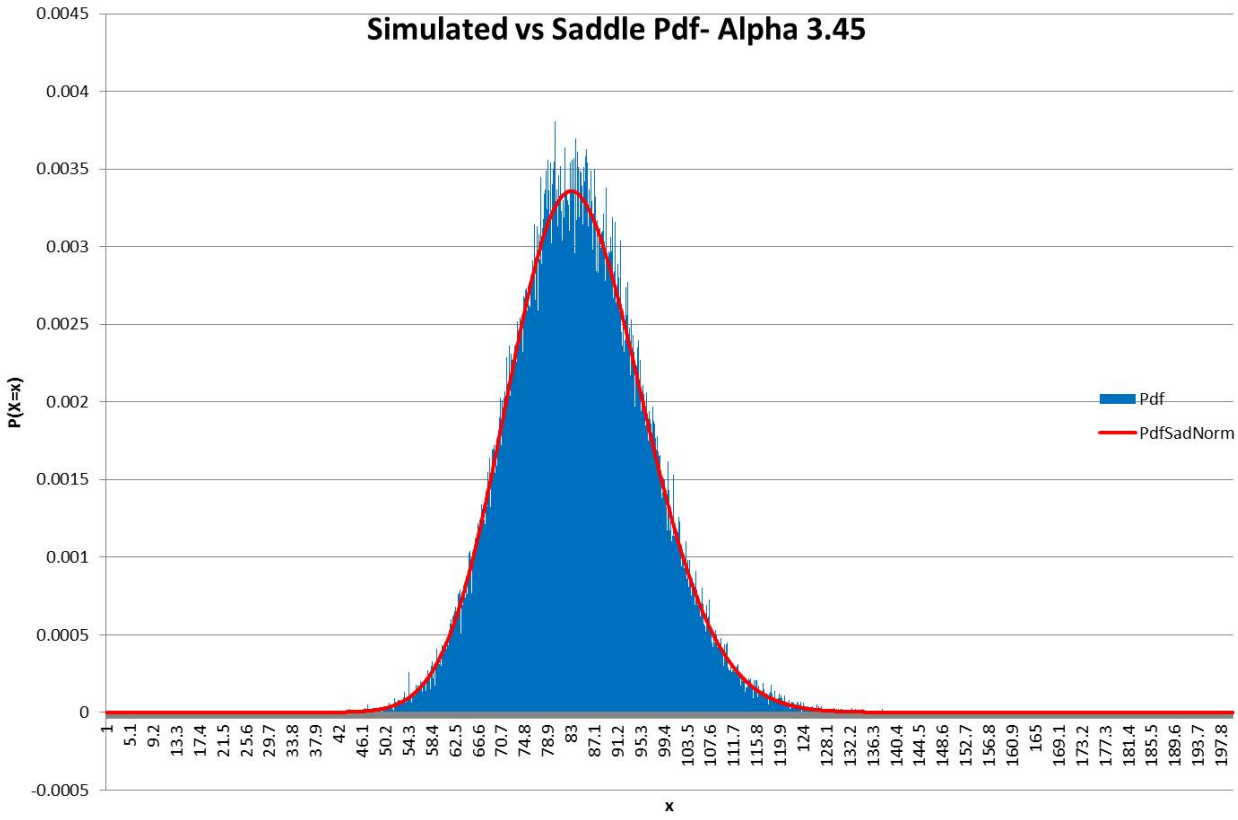


Figure 3.4: Pareto Pdf Simulation vs Saddlepoint Approximation - Alpha 3.45

As the saddlepoint $\hat{s}$ gets closed to 0, which happens when $y \to \mu_K$ in (3.2.30), an approximated closed form solution can be found to solve the saddlepoint equation (3.2.34).

Proposition 4. *As $y \to \mu_K$ in (3.2.30), the saddlepoint $\hat{s}$ can be approximated as*

$$\hat{s} = \left(\frac{\tilde{y} + 1 - \frac{\alpha}{\alpha - 1}}{\alpha^2 (-1)^{\alpha} \Gamma(-\alpha)} \right)^{\frac{1}{\alpha - 1}}, \tag{3.2.35}$$

with $\tilde{y} = \frac{y}{\alpha}$.

Proof. The proof follows from expanding the exponential and gamma terms around 0 in (3.2.34).

The Taylor expansion of the exponential function as $x \to 0$ is

$$e^x \simeq 1 + x + \frac{x^2}{2!} + \dots ,$$

while the Incomplete Gamma function admits the expansion

$$\Gamma(a,z) \simeq \Gamma(a) - \frac{z^a}{a} - \frac{z^{a+1}}{a+1} + \frac{z^{a+2}}{2(a+2)} + \dots ,$$

when $z \to 0$. In the equation below, replace the gamma and exponential function by the expressions above

$$\begin{aligned} \psi'_K(-i\hat{s}) &= \alpha^2(-i)(-\hat{s})^{\alpha-1}\Gamma(-\alpha,-\hat{s}) - i\alpha\frac{e^{\hat{s}}}{\hat{s}} \\ &\simeq \alpha^2(-i)(-\hat{s})^{\alpha-1}\left(\Gamma(-\alpha) - \frac{(-\hat{s})^{-\alpha}}{-\alpha} - \frac{(-\hat{s})^{-\alpha+1}}{-\alpha+1}\right) - i\alpha\left(\frac{1}{\hat{s}}+1\right) \end{aligned}$$

Then solve $\widetilde{\mathcal{K}}'_{\widetilde{Y}}(-i\hat{s}) = iy$ to obtain (4). □

These approximations were computed under the assumption that the rate of the Poisson process is large. When LT is small, one must be careful about the point mass at $\mathcal{N}(T) = 0$ when using saddlepoint approximations. In fact, the compound random variable Y is not continuous at $\mathcal{N}(T) = 0$. Define Y^* as the continuous compound random variable. Lemma gives its characteristic function.

Lemma 9.

$$\psi_{Y^*}(\theta) = \frac{\exp\left\{LT\left[\alpha(-i\theta)^\alpha\Gamma(-\alpha,-i\theta) - 1\right]\right\} - e^{-LT}}{1 - e^{-LT}} \tag{3.2.36}$$

Proof. Jensen [61] proposes to condition on the value of $\mathcal{N}(T)$ to compute $\psi_{Y^*}(\theta)$.

$$\begin{aligned} \psi_Y(\theta) = \mathbb{E}\left[e^{\theta Y}\right] &= \mathbb{E}\left[e^{\theta Y}|\mathcal{N}(T) > 0\right]\mathbb{P}\left\{\mathcal{N}(T) > 0\right\} + \mathbb{P}\left\{\mathcal{N}(T) = 0\right\} \\ &= \psi_{Y^*}(\theta)\left(1 - e^{-LT}\right) + e^{-LT} \end{aligned}$$

Replace $\psi_Y(\theta)$ by its value in (3.2.29) and (3.2.36) follows. □

When LT is large, using the mgf (3.2.36) instead of (3.2.29) does not have much impact on the approximated saddlepoint pdf. However, when LT is small, the modification improves the approximation. For example, when $LT = 10$, $\mathbb{P}\left\{\mathcal{N}(T) = 0\right\} = e^{-LT} \simeq 0.0000454$, which is negligeable. However when $LT = 2$, $\mathbb{P}\left\{\mathcal{N}(T) = 0\right\} = e^{-LT} \simeq 0.135$, and cannot be ignored.

Remark 3. *Few remarks are in order:*

- *saddlepoint approx works very well even for very small values of LT...e.g. LT=10*

- *The greater the tail index, the more accurate the saddlepoint approximation is because second and higher moment exists.*
- *The numerical results section proves that the saddlepoint approximation is accurate for* $\alpha > 2.8$. *In other words, the tail approximation with the saddlepoint method does not require that the third or fourth moment exist.*
- *When* α *is close to 2, the saddlepoint approximation can be improved by increasing the intensity function of the Poisson process eg* $LT = 80$ *produces better results than* $LT = 60$ *for* $\alpha = 2.8$.

The saddlepoint approximation method requires that the Laplace transform of the random variable exists in a neighborhood of zero. For heavy-tail distributions with infinite variance, in this case the Pareto with tail index $1 < \alpha < 2$, this assumption fails [61]. In fact, as the saddlepoint $\hat{s}$ gets closed to zero (which happens when $y \to \mathbb{E}[Y]$), the second derivative of the cgf, i.e. $\mathcal{K}''(\hat{s})$, tends to ∞. This causes the saddlepoint tail to decreases faster than what heavy-tail requires. As illustrated in Figure...the saddlepoint approximation works only when $y < \mu_K$.

As illustrated in Proposition 3, the pdf of compound random variable can be written in its approximated version using saddlepoint approximation. This greatly simplifies the expression of likelihood ratio test. The next section presents the approach.

3.2.4 Application of Saddlepoint Approximation to Hypothesis Testing

Following the procedure described in Proposition 3, the distribution of Y, given H_0, can be estimated using $\zeta\varepsilon$ instead of L, i.e.

$$f_{\bar{Y}}(y) \approx \sqrt{\frac{\zeta\varepsilon T}{-2\pi\psi_K''(-i\hat{s}_0)}} \exp\left\{\zeta\varepsilon T\left[\psi_K(-i\hat{s}_0) - y\hat{s}_0 - 1\right]\right\} . \tag{3.2.37}$$

Under H_0, cluster size are Pareto random variables with tail index α_0. The saddlepoint $\hat{s}_0$ in (3.2.37) can be computed from the saddlepoint equation (3.2.31) by replacing α by α_0. Let $\widetilde{Y}_0$ and $\widetilde{Y}_1$ denote the random variable $\widetilde{Y}$ given H_0 and H_1, respectively. The likelihood ratio can be written as

$$\mathbb{L}(y) = \left(\frac{L}{\zeta\varepsilon}\right)^{1/2} \left(\frac{\psi_K''(-i\hat{s})}{\psi_K''(-i\hat{s}_0)}\right)^{1/2} \frac{\exp\left\{LT\left[\psi_K(-i\hat{s}) - y\hat{s} - 1\right]\right\}}{\exp\left\{\zeta\varepsilon T\left[\psi_K(-i\hat{s}_0) - y\hat{s}_0 - 1\right]\right\}}$$

Solving the Neyman-Pearson test $\mathbb{L}(y) > \eta$ yields

$$\begin{aligned} e^{-y(\hat{s}-\hat{s}_0)} &> \eta \times \left(\frac{L}{\zeta\varepsilon}\right)^{1/2} \left(\frac{\psi_K''(-i\hat{s})}{\psi_K''(-i\hat{s}_0)}\right)^{1/2} \times \frac{\exp\left\{LT\left[\psi_K(-i\hat{s})-1\right]\right\}}{\exp\left\{\zeta\varepsilon T\left[\psi_K(-i\hat{s}_0)-1\right]\right\}} \\ y &\leq -\frac{\ln\tilde{\eta}}{(\hat{s}-\hat{s}_0)} \end{aligned}$$

The Neyman-Pearson threshold η must satisfy

$$\mathbb{P}\left\{\bar{Y} \geq \frac{\ln\tilde{\eta}}{(\hat{s}_0-\hat{s})} | H_0\right\} = \varphi\,, \tag{3.2.38}$$

where

$$\tilde{\eta} = \eta \times \left(\frac{L}{\zeta\varepsilon}\right)^{1/2} \left(\frac{\psi_K''(-i\hat{s})}{\psi_K''(-i\hat{s}_0)}\right)^{1/2} \times \frac{\exp\left\{LT\left[\psi_K(-i\hat{s})-1\right]\right\}}{\exp\left\{\zeta\varepsilon T\left[\psi_K(-i\hat{s}_0)-1\right]\right\}}$$

Solving (3.2.38) requires the complementary cdf of $\bar{Y}$. $\bar{F}_{\bar{Y}}(y) = 1 - \mathbb{P}\left\{\bar{Y} = y\right\}$ can be computed by numerical integration of the pdf given by the saddlepoint approximation in (3.2.30). Another option is to apply the saddlepoint technique directly to approximate $\bar{F}_{\bar{Y}}(y)$.

As shown in (3.2.38), we are interested in the complementary cdf under H_0. Given H_0, recall $Y = \sum_{j=1}^{U(T)} K_j = \zeta\varepsilon T\bar{Y}$, with characteristic function $\exp\left\{\zeta\varepsilon T\left[\psi_K(\theta)-1\right]\right\}$. Let $\mathcal{K}_Y(\theta) = \ln\psi_Y(\theta) = \zeta\varepsilon T\left[\psi_K(\theta)-1\right]$. Using results from [75], the complementary cdf of Y can be approximated as

$$\bar{F}_Y(y) = \begin{cases} 1 - \Phi(\hat{w}) + \phi(\hat{w})\left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right) & y \neq \mathbb{E}[Y] \\ \frac{1}{2} + \frac{\mathcal{K}_Y'''(0)}{6\sqrt{2\pi}\left(\mathcal{K}_Y''(0)\right)^{3/2}} & y = \mathbb{E}[Y] \end{cases} \tag{3.2.39}$$

where Φ and ϕ are the cdf and pdf of the standard normal distribution, respectively, and

$$\begin{aligned} \hat{w} &= \operatorname{sign}(\hat{s})\sqrt{2\hat{s}y - 2\mathcal{K}_Y(-i\hat{s})} \\ \hat{u} &= \hat{s}\sqrt{-\mathcal{K}_Y''(-i\hat{s})} \end{aligned}$$

$\hat{s}$ is the saddlepoint that satisfies the equation

$$\mathcal{K}_Y'(-i\hat{s}) = iy \triangleq \zeta\varepsilon T\left[-\alpha^2(-c\hat{s})^{\alpha-1}\Gamma\left(-\alpha,-c\hat{s}\right) - \alpha_1\frac{e^{c\hat{s}}}{(c\hat{s})}\right] = y.$$

Furthermore,

$$\begin{aligned} \mathcal{K}_Y(-i\hat{s}) &= \alpha\zeta\varepsilon T(-c\hat{s})^{\alpha}\Gamma(-\alpha,-c\hat{s}) - \zeta\varepsilon T \\ \mathcal{K}_Y''(-i\hat{s}) &= \zeta\varepsilon T\left[-\alpha_1^2(\alpha_1-1)(-c_1\hat{s})^{\alpha_1-2}\Gamma\left(-\alpha_1,-c_1\hat{s}\right) - (\alpha_1-\alpha_1^2)\frac{e^{c_1\hat{s}}}{(c_1\hat{s})^2} + \alpha\frac{e^{c_1\hat{s}}}{c_1\hat{s}}\right] \end{aligned}$$

Note that if $y = \mathbb{E}[Y]$, then $\mathcal{K}'_Y(0) = \mathbb{E}[Y]$ and the saddle point $\hat{s} = 0$ [17]. In that case $\hat{w} = 0$ and the first equation in (3.2.39) is undefined. Therefore, the cdf must be approximated using the second formula.

The main assumption so far was that packet flows arrive at the control router according to an homogeneous Poisson process. In reality, the intensity function of the Poisson process is not constant but varies with time. The following section addresses that case.

3.2.5 Inhomogeneous Marked Poisson

Assume the packet flow process is an inhomogeneous Poisson process. This model is a more realistic representation of the network traffic. As described in the previous section, each flow are marked with two random variables. Cluster sizes are Pareto distributed and the Bernoulli mark is such that a legitimate and attack packet flow (and all data packets in the flow) have unknown source IP address with probability ε and ς, respectively. The thinning property of Poisson processes applies in the inhomogeneous case as well. Hence, the legitimate and attack process of flows with unknown source IP addresses, i.e. $\{U(t), t \geq 0\}$ and $\{A(t), t \geq 0\}$, have intensity functions $\varepsilon\zeta(t)$ and $\varsigma\lambda(t)$. Under H_1, the two independent inhomogeneous Poisson processes are additive [109]. The corresponding binary hypothesis testing is

$$\begin{cases} H_0 : Y \sim \text{Compound Poisson with intensity } \varepsilon\zeta(t) \\ H_1 : Y \sim \text{Compound Poisson with intensity } \varepsilon\zeta(t) + \varsigma\lambda(t) \end{cases} \tag{3.2.40}$$

Let $L(t) = \varepsilon\zeta(t) + \varsigma\lambda(t)$. Paxson and Floyd [100] emphasize that TCP traffic can be well modeled as an inhomogeneous Poisson process with a piecewise constant intensity function. Therefore results from the Section 3.2 apply, assuming that the observation window of size T is chosen adequately. Otherwise, saddlepoint approximations can be directly applied in the inhomogeneous Poisson case [41]. Given H_1, the total number of data packets with unknown source IP address is $Y = \sum_{j=1}^{\mathcal{N}(t)} K_j$, where $\mathcal{N}(t)$ follows a Poisson distribution with intensity $\Lambda(t)$. From [114], Y has characteristic function

$$\begin{aligned} \psi_Y(\theta) &= \exp\left\{ \int_0^T \left[\psi_K(\theta) - 1\right] L(\upsilon)\, d\upsilon \right\} \\ &= \exp\left\{ \Lambda(T) \left[\psi_K(\theta) - 1\right] \right\}, \end{aligned} \tag{3.2.41}$$

where $\Lambda(T)$ is the mean value function of the inhomogeneous process defined as

$$\Lambda(T) = \int_0^T L(\upsilon)\, d\upsilon. \tag{3.2.42}$$

So far, we have addressed the case where cluster sizes were distributed according to a pure Pareto distribution. However, Section 2.1 emphasizes that mixture distributions are more adequate models for cluster size. Therefore the next section applies the saddlepoint approximation method to mixture distributions. Provided that the variance of the mixture random variable is finite, the first order saddlepoint can approximate the pdf f_K of cluster size K.

3.3 Marked Poisson with Mixture Cluster Size

Consider that cluster sizes $\{K_j\}$ are mixture random variables. The distribution function of a finite mixture distribution with m mixture components can be written as

$$F_K(x) = \sum_{i=1}^{m} q_i F_i(x). \tag{3.3.1}$$

The corresponding characteristic function is

$$\psi_K(\theta) = \sum_{i=1}^{m} q_i \psi_i(\theta), \tag{3.3.2}$$

where the weights q_i satisfy $0 \leq q_i \leq 1$ and $\sum_{i=1}^{m} q_i = 1$.
We present the saddlepoint approximation for two particular cases: the mixture Pareto/Pareto and mixture discrete /Pareto distributions. More complex mixtures such as discrete/Pareto/Pareto (see distribution in (2.1.3)) can be approximated similarly. Saddlepoint approximations for the mixture distribution basically involve the same steps as with single component distributions. Alternatively, the FFT method can also be applied to approximate the mixture distribution function.

3.3.1 Mixture Pareto-Pareto

Assume the distribution of cluster sizes $\{K_j\}$ is composed of two Pareto distributions with parameters (α_1, c_1) and (α_2, c_2), respectively. That is,

$$F_K(k) = (1-q)\left[1 - \left(\frac{c_1}{k}\right)^{\alpha_1}\right] + q\left[1 - \left(\frac{c_2}{k}\right)^{\alpha_2}\right].$$

Then the characteristic function of K_j, for any flow j, can be written as

$$\psi_K(\theta) = (1-q)\alpha_1(-i\,c_1\,\theta)^{\alpha_1}\Gamma(-\alpha_1, -ic_1\,\theta) + q\alpha_2(-i\,c_2\,\theta)^{\alpha_2}\Gamma(-\alpha_2, -ic_2\,\theta), \tag{3.3.3}$$

with $0 \leq q \leq 1$.

Define the average number of data packets with unknown source IP addresses as $\bar{Y} = \frac{1}{LT}\sum_{j=1}^{\mathcal{N}(T)} K_j$. We focus on computing the distribution of $\bar{Y}$ given H_1, the procedure for H_0 is similar. Proposition 3 shows how to compute the pdf of $\bar{Y}$ with Pareto cluster sizes if at least the second moment exists. Approximating $f_{\bar{Y}}$ with mixture cluster size follows (3.2.30) using (3.3.3) as the characteristic function. The saddlepoint $\hat{s}$ must satisfy

$$\begin{aligned}\psi'_K(-i\hat{s}) = iy &\triangleq (1-q)\left[-\alpha_1^2(-c_1\hat{s})^{\alpha_1-1}\Gamma\left(-\alpha_1,-c_1\hat{s}\right) - \alpha_1\frac{e^{c_1\hat{s}}}{(c_1\hat{s})}\right] \\ &+ q\left[-\alpha_2^2(-c_2\hat{s})^{\alpha_2-1}\Gamma\left(-\alpha_2,-c_2\hat{s}\right) - \alpha_2\frac{e^{c_2\hat{s}}}{(c_2\hat{s})}\right] = y\ .\end{aligned}$$

The saddlepoint $\hat{s}$ is numerically computed using a rootfinder. Saddlepoint approximations also require the second derivative of the cgf as

$$\begin{aligned}\psi''_K(-i\hat{s}) &= (1-q)\left[-\alpha_1^2(\alpha_1-1)(-c_1\hat{s})^{\alpha_1-2}\Gamma\left(-\alpha_1,-c_1\hat{s}\right) - (\alpha_1-\alpha_1^2)\frac{e^{c_1\hat{s}}}{(c_1\hat{s})^2} + \alpha\frac{e^{c_1\hat{s}}}{c_1\hat{s}}\right] \\ &+ q\left[-\alpha_2^2(\alpha_2-1)(-c_2\hat{s})^{\alpha_2-2}\Gamma\left(-\alpha_2,-c_2\hat{s}\right) - (\alpha_2-\alpha_2^2)\frac{e^{c_2\hat{s}}}{(c_2\hat{s})^2} + \alpha\frac{e^{c_2\hat{s}}}{c_2\hat{s}}\right]\end{aligned}$$

Tail probabilities, i.e. $\bar{F}_{\bar{Y}}$ the complementary cdf of the compound Pareto random variable, can be approximated by the Luganni-Rice formula [75] as described in (3.2.39). The saddlepoint approximation may not be always applicable. In that case, the FFT algorithm can be used to approximate $f_{\bar{Y}}(y)$ based on the mixture characteristic function given in (3.3.3).

If $f_{\bar{Y}}(y)$, the pdf of the average number of unknown source IP addresses ($\bar{Y} = \frac{1}{LT}\sum_{j=1}^{\mathcal{N}(T)} K_j$) is evaluated for y large, then the saddlepoint approximation can be applied to the Pareto component with the smallest tail index if the latter is greater than 2. Otherwise, the pdf $f_{\bar{Y}}(y)$ can be approximated by a stable distribution if $0 < \alpha < 2$ with the minimum tail index. In fact, the asymptotic tail of the sum of two independent power laws is dominated by the component with the heaviest tail, i.e., the power law with the smallest tail index [78, 40]. Proposition gives the small θ expansion of the characteristic function of a mixture of two Pareto distributions.

Proposition 5. *Define the normalized cluster size as*

$$\hat{K}_i = \frac{K_i - \mu_K}{(LT)^{1/\alpha}}\ ,$$

with

$$\mu_K = (1-q)\frac{\alpha_1 c_1}{\alpha_1 - 1} + q\frac{\alpha_2 c_2}{\alpha_2 - 1}. \tag{3.3.4}$$

As $LT \to \infty$, i.e. $\theta \to 0$, the characteristic function of the normalized number of unknown source

IP addresses, $Y = \sum K_j$*, satisfies*

$$\psi_{\hat{Y}}(\theta) \approx (1-q)\alpha_1(-ic_1\theta)^{\alpha_1}\Gamma(-\alpha_1) + q\alpha_2(-ic_2\theta)^{\alpha_2}\Gamma(-\alpha_2) \tag{3.3.5}$$

Let $\alpha_m = \min\{\alpha_1, \alpha_2\}$*, then*

$$\psi_{\hat{Y}}(\theta) \approx \alpha_m(-ic\theta)^{\alpha}_{m}\Gamma(-\alpha_m) \tag{3.3.6}$$

Proof. Using the property of characteristic functions, i.e. $\psi_{aX+b}(\theta) = e^{i\theta b}\psi_X(a\theta)$, we can write

$$\psi_{\hat{K}}(\theta) = \exp\left\{-i\theta\mu_K (LT)^{-\frac{1}{\alpha}}\right\}\psi_K\left(\frac{\theta}{(LT)^{1/\alpha}}\right), \tag{3.3.7}$$

where

$$\psi_K(\theta) = (1-q)\alpha_1(-i\,c_1\,\theta)^{\alpha_1}\Gamma(-\alpha_1, -ic_1\,\theta) + q\alpha_2(-i\,c_2\,\theta)^{\alpha_2}\Gamma(-\alpha_2, -ic_2\,\theta). \tag{3.3.8}$$

As $LT \to \infty$, $\theta \to 0$ and the characteristic function of the normalized cluster size $\hat{K_j}$ can be expanded as

$$\begin{aligned}
\psi_{\hat{K}}(\theta) &= \left[1 - \frac{i\theta\mu_K}{(LT)^{1/\alpha}} - \frac{1}{2}\left(\frac{\theta\,\mu_K}{(LT)^{1/\alpha}}\right)^2 + \ldots\right] \\
&\times \left[(1+q)\left(1 + \alpha_1\frac{(-ic_1\theta)^{\alpha_1}}{LT}\Gamma(-\alpha_1) + \frac{\alpha_1}{\alpha_1 - 1}\frac{(ic_1\theta)}{(LT)^{1/\alpha_1}} + \ldots\right)\right. \\
&+ \left. q\left(1 + \alpha_2\frac{(-ic_2\theta)^{\alpha_2}}{LT}\Gamma(-\alpha_2) + \frac{\alpha_2}{\alpha_2 - 1}\frac{(ic_2\theta)}{(LT)^{1/\alpha_2}} + \ldots\right)\right]
\end{aligned}$$

Theorem 5 indicates that the effect of the Poisson is negligible. Thus, for simplicity consider $\bar{Y} = \sum_{j=1}^{LT}\hat{K}_j$. The logarithm function admits the following Taylor expansion as $\ln(1+x) \approx x - \frac{x^2}{2} + \ldots$, or alternatively $\ln(x) \approx (x-1) - \frac{(x-1)^2}{2} + \ldots$.

Therefore,

$$\begin{aligned}
\psi_{\hat{Y}}(\theta) = \left[\psi_{\hat{K}}(\theta)\right]^{LT} &= \exp\left\{LT \ln \psi_{\hat{K}}(\theta)\right\} \\
&= \exp\left\{ LT \left[-\frac{i\theta\mu_K}{(LT)^{1/\alpha}} - \frac{1}{2}\left(\frac{\theta\,\mu_K}{(LT)^{1/\alpha}}\right)^2 + \ldots \right] \right. \\
&+ LT \ln \left[(1+q)\left(1 + \alpha_1 \frac{(-ic_1\theta)^{\alpha_1}}{LT}\Gamma(-\alpha_1) + \frac{\alpha_1}{\alpha_1 - 1}\frac{(ic_1\theta)}{(LT)^{1/\alpha_1}} + \ldots\right)\right. \\
&\left.\left. + q\left(1 + \alpha_2 \frac{(-ic_2\theta)^{\alpha_2}}{LT}\Gamma(-\alpha_2) + \frac{\alpha_2}{\alpha_2 - 1}\frac{(ic_2\theta)}{(LT)^{1/\alpha_2}} + \ldots\right)\right]\right\} \\
&= \exp\left\{ \left[-\frac{i\theta\mu_K}{(LT)^{1/\alpha-1}} - \frac{1}{2}\frac{\theta^2\,\mu_K^2}{(LT)^{2/\alpha-1}} + \ldots \right] \right. \\
&+ (1-q)\left(LT + \alpha_1(-ic_1\theta)^{\alpha_1}\Gamma(-\alpha_1) + \frac{\alpha_1}{\alpha_1 - 1}\frac{(ic_1\theta)}{(LT)^{1/\alpha_1 - 1}} + \ldots\right) \\
&\left. + q\left(LT + \alpha_2(-ic_2\theta)^{\alpha_2}\Gamma(-\alpha_2) + \frac{\alpha_2}{\alpha_2 - 1}\frac{(ic_2\theta)}{(LT)^{1/\alpha_2 - 1}} + \ldots\right) - LT\right\}
\end{aligned}$$

As $LT \to \infty$, $\psi_{\hat{Y}}(\theta) \approx (1-q)\alpha_1(-ic_1\theta)_1^{\alpha}\Gamma(-\alpha_1) + q\alpha_2(-ic_2\theta)_2^{\alpha}\Gamma(-\alpha_2)$ since all other terms tend to zero. The result follows by replacing μ_K by its expression given in (3.3.4). If $0 < \alpha_1, \alpha_2 < 2$, the characteristic function describes the mixture of two stable distributions (see theorems 6 and 7), which have asymptotic power laws. The mixture of power law distributions is dominated by the power law with the smallest tail index. □

3.3.2 Mixture Discrete-Pareto

Consider the following mixture discrete/Pareto distribution

$$F_K(k) = (1-q)\sum_{x_i \le k} \mathbb{P}\{K = x_i\} + q\left[1 - \left(\frac{c}{k}\right)^{\alpha}\right],$$

where $x_i \in \{1, 2, \ldots, \lfloor c \rfloor - 1\}$ and with $0 \le q \le 1$. The corresponding characteristic function can be written as

$$\psi_K(\theta) = (1-q)\sum_{i=1}^{\lfloor c \rfloor - 1} e^{i\theta x_i}\mathbb{P}\{K = x_i\} + q\alpha(-i\,c\,\theta)^{\alpha}\Gamma(-\alpha, -ic\,\theta), \tag{3.3.9}$$

with derivative

$$\psi'_K(\theta) = (1-q)\sum_{i=1}^{\lfloor c\rfloor-1} ix_i e^{i\theta x_i}\mathbb{P}\left\{K=x_i\right\} + q\left[-i\alpha^2(-ic\theta)^{\alpha-1}\Gamma\left(-\alpha,-ic\theta\right) - \alpha\frac{e^{ic\theta}}{(c\theta)}\right],$$

Saddlepoint approximations applies to discrete random variables. For example, Davison and Kuonen [26] compute the distribution of a bootstrap statistic, using the saddlepoint method. Huang et al. [56] also applies the technique to compute the pdf of the Bernoulli mixture model. The saddlepoint equation is

$$\psi'_K(-i\hat{s}) = iy \triangleq (1-q)\sum_{i=1}^{\lfloor c\rfloor-1} x_i e^{c\hat{s}x_i}\mathbb{P}\left\{K=x_i\right\} + q\left[-\alpha^2(-c\hat{s})^{\alpha-1}\Gamma\left(-\alpha,-c\hat{s}\right) + i\alpha\frac{e^{c\hat{s}}}{(c_2\hat{s})}\right] = y.$$

The approximated pdf is given in (3.2.30) where the second derivative becomes

$$\begin{aligned}\psi''_K(-i\hat{s}) &= -(1-q)\sum_{i=1}^{\lfloor c\rfloor-1} x_i^2 e^{c\hat{s}x_i}\mathbb{P}\left\{K=x_i\right\} \\ &+ q\left[-\alpha^2(\alpha-1)(-c\hat{s})^{\alpha-2}\Gamma\left(-\alpha,-c\hat{s}\right) - (\alpha-\alpha^2)\frac{e^{c\hat{s}}}{(c\hat{s})^2} + \alpha\frac{e^{c\hat{s}}}{c\hat{s}}\right].\end{aligned}$$

Note that saddlepoint approximation smoothes over discrete distribution functions. Alternatively, Davison and Mastropietro [27] propose that the discrete component be first removed from the mgf. The saddlepoint approximation can then be applied to continuous component alone. Once the approximated continuous cdf is obtained, the discrete component can be reinserted. However, note that hypothesis testing requires tail probabilities with small limits on error probabilities. In fact, φ and ξ can be as low as 0.05 or 0.01. We are then interested in tail probabilities as $\bar{F}(y) = 1-\mathbb{P}\left\{Y=y\right\} = \varphi = 0.05, 0.025$. Smoothing over the discrete component of the mixture distribution does not affect the tail probabilities [27]. Proposition 6 demonstrates that the Pareto component dominates the upper tail, suggesting that the discrete distribution has a limited contribution.

Proposition 6. *Define the normalized cluster size as*

$$\hat{K}_i = \frac{K_i - \mu_K}{(LT)^{1/\alpha}},$$

with

$$\mu_K = (1-q)\sum_{i=1}^{\lfloor c\rfloor-1} x_i\mathbb{P}\left\{K=x_i\right\} + q\frac{\alpha c}{\alpha-1}. \tag{3.3.10}$$

As $LT\to\infty$, i.e. $\theta\to 0$, the characteristic function of the normalized number of unknown source

IP addresses, $Y = \sum K_j$*, satisfies*

$$\psi_{\hat{Y}}(\theta) \approx q\alpha(-ic\theta)^{\alpha}\Gamma(-\alpha) \tag{3.3.11}$$

Proof. The proof is similar to that of Proposition 5. Formula (3.3.7) holds with

$$\psi_K(\theta) = (1-q)\sum_{i=1}^{\lfloor c\rfloor-1} e^{i\theta x_i}\mathbb{P}\{K=x_i\} + q\alpha(-i\,c\,\theta)^{\alpha}\Gamma(-\alpha,-ic\,\theta), \tag{3.3.12}$$

Let $p_i = \mathbb{P}\{K = x_i\}$.
Around zero, the characteristic function of the normalized cluster size $\hat{K}_j$ can be expanded as

$$\begin{aligned}
\psi_{\hat{K}}(\theta) &= \left[1 - \frac{i\theta\mu_K}{(LT)^{1/\alpha}} - \frac{1}{2}\left(\frac{\theta\,\mu_K}{(LT)^{1/\alpha}}\right)^2 + \dots\right] \\
&\times \left[(1+q)\sum_{i=1}^{\lfloor c\rfloor-1} p_i\left(1 + \frac{i\theta x_i}{(LT)^{1/\alpha}} - \frac{1}{2}\frac{\theta^2 x_i^2}{(LT)^{2/\alpha}} + \dots\right)\right. \\
&+ \left. q\left(1 + \alpha\frac{(-ic\theta)^{\alpha}}{LT}\Gamma(-\alpha) + \frac{\alpha}{\alpha-1}\frac{(ic\theta)}{(LT)^{1/\alpha}} + \dots\right)\right]
\end{aligned}$$

Based on Theorem 5, consider $\bar{Y} = \sum_{j=1}^{LT}\hat{K}_j$. From the Taylor expansion of the logarithm function,

$$\begin{aligned}
\psi_{\hat{Y}}(\theta) = \left[\psi_{\hat{K}}(\theta)\right]^{LT} &= \exp\left\{LT\ln\psi_{\hat{K}}(\theta)\right\} \\
&= \exp\left\{LT\left[-\frac{i\theta\mu_K}{(LT)^{1/\alpha}} - \frac{1}{2}\left(\frac{\theta\,\mu_K}{(LT)^{1/\alpha}}\right)^2 + \dots\right]\right. \\
&+ LT\ln\left[\sum_{i=1}^{\lfloor c\rfloor-1}(1-q)p_i\left(1 + \frac{i\theta x_i}{(LT)^{1/\alpha}} - \frac{1}{2}\frac{\theta^2 x_i^2}{(LT)^{2/\alpha}} + \dots\right)\right. \\
&+ \left.\left. q\left(1 + \alpha\frac{(-ic\theta)^{\alpha}}{LT}\Gamma(-\alpha) + \frac{\alpha}{\alpha-1}\frac{(ic\theta)}{(LT)^{1/\alpha}} + \dots\right)\right]\right\} \\
&= \exp\left\{\left[-\frac{i\theta\mu_K}{(LT)^{1/\alpha-1}} - \frac{1}{2}\frac{\theta^2\,\mu_K^2}{(LT)^{2/\alpha-1}} + \dots\right]\right. \\
&+ \sum_{i=1}^{\lfloor c\rfloor-1}(1-q)p_i\left(LT - \frac{i\theta x_i}{(LT)^{1/\alpha-1}} - \frac{1}{2}\frac{\theta^2 x_i^2}{(LT)^{2/\alpha-1}} + \dots\right) \\
&+ \left. q\left(LT + \alpha(-ic\theta)^{\alpha}\Gamma(-\alpha) + \frac{\alpha}{\alpha-1}\frac{(ic\theta)}{(LT)^{1/\alpha-1}} + \dots\right) - LT\right\}
\end{aligned}$$

As $LT \to \infty$, $\psi_{\hat{Y}}(\theta) \approx q\alpha(-ic\theta)^{\alpha}\Gamma(-\alpha)$ since all other terms tend to zero. The result follows by

replacing μ_K by its expression given in (3.3.10).
We conclude the the Pareto component dominates in the tails. □

In order to deduce quantile values, we have shown how to use different approximation tools to evaluate the pdf of the random variable of interest. Instead of modeling the entire distribution, the focus could be on understanding the upper tail behavior of distribution functions. The next section present some well-known results in large deviation theory for heavy tailed distributions and show how they can be applied in hypothesis testing.

3.4 Detection and Large Deviations for Heavy-Tailed Distributions

Recall the binary hypothesis testing based on the percentage of unknown source IP addresses can be formulated as

$$\begin{cases} H_0 : Y \sim \text{Compound Poisson with mean } (\zeta\varepsilon)T \\ H_1 : Y \sim \text{Compound Poisson with mean } (\zeta\varepsilon + \lambda\varsigma)T \ . \end{cases} \tag{3.4.1}$$

Assume the Neyman-Pearson test can be expressed in terms of $Y = \sum K_j$, the total number of unknown source IP addresses. This is usually the case, when the conditional pdfs, given H_0 and H_1, have closed forms and the likelihood ratio can be expressed as a invertible function of the test statistic. For instance, this is the case when we monitor the percentage of distinct source IP addresses (see Section 3.1.1). Then, the binary hypothesis test compares two Poisson random variables. The log-likelihood ratio can be inverted and the Neyman-Pearson test $\mathbb{L}(y) \underset{H_1}{\overset{H_0}{\lessgtr}} \eta$ is written as $y \underset{H_1}{\overset{H_0}{\lessgtr}} \tilde{\eta}$. Assume $Y = y \underset{H_1}{\overset{H_0}{\lessgtr}} \tilde{\eta}$ is a valid test when Y is a compound Poisson-Pareto random variable. Then the NP threshold, $\tilde{\eta}$, must be chosen such that $\mathbb{P}\{Y > \tilde{\eta}|H_0\} = \varphi$. The next sections show how to use large deviations results to deduce quantiles and ultimately compute $\tilde{\eta}$.

3.4.1 Compound Poisson and Heavy-Tail Distribution

Let $S_n = \sum_{i=1}^n X_i$, where $\{X_i\}$ are iid random variables with cdf F_X, which has a heavy tail i.e., $\bar{F}_X(x) = 1 - F_X(x) = x^{-\alpha}L(x)$. Recall $L(x)$ is a slowly varying function. For the Pareto distribution with shape and scale parameters, α and c, $L(x) = c^\alpha$, a positive and constant function. Assume the tail index satisfies the condition $\alpha > 1$ so that the random variables $\{X_i\}$ have a common finite mean μ_X. Using the subexponentiality property (3.2.1), Nagaev [90] proves that as $n \to \infty$,

$$\mathbb{P}\{S_n - \mathbb{E}[S_n] > x\} \sim \mathbb{P}\{\max\{X_1, X_2, \ldots, X_n\} > x\} \sim n\bar{F}(x) \ , \tag{3.4.2}$$

with $\mathbb{E}[S_n] = n\mu_X$.

Large deviation results for random sums derive from the case where the number of summands is a fixed number n. Klüppelberg and Mikosch [68] extend the result (3.4.2) to the case where the number of summand is a random variable. In particular, they address the case where the number of summands is a Poisson random variable as described above. Let $\lambda(t)$ to be the intensity function of $\mathcal{N}(t)$, a Poisson process. Recall $\mathcal{N}(t)$ is homogeneous if $\lambda(t) = \lambda$ for all t, where λ is a positive constant; else it is inhomogeneous. Let $S(t) = \sum_{i=1}^{\mathcal{N}(t)} X_i$, where $\{X_i\}$ and $\mathcal{N}(t)$ are assumed to be independent as well. $\{X_i\}$ are iid random variables with the common cdf F_X defined above. If $\lambda(t) \to \infty$ then for every fixed $\rho > 0$ and $x \geq \rho\lambda(t)$,

$$\mathbb{P}\{S(t) - \mu(t) > x\} \sim \lambda(t)\bar{F}(x), \tag{3.4.3}$$

where $\mu(t) = \mathbb{E}[S(t)] = \mu_X\lambda(t)$. Equation (3.4.3) holds if $\{X_i\}$ have a finite mean. The condition is violated if the Pareto cluster size have a tail index such that $0 < \alpha < 1$.

Based on the tail approximation (3.4.3), we deduce the Neyman-Pearson threshold $\tilde{\eta}$.

Proposition 7. *Assume the Neyman-Pearson test for the binary hypothesis test in (3.4.1) can be expressed as* $Y = y \underset{H_1}{\overset{H_0}{\lessgtr}} \tilde{\eta}$ *where* $Y = \sum_{j=1}^{\mathcal{N}(t)} K_j$ *and* $K_j \sim Par(\alpha, c)$*, with* $\alpha > 1$*. Then*

$$\tilde{\eta} \approx c\left(\frac{\zeta\varepsilon T}{\varphi}\right)^{1/\alpha} + (\zeta\varepsilon)T\frac{\alpha c}{\alpha - 1} \tag{3.4.4}$$

Proof. The complementary cdf of the Pareto distribution is $\bar{F}(k) = \left(\frac{c}{k}\right)^{\alpha}$. Moreover, the NP threshold, $\tilde{\eta}$, must be chosen such that $\mathbb{P}\{Y > \tilde{\eta}|H_0\} = \varphi$, we deduce that $\mu_Y = (\zeta\varepsilon)T\mu_K$, where μ_K is the Pareto mean. Then by (3.4.3),

$$\mathbb{P}\{Y - \mu_Y > y\} \sim \zeta\varepsilon T\left(\frac{c}{y}\right)^{\alpha},$$

Let φ be the limit on the probability of false alarm. We solve $\zeta\varepsilon T\left(\frac{c}{y}\right)^{\alpha} = \varphi$ to deduce $y = c\left(\frac{\zeta\varepsilon T}{\varphi}\right)^{1/\alpha}$. Since $Y - \mu_Y > y$ then $Y > \tilde{\eta} \approx y + \mu_Y$. □

Note that in the mixture distribution case, the complementary cdf can be computed as

$$\bar{F}_K(k) = 1 - F_K(k) = q\sum_{x_i \leq k}(1 - \mathbb{P}\{K = x_i\}) + q_1\left(\frac{c_1}{k}\right)^{\alpha_1} + q_2\left(\frac{c_2}{k}\right)^{\alpha_2}. \tag{3.4.5}$$

In that case the NP threshold can be computed by numerically inverting Expression (3.4.5).

Section 3.2.2 shows that the compound Pareto random variable can be approximated by a stable

distribution if the tail index satisfies $0 < \alpha < 2$. Fofack and Nolan [38] present an extensive study on large deviations, tail probabilities and moments of stable distributions. More details can also be found in [70, 92, 110]. In the following section, we propose to compute the Neyman-Pearson threshold based on large deviation results for stable distributions.

3.4.2 Tail Probability for Stable Distribution

Given $\alpha < 2$, Lévy [70] proves that the tails of stable distributions behave asymptotically as a Pareto law. That is if $X \sim \mathcal{S}_\alpha(1, \beta, 0)$ then

$$\mathbb{P}\{X > x\} = 1 - F_X(x) \to \mathcal{C}_\alpha(1+\beta)x^{-\alpha}\,, \tag{3.4.6}$$

where

$$\mathcal{C}_\alpha = \left(2\int_0^\infty x^{-\alpha}\sin x\,dx\right)^{-1} = \frac{1}{\pi}\Gamma(\alpha)\sin\frac{\pi\alpha}{2}\,. \tag{3.4.7}$$

The larger the tail index, the slower the convergence [127]. Nolan [92] extends (3.4.6) to stable random variables $X \sim \mathcal{S}_\alpha(\gamma, \beta, 0)$ and shows that

$$\mathbb{P}\{X > x\} \sim \gamma^\alpha \mathcal{C}_\alpha(1+\beta)x^{-\alpha}\,, \tag{3.4.8}$$

where $\mathcal{C}_\alpha$ is computed as in (3.4.7). Equation (3.4.8) corresponds exactly to the upper tail of a Pareto distribution. One major disadvantage of this tail approximation results from not knowing the point at which the asymptotic is accurate [38]. Although they could not provide a precise way about the asymptotic point, Fofack and Nolan [38] attempt to compare the tails of stable and Pareto distribution to understand when (3.4.8) holds.

Using the same approach as in Proposition 7, Proposition 8 computes the NP threshold as a function of the tail probability in (3.4.6).

Proposition 8. *Given H_0, define the normalized number of unknown source IP addresses as*

$$\widetilde{Y} = \begin{cases} \frac{1}{(\zeta\varepsilon T)^{1/\alpha}} \sum_{j=1}^{U(T)} K_j & \text{if } 0 < \alpha < 1 \\ \frac{1}{(\zeta\varepsilon T)^{1/\alpha}} \left(\sum_{j=1}^{U(T)} K_j - \zeta\varepsilon T\mu_K\right) & \text{if } 1 < \alpha < 2 \end{cases} \tag{3.4.9}$$

If the Neyman-Pearson test can be expressed as $\widetilde{Y} = y \underset{H_1}{\overset{H_0}{\lessgtr}} \tilde{\eta}$, then

$$\tilde{\eta} \approx \left(\frac{\varphi}{2\gamma^\alpha \mathcal{C}_\alpha}\right)^{-1/\alpha}, \tag{3.4.10}$$

where $\mathcal{C}_\alpha$ is computed as in (3.4.7) and

$$\gamma^\alpha = \Gamma(1-\alpha)\cos\left(\frac{\pi\alpha}{2}\right). \tag{3.4.11}$$

Proof. Theorem 6 and Theorem 7 show that the pdf of $\widetilde{Y}$ can be approximated by a stable distribution with parameters $\gamma^\alpha = \Gamma(1-\alpha)\cos\left(\frac{\pi\alpha}{2}\right)$; $\beta = 1$; $\delta = 0$. Equation (3.4.10) follows by setting the tail value equal to φ, the desired limit on false alarm probabilities, and inverting the resulting expression. □

The threshold (3.4.10) can be also applied when the cluster size are a mixture random variable. Propositions 5 and 6 show that the Pareto component with the heaviest tail dominates the upper tail distribution.

3.5 Summary

In this chapter, a fixed sample size detection system was presented. The model is developed based on the assumption that network traffic follows a marked Poisson process. The statistic of interest is the percentage of unknown source IP addresses in a given (fixed length) observation window. The detection problem is formulated as a binary hypothesis problem with the objective of discriminating between two compound Poisson processes.

The distribution of the compound random variable is approximated either by a stable distribution or using saddlepoint approximations. The latter technique is remarkably accurate and relatively easy to implement while previous methods tend to be very involved. Compound Poisson Pareto random variables can be found in other areas beside network traffic and DDoS detection. Additional applications include finance and option pricing where stock returns can be modeled with heavy tailed distributions. Another application is found in insurance where the total insurance claim amount can be modeled as a compound random variable with heavy tail claim sizes. Therefore, the tools presented here can apply in these areas as well.

The proposed mechanism is developed to detect stealthy DDoS attacks. For instance, the detection model takes into account the possibility that attack data packets may be modified so that they seem to originate from known source IP addresses. Moreover, by allowing the arrival of packet flows to be an inhomogeneous Poisson process, the detection system addresses the case where DDoS attacks have a dynamic rate.

By allowing observation windows lengths to vary, as opposed to having a predetermined size, the detection accuracy and delay can be improved. The next chapter presents such detection models where the sample size is a random variable.

CHAPTER 4

SEQUENTIAL DETECTION MODEL

This section models the DDoS detection problem as a sequential hypothesis testing problem. As opposed to the previous section, the sample size is stochastic. The arrival of packet flows is again modeled as a marked Poisson process (T_j, K_j, B_j) where T_j represents the arrival time of the first data packet in each flow j, the mark B_j denotes whether flow j has a known or unknown source IP address, and the mark K_j denotes the cluster size of flow j, i.e. the total number of data packets in the flow. Moreover, the alternative hypothesis is formulated as a composite hypothesis given that the rate of the attack packet flows is actually unknown,. We then seek a uniformly most powerful test (UMP), which is a generalization of Neyman-Pearson criterion [103]. UMP tests indicate that the NP threshold is valid regardless of the value of the parameter of the composite hypothesis.

4.1 Poisson Processes and Uniformly Most Powerful Tests

Assume Y_1, Y_2, ..., Y_n are iid samples from an homogeneous Poisson process with rate θ. The binary hypothesis testing to distinguish two Poisson processes can be formulated as

$$\begin{cases} H_0 : Y_i \sim \text{Poisson intensity } \theta_0 \\ H_1 : Y_i \sim \text{Poisson intensity } \theta_1,\ \theta_1 > \theta_0. \end{cases} \tag{4.1.1}$$

H_0 is a simple hypothesis while H_1 is composite and for any value of θ_1, we seek an UMP test. If the distribution of the composite hypothesis parameter is known, the binary test can be solved using a Bayesian approach with uniform or non-uniform costs [88]. The objective is then to minimize the posterior cost. In reality, prior distributions of hypothesis are generally unknown, rendering Bayesian constraints unpractical [115]. An alternative to Bayesian methods is the UMP test. Unlike the Bayesian case, UMP tests don't assume any distribution function for θ_1. Note that UMP tests don't always exist, especially when the composite hypothesis is two-sided. Next we prove the existence of an UMP test for our detection problem (4.1.1). Assume the random variables $\{Y_i\}$ are collected for successive non-overlapping intervals, each of length h. Then $\{Y_i\}$ are iid Poisson random variables with mean $h\theta$ and the likelihood ratio corresponding to the test (4.1.1) can be

written as

$$\mathbb{L}(y) = \prod_{i=1}^{n} \frac{\mathbb{P}\{Y_i = y_i | H_1\}}{\mathbb{P}\{Y_i = y_i | H_0\}} \tag{4.1.2}$$

The likelihood test statistic always depends on observations through a sufficient statistic [88]. So whenever possible, a sufficient statistic should be derived from the likelihood ratio as shown in Lemma 11 [103]. The derivation requires the following Factorization theorem.

Lemma 10 (Factorization Theorem [37]). *$S(Y)$ is a sufficient statistic for θ if and only if there exist functions g_θ and h such that*

$$\mathbb{P}\{Y_i = y_i, i = 1, 2, \ldots, n | \theta\} = g_\theta\left(S(y_i)\right) h(y_i)$$

Lemma 11. *Let θ be the parameter of a Poisson distribution. $S(Y) = \sum_{i=1}^{n} Y_i$ is a sufficient statistic for θ.*

Proof.

$$\mathbb{P}\{Y_i = y_i, i = 1, 2, \ldots, n | \theta\} \quad = \quad \prod_{i=1}^{n} e^{-\theta h} \frac{(\theta h)^{y_i}}{y_i!} = \frac{1}{y_1! y_2! \ldots y_n!} e^{-n\theta h} (\theta h)^{\sum_{i=1}^{n} y_i}$$

(4.1.3) factors as $h(y_1, y_2, \ldots, y_n) g(S(Y), \theta)$, where

$$h(y_1, y_2, \ldots, y_n) = \frac{1}{y_1! y_2! \ldots y_n!} \text{ and } g(S(Y), \theta) = e^{-n\theta h} (\theta h)^{S(Y)}$$

By the Factorization Theorem (see Lemma 10), $S(Y)$ is a sufficient statistic. □

Given a sufficient statistic, the Karlin-Rubin Theorem states conditions for the existence of an UMP test.

Lemma 12 (Karlin-Rubin Theorem [95]). *Let S be a scalar test statistic whose density, under both hypothesis, is parameterized by a scalar parameter θ. Assume the likelihood ratio statistic $\frac{\mathbb{P}\{S|\theta_1\}}{\mathbb{P}\{S|\theta_0\}}$ is a monotone non-decreasing function of S for every pair (θ_0, θ_1), with $\theta_1 > \theta_0$. For all $(\theta_0, \theta_1 > \theta_0)$, S has a monotone likelihood ratio and the threshold test $S \underset{H_1}{\overset{H_0}{\lessgtr}} \eta$ maximizes the probability of detection for a given probability of false alarm. $S \underset{H_1}{\overset{H_0}{\lessgtr}} \eta$ is the uniformly most powerful test among all tests with the same probability of false alarm.*

Theorem 8. *The likelihood ratio in (4.1.2) is a non-decreasing function of the sufficient statistic $S(Y) = \sum_{i=1}^{n} Y_i$. The UMP test for the binary hypothesis test described in (4.1.1) can be formulated as*

$$S(Y) \underset{H_1}{\overset{H_0}{\lessgtr}} \frac{\ln \eta + nh(\theta_1 - \theta_0)}{\ln \theta_1 - \ln \theta_0} = \kappa \tag{4.1.3}$$

According to the Neyman-Pearson criterion [103], the threshold κ is selected such that

$$\mathbb{P}\left\{S(Y) > \kappa | H_0\right\} = \varphi \tag{4.1.4}$$

Proof. Define

$$\mathbb{L}_i(y) = \frac{\mathbb{P}\left\{Y_i = y_i | H_1\right\}}{\mathbb{P}\left\{Y_i = y_i | H_0\right\}} = e^{-h(\theta_1 - \theta_0)} \left(\frac{\theta_1 h}{\theta_0 h}\right)^{y_i}$$

The likelihood ratio is a function of $S(Y)$ and is given by

$$\mathbb{L}(y) = \prod_{i=1}^{n} \mathbb{L}_i(y) = \frac{g(S(Y), \theta_1 h)}{g(S(Y), \theta_0 h)} = e^{-nh(\theta_1 - \theta_0)} \left(\frac{\theta_1}{\theta_0}\right)^{S(Y)}$$

For a threshold η, the Neyman Pearson test is

$$\ln \mathbb{L}(y) = \sum_{i=1}^{n} \ln \mathbb{L}_i(y) \underset{H_1}{\overset{H_0}{\lessgtr}} \ln \eta \ ,$$

which is equivalent to the test with the sufficient statistic in (4.1.3). The Karlin-Rubin Theorem indicates that the test (4.1.3) is UMP for the binary hypothesis test formulated in (4.1.1). □

Theorem 8 demonstrates that the hypothesis test (4.1.3) is valid regardless of the value of the composite hypothesis value, namely θ_1. Now, we must determine the necessary thresholds to perform Neyman-Pearson tests. Unlike the previous chapter which considers fixed sample size hypothesis test, the next section describes a sequential HT procedure and the corresponding thresholds.

4.2 Overview Sequential Hypothesis Testing

Internet packet flows arrive sequentially and continuously at the control router. In this case, a sequential HT may be more indicated than fixed-sample testing. Moreover, multiple samples carry more information than a single one. We can identify two sequential decision problems. The first is the sequential hypothesis testing. The objective is to distinguish a "normal" from an "abnormal" process using as few observations as possible. The second is the sequential change-point problem, which consists of detecting a point at which an otherwise "normal" process starts to become "abnormal". Both models are presented in the next sections.

4.2.1 Wald's Sequential Probability Ratio Test (SPRT)

This section adopts the results and notations from [103]. Consider Y_1, Y_2, ..., Y_n, a sequence of iid observations. $\{Y_i\}$ are distributed as follows

$$\begin{cases} H_0 : Y_i \sim f_0(y) \\ H_1 : Y_i \sim f_1(y), \end{cases} \tag{4.2.1}$$

where H_0 and H_1 are simple hypothesis. Based on n observations, the joint likelihood ratio is defined as [88]:

$$\mathbb{L}(y) = \prod_{i=1}^{n} \frac{f_1(y_i)}{f_0(y_i)} \tag{4.2.2}$$

Wald [121] formulates the sequential probability ratio test (SPRT) based on the likelihood ratio. The SPRT requires two thresholds τ_0 and τ_1. The essential element of SPRT is to set the thresholds such that desired levels of type I and type II errors are achieved. Define the sequential decision rule $(\phi, \mathcal{D})$. Given a realized observation sequence $\{y_i, i = 1, 2 \ldots\}$, the binary variable ϕ indicates when to stop sampling in the following manner. If $\phi_n(y_1, y_2, \ldots, y_n) = 0$, then take the $(n+1)^{st}$ sample. Else, stop the sample collection and make a decision.

$$\phi_n(y_1, y_2, \ldots, y_n) = \begin{cases} 0 & \text{if } \tau_0 < \mathbb{L}(y) < \tau_1 \\ 1 & \text{otherwise} \end{cases} \tag{4.2.3}$$

$\mathcal{D}_n(y_1, y_2, \ldots, y_n)$ constitutes the final decision in that it specifies what decision to make when the sampling procedure is stopped.

$$\mathcal{D}_n(y_1, y_2, \ldots, y_n) = \begin{cases} 1 & \text{if } \mathbb{L}(y) \geq \tau_1 \\ 0 & \text{if } \mathbb{L}(y) \leq \tau_0 \end{cases} \tag{4.2.4}$$

When $\mathcal{D}_n(y_1, y_2, \ldots, y_n) = 0$, an additional sample must be collected to update the likelihood ratio in (4.2.2). Else if $\mathcal{D}_n(y_1, y_2, \ldots, y_n) = 1$, the sampling procedure is stopped, and a decision must be made. Given a decision rule with thresholds (τ_0, τ_1), Wald's sequential decision test [121] can be summarized as follows. If

$$\mathbb{L}(y) \begin{cases} \leq \tau_0 & \text{declare } H_0 \text{ and stop testing} \\ \geq \tau_1 & \text{declare } H_1 \text{ and stop testing} \\ \in (\tau_0, \tau_1) & \text{take an additional sample } Y_{n+1} \end{cases} \tag{4.2.5}$$

In other words, the observation space is partitioned into three mutually exclusive regions, each corresponding to a decision of collecting additional samples, raising an alarm to signal an attack,

or declaring that the sample is legitimate. The boundaries, τ_0 and τ_1 in (4.2.5), must satisfy $0 < \tau_0 \leq 1 \leq \tau_1 < \infty$. The decision is arbitrary if $\tau_0 = \tau_1$ [103].

Define the random stopping time as $\mathcal{T} = \min\{n|\phi_n(y_1, y_2, \ldots, y_n) = 1\}$. Given $\{\mathcal{T} \leq t\}$, $\mathcal{T}$ depends only on Y_t but not on Y_{t+1}, Y_{t+2}, ... Therefore, $\mathcal{T}$ is a Markov stopping time with respect to the sequence of iid random variables $\{Y_i\}$ [115]. Wald [121] proves that the sequential likelihood ratio test eventually terminates with probability one. Sequential HT implies that the sample size, i.e. the stopping time $\mathcal{T}$, is stochastic. The desired performance is fixed and the number of samples varies accordingly. Let P_M and P_{FA} be the probabilities of miss and false alarm (FA), respectively. Given any decision rule, they are defined as

$$\begin{aligned} P_{FA}(\phi, \mathcal{D}) &= \mathbb{P}\{\mathcal{D}_\tau(y_1, y_2, \ldots, y_\tau) = 1|H_0\} \\ P_M(\phi, \mathcal{D}) &= \mathbb{P}\{\mathcal{D}_\tau(y_1, y_2, \ldots, y_\tau) = 0|H_1\} \end{aligned} \tag{4.2.6}$$

A decision rule with thresholds (τ_0, τ_1) generates the following error probabilities

$$\begin{aligned} \varphi &= \mathbb{P}\{\mathbb{L}(y) \geq \tau_1|H_0\} \\ \xi &= \mathbb{P}\{\mathbb{L}(y) \leq \tau_0|H_1\} \end{aligned}$$

The performance of a sequential test is measured in terms of the rate of false alarms and false clears. A decision can be reached with few samples or additional samples may be required to make an educated decision. The duration of the sequential test depends on the signal-to-noise ratio. Another quantity of interest in the sequential hypothesis testing is the expected stopping time, that is the expected number of samples that must be collected before a decision can be made. Wald's sequential probability ratio test minimizes the expected number of collected samples given that probability errors do not exceed pre-determined bounds. For the same performance, the Wald-Wolfowitz Theorem, stated below, states the optimality properties of the sequential probability ratio test.

Theorem 9 (Wald-Wolfowitz Theorem [103])**.** *Let $(\phi_0, \mathcal{D}_0)$ be the sequential likelihood ratio test described in (4.2.5), namely SQRT(τ_0,τ_1). Consider any other sequential decision rule $(\phi, \mathcal{D})$ such that*

$$\begin{aligned} P_{FA}(\phi, \mathcal{D}) &\leq P_{FA}(\phi_0, \mathcal{D}_0) \\ P_M(\phi, \mathcal{D}) &\leq P_M(\phi_0, \mathcal{D}_0). \end{aligned}$$

Then

$$\mathbb{E}[\mathcal{T}(\phi)|H_k] \geq \mathbb{E}[\mathcal{T}(\phi_0)|H_k], \tag{4.2.7}$$

with $k = 0, 1$.

The Wald-Wolfowitz Theorem states that for given P_{FA} and P_M, SQRT(τ_0,τ_1) is the detector with the smallest average sample size. Sequential test allows us to arbitrarily select the error probabilities P_F and P_M, which a great advantage. The question is how to determine the boundaries τ_0 and τ_1.

Let φ and ξ be the desired limits on type I and type II error probabilities, respectively. Wald [121] proves that the following inequalities hold

$$\tau_1 \leq \frac{(1-\xi)}{\varphi} \text{ and } \tau_0 \geq \frac{\xi}{(1-\varphi)}, \tag{4.2.8}$$

Equation (4.2.8) also holds for the case of non-iid random variables [121]. In the non-iid case however, the likelihood ratio is no longer the product of ratio of pdfs. Note that as the error probabilities φ and ξ decrease, τ_0 gets smaller while τ_1 gets larger. This implies that the smaller the bounds on the error probabilities, the more samples are required to make a decision [103]. Using (4.2.8), the boundaries τ_0 and τ_1 can be approximated to obtain the desired errors φ and ξ. Assume the excess over boundaries defined as [121]

$$\mathbb{L}(y) - \tau_1 \text{ and } \tau_0 - \mathbb{L}(y), \tag{4.2.9}$$

are negligible. Then, the inequalities (4.2.8) are exact, i.e.,

$$\tau_1 = \frac{(1-\xi)}{\varphi} \text{ and } \tau_0 = \frac{\xi}{(1-\varphi)}, \tag{4.2.10}$$

Equalities (4.2.10) are known as *Wald's approximations*. If φ^* and ξ^* are the fixed desired error probabilities then using (4.2.10) yields the following boundaries

$$\tau_1^* = \frac{(1-\xi^*)}{\varphi^*} \text{ and } \tau_0^* = \frac{\xi^*}{(1-\varphi^*)}. \tag{4.2.11}$$

Using Wald's approximations instead of the true decision boundaries increases the expected number of samples that need to be collected before a decision is reached. Observe that the boundaries do not depend on the distribution of the $\{Y_i\}$, which is an advantage of sequential over fixed sample size hypothesis testing [103]. Furthermore, the boundaries can be easily modified if instead of the likelihood ratio in (4.2.2), one chooses to use the log-likelihood ratio. The equivalent decision boundaries are $\ln \tau_0$ and $\ln \tau_1$. Note however that the expected sample size (i.e., expected stopping time), which is a performance indicator, depends on the distribution of observations. Wald's Fundamental Identity can be applied to extend these results to general sequential hypothesis testing problems with random variable of the form $g(Y_i)$, where $g : \mathbb{R} \to \mathbb{R}$.

Theorem 10 (Wald's Fundamental Identity [121])**.** *Consider the decision rule* $(\phi, \mathcal{D})$*. Let* $\mathcal{T} = \min\{n | \phi_n(y_1, y_2, \ldots, y_n) = 1\}$ *and* $S_n = \sum_{i=1}^{n} g(Y_i)$*.*

Define $\mathcal{M}_k(\theta) = \mathbb{E}\left[\exp\left\{\theta g(Y_i)\right\}|H_k\right]$ *as the moment generating function of the random variable* $g(Y_i)$ *given the hypothesis* H_k, $k = 0, 1$. *If* $\mathbb{P}\left\{g(Y_i) = 0|H_k\right\} \neq 1$ *and* $\mathbb{P}\left\{|g(Y_i)| < \infty|H_k\right\} = 1$ *then*

$$\mathbb{E}\left[e^{\theta S_{\mathcal{T}}}\left\{\mathcal{M}_k(\theta)\right\}^{-\mathcal{T}}|H_k\right] = 1, \tag{4.2.12}$$

for $\theta \in \mathbb{R}$ *such that* $\mathcal{M}(\theta) < \infty$.

Corollary 1. *Suppose* $\mathcal{M}_k(\theta) < \infty$ *near* $\theta = 0$. *Let* $S_n = \sum_{i=1}^{n} g(Y_i)$, *with* $g : \mathbb{R} \to \mathbb{R}$. *Define* $\mu_k = \mathbb{E}\left[g((Y_i)|H_k\right]$ *and* $\sigma_k^2 =$ *Var*$\left[g((Y_i)|H_k\right]$, *for* $k = 0, 1$. *Then*

$$\begin{aligned} \mathbb{E}\left[S_{\mathcal{T}}|H_k\right] &= \mu_k \mathbb{E}\left[\mathcal{T}|H_k\right] \\ \mathbb{E}\left[\left(S_{\mathcal{T}} - \mathcal{T}\mu_k\right)^2|H_k\right] &= \sigma_k^2 \mathbb{E}\left[\mathcal{T}|H_k\right], \end{aligned}$$

where $\mathcal{T}$ *is the stopping time.*

Later, we apply Wald's Fundamental Identity to extend Wald's approximations to sequential test with statistic of the form $S_n = \sum_{i=1}^{n} g(Y_i)$, $g : \mathbb{R} \to \mathbb{R}$. Corollary 1 allows computing the expected sample size for the test with statistic S_n. In applying the sequential hypothesis test in our DDoS detection problem, we must emphasize that one SPRT is not enough. In fact, network traffic must be monitored continuously, therefore, successive applications of Wald's tests are necessary [135]. In other words, when the test ends on the lower boundary, i.e, $\mathbb{L}(y) \leq \tau_0$ and H_0 is accepted, current samples can be discarded and a new sequential test started. Of course, when the test ends on the upper boundary, i.e. $\mathbb{L}(y) \geq \tau_1$, the testing procedure is stopped and an alarm is triggered. Sequential change-point detectors are developed based on this idea of continuous monitoring. The next section presents two change-point detection procedures.

4.2.2 Change-Point Detection

The change-point detection is similar to applying a sequence of Wald's sequential tests. Again consider $\{Y_i, i \geq 1\}$, a sequence of iid observations. Let $f_0(y_1, y_2, \ldots, y_n)$ be the joint pdf of the observed random variables until an event occurs. In our case, an event would be the start of the DDoS attack. After the change, which happens at an unknown point in time $\mathcal{T}$, the distribution becomes $f_1(y_1, y_2, \ldots, y_n)$. A binary hypothesis test can be formulated based on the stopping time $\mathcal{T}$, i.e.,

$$\begin{cases} H_0 : \text{change at } 0 \leq \mathcal{T} < n \\ H_1 : \text{no change for } \mathcal{T} \geq n \end{cases} \tag{4.2.13}$$

Then the likelihood ratio is

$$\mathbb{L}(y) = \frac{\prod_{i=1}^{\mathcal{T}} f_0(y_i) \prod_{i=\mathcal{T}+1}^{n} f_1(y_i)}{\prod_{i=1}^{n} f_0(y)} = \frac{\prod_{i=\mathcal{T}+1}^{n} f_1(y_i)}{\prod_{i=\mathcal{T}+1}^{n} f_0(y)} \tag{4.2.14}$$

A sequential change-point detection procedure consists of identifying the optimal stopping time $\mathcal{T}$, i.e. the time at which an alarm is triggered to indicate a change in distribution ($f_0(y)$ to $f_1(y)$). The performance of a change point detector is measured in terms of the rate of false alarms and the detection delay. The latter corresponds to the expected number of samples that must be collected before the likelihood ratio exceeds a preset threshold.

Page's CUSUM [98] algorithm corresponds to applying a sequence of Wald sequential tests with log-boundaries fixed at 0 and Υ [74]. For $n \geq \mathcal{T}$, the Page's test statistic is

$$\mathcal{R}_{n,\mathcal{T}} = \sum_{k=\mathcal{T}}^{n} \log \frac{f_1(Y_i)}{f_0(Y_i)} \,. \tag{4.2.15}$$

The stopping time $\mathcal{T}$ is defined as

$$\mathcal{T} = \min \{n \geq 1 \,|\, \mathcal{U}_n \geq \Upsilon\} \,, \tag{4.2.16}$$

where

$$\mathcal{U}_n = \max_{0 \leq \tau \leq n} \mathcal{R}_{n,\mathcal{T}}.$$

Since observations $\{Y_i, i \geq 1\}$ are iid, the test statistic $\mathcal{U}_n$ can be computed recursively as

$$\mathcal{U}_n = \begin{cases} 0 & \text{if } n = 0 \\ \max\left\{0, \mathcal{U}_{n-1} + \log \frac{f_1(Y_n)}{f_0(Y_n)}\right\} & \text{if } n \geq 1 \end{cases}$$

Basically, the procedure compares the maximum log-likelihood ratio to a threshold. Samples are continuously collected to update the likelihood ratio until reaching a decision in favor of H_1 , i.e. $\mathcal{U}_n \geq \ln \tau_1$. Then the testing procedure is stopped and an alarm is triggered. The objective is to minimize the worst average detection delay. Let $\mathcal{T}^*$ denote the time required for a single SPRT to stop. Then

$$\begin{aligned} \mathbb{E}\left[\mathcal{T}|H_0\right] &= \frac{1}{\varphi}\mathbb{E}\left[\mathcal{T}^*|H_0\right] \\ \mathbb{E}\left[\mathcal{T}|H_1\right] &= \frac{1}{1-\xi}\mathbb{E}\left[\mathcal{T}^*|H_1\right], \end{aligned}$$

where φ and ξ represent the type I and type II error probabilities, respectively.

An alternative change-point procedure is due to Shiryaev, Roberts and Pollack [116]. Let $\mathcal{R}_{n,\mathcal{T}}$ be defined as in (4.2.15). The test statistic is given as

$$\mathcal{R}_{SP} = \sum_{\mathcal{T}=1}^{n} e^{\mathcal{R}_{n,\mathcal{T}}} = \sum_{\mathcal{T}=1}^{n} \prod_{i=\mathcal{T}}^{n} \frac{f_1(Y_i)}{f_0(Y_i)}$$

Assuming that observations are iid, $\mathcal{R}_n$ satisfies the recursion

$$\mathcal{R}_{n+1} = (\mathcal{R}_n + 1)\,\frac{f_1(Y_{n+1})}{f_0(Y_{n+1})}$$

The stopping time $\mathcal{T}$ must satisfy

$$\mathcal{T} = \min\{n \geq 1 | \mathcal{R}_n \geq \Upsilon\} \tag{4.2.17}$$

Both change-point detection algorithms seek to minimize the average detection delay for a fixed average time between false alarms T_{FA}. The threshold Υ can be set as $\Upsilon = \ln T_{FA}$.

One difference between Wald's sequential likelihood ratio test and sequential change-point is that the levels of false alarm and detection rates are prefixed and the decision boundaries must be selected accordingly. Whereas, with the latter, the rate of false alarm is a performance quantity to be minimized along with the detection delay. Another difference is the number of thresholds. Wald's sequential test uses a lower and upper threshols while the change-point algorithms utilizes only an upper bound, taking an action only when the test statistic exceeds the upper threshold. Sequential HT requires that the distribution function be known under both hypothesis, which may be a disadvantage. The next sections demonstrate the application of sequential HT in different instances of our DDoS detection problem. First, we briefly address the sampling procedure.

4.2.3 Sampling Procedure

As in the precedent chapter, we assume that the process of packet flows is a marked Poisson process. When $\{Y_i, i \geq 1\}$, the sequence of observations is distributed according to a Poisson process, we can identify two sampling techniques. The first is the time-discrete sampling. Kazakos and Papantoni-Kazakos [65] propose a time-discrete sequential HT where observations are collected for each successive non-overlapping time intervals of length h. The second sampling can be viewed as the "standard" sequential HT where samples are collected individually, not by time intervals. Application to our DDoS detection problem shows that one observation with the time-discrete sampling corresponds to a set of flows (more than one flow) that arrive in the interval h. One the contrary, the standard sampling implies that we collect one flow after another. The latter method is particularly useful when dealing with the cluster size of each flow. Computations can be simplified if we consider that packet flows are independent and collect one flow at the time. Both methods are illustrated in subsequent sections.

4.3 Detection Model Based on Distinct Source IP Addresses

Let $\{U(t), t \geq 0\}$ and $\{A(t), t \geq 0\}$ be the observed process of legitimate and attack packet flows. Assume both processes are homogeneous Poisson processes. This section discusses a simpler version of the detection problem where we monitor the percentage of *distinct* number of unknown source IP addresses. As mentioned in the previous chapter, all data packets in a flow share the same source IP address, therefore the flow cluster size is irrelevant for this detection model. The marked Poisson process described in Section 2.1 simplifies to (T_j, B_j) where the mark B_j is a Bernoulli random variable and T_j represents the arrival time of the first data packet in each flow. Recall the Bernoulli mark B_j indicates the probability of that the packet flow j has an unknown source IP address. B_j has parameter ε for legitimate and ϱ for attack traffic, respectively. Since we don't consider cluster size the time-discrete sampling is the most adequate sampling method.

Consider that an observation time interval of length T is partitioned into n intervals $I_i = [(i-1)h, ih]$, $i = 1, 2, \ldots, n$. The percentage of unknown source IP addresses is computed for each interval of length h. Let $Y_1, Y_2, \ldots, Y_n$ be the number of observed data packet with unknown IP address in I_i. From the previous chapter, we concluded that under H_1, the independent legitimate and attack Poisson processes were additive. Applying the thinning property of Poisson processes using the Bernoulli mark separated the initial packet flow process into two independent Poisson processes: one with unknown and the other with known source IP addresses; we focus on the former. Let $L = \zeta\varepsilon + \lambda\varrho$. The binary hypothesis testing to detect whether a DDoS attack is taking place or not can be formulated as follows.

$$\begin{cases} H_0 : Y_i \sim \text{Poisson intensity } \zeta\varepsilon h \\ H_1 : Y_i \sim \text{Poisson intensity } Lh,\ L > \zeta\varepsilon \end{cases} \tag{4.3.1}$$

H_0 is modeled as a simple hypothesis and represents the hypothesis that only legitimate packet flows are observed. H_1, on the other hand, denotes the hypothesis that an attack is taking place in addition to the legitimate traffic. H_1 is formulated as a composite hypothesis since the rate of the attack flow is unknown. Observations Y_i's are iid Poisson random variables with parameter Lh and $\zeta\varepsilon h$, under H_1 and H_0, respectively. The objective of the detection model is to provide an UMP test for (4.3.1) and apply a sequential detection procedure to distinguish the two Poisson processes.

4.3.1 UMP Test and Pdf Sufficient Statistic

For any value of $L > \zeta\varepsilon$, we can apply Theorem 8 to prove that a uniformly most powerful (UMP) test exists for (4.3.1). The log-likelihood ratio of the binary hypothesis test defined in (4.3.1) can be written as

$$\ln \mathbb{L}(y) = -nh\,(L - \zeta\varepsilon) + S(Y) \ln\left(\frac{L}{\zeta\varepsilon}\right), \tag{4.3.2}$$

where $S(Y) = \sum_{i=1}^{n} Y_i$. Lemma 11 shows that $S(Y)$ is a sufficient statistic for the Poisson parameter $h\zeta\varepsilon$ or Lh. We can apply Theorem 8 to prove that an UMP test exists for the binary test above. The proof basically reduces to proving that the value of the threshold κ in (4.1.3) depends solely on the distribution function of $S(Y)$ and on the type I error probability φ The Neyman-Pearson test can be formulated as in (4.1.3) with $\theta_1 = Lh$ and $\theta_0 = h\zeta\varepsilon$, i.e.

$$S(Y) \underset{H_1}{\overset{H_0}{\lessgtr}} \frac{\ln \eta + nh\,(L - \zeta\varepsilon)}{\ln L - \ln \zeta\varepsilon} = \kappa \tag{4.3.3}$$

The threshold κ effectively depends on $\bar{F}_{S|H_0}(y)$ and on the value of φ.

Therefore, (4.3.3) is an UMP test for the binary hypothesis test formulated in (4.3.1). Next, we briefly address the steps to compute the probability density of the sufficient statistic $S(Y)$ under H_0 (see (4.1.4)).

Given that Y_i's are iid Poisson with rate $h\zeta\varepsilon$ under H_0, $S(Y)$ is also Poisson distributed with rate $nh\zeta\varepsilon$ [109]. This follows since

$$\psi_S(\theta) = (\psi_Y(\theta))^n = \left(e^{h\zeta\varepsilon\left(e^{i\theta}-1\right)}\right)^n = e^{nh\zeta\varepsilon\left(e^{i\theta}-1\right)}$$

In the equation above, $\psi_S(\theta)$ represents the characteristic function of the sufficient statistic. Therefore, we deduce that the value of the threshold κ in (4.1.3) depends only on the fixed error probability φ and the rate $nh\zeta\varepsilon$. κ is the solution of $F_S(\kappa) = 1 - \varphi$, where

$$F_S(\kappa) = \frac{\Gamma(\lfloor \kappa + 1 \rfloor, nh\zeta\varepsilon)}{\lfloor \kappa \rfloor !}$$

is the cdf of $S(Y)$ and $\Gamma(a, z)$ is the incomplete gamma function. $F_S(\kappa)$ can not be easily inverted so an approximation must be found. The central limit theorem (CLT) indicates that for a large n, the distribution of $S(Y)/n$ can be approximated by a normal distribution.

$$\frac{S(Y)}{n} \sim \mathcal{N}\left(h\zeta\varepsilon, \frac{h\zeta\varepsilon}{n}\right) \tag{4.3.4}$$

$$\mathbb{P}\left\{S(Y) > \kappa | H_0\right\} = \varphi \Rightarrow \mathbb{P}\left\{\frac{S(Y)/n - h\zeta\varepsilon}{\sqrt{h\zeta\varepsilon/n}} > \frac{\kappa - h\zeta\varepsilon}{\sqrt{h\zeta\varepsilon/n}}\right\} = \varphi$$

This is equivalent to

$$\Phi\left(\frac{\kappa - h\zeta\varepsilon}{\sqrt{h\zeta\varepsilon/n}}\right) = 1 - \varphi \Rightarrow z_\varphi = \frac{\kappa - h\zeta\varepsilon}{\sqrt{h\zeta\varepsilon/n}}\,,$$

where Φ is the cdf of the standard normal distribution. In that case, reject H_0 if

$$\frac{S(Y)}{n} \geq h\zeta\varepsilon + z_\varphi \sqrt{\frac{h\zeta\varepsilon}{n}} \tag{4.3.5}$$

Suppose that $S(Y)$ is continuously sufficient. In other words, the Factorization Theorem 10 and subsequently Neyman-Pearson test (see Theorem 8) applies at every step of the sequential test [44]. Then Wald's sequential test, described in (4.2.5), can be reformulated in terms of the sufficient statistic as

$$\begin{cases} H_0 & \text{if } S(Y) \leq \kappa_0 \\ H_1 & \text{if } S(Y) \geq \kappa_1 \end{cases} \tag{4.3.6}$$

Else, if $\kappa_0 < S(Y) < \kappa_1$, update $S(Y)$ with sample Y_{n+1}. Denote $\mathcal{S}(\kappa_0, \kappa_1)$, the sequential hypothesis test with decision boundaries κ_0 and κ_1, with $\kappa_0 < \kappa_1$. Ferguson [37] emphasizes that the use of sufficient statistics reduces the complexity of the hypothesis testing problem. Furthermore, Wald [121] argues that the SPRT as defined in (4.2.5) satisfies the requirements for composite hypothesis testing with one-sided alternative hypothesis. The following section shows how to compute the bounds κ_0 and κ_1.

4.3.2 Decision Boundaries SPRT

At the stopping time of any sequential hypothesis test, the likelihood ratio $\mathbb{L}(y)$ crosses the decision boundaries in three possible scenarios: no excess, excess over one decision boundary, and finally, excess over both decision boundaries. With time continuous random processes, as the Poisson process in our case, the second or first scenario is the most likely [44]. We first address the case of no excess over boundaries in Proposition 9.

Proposition 9. *Assume the excess over boundaries (4.2.9) are negligible. The decision boundaries in the test (4.3.6) can be defined in terms of Wald's approximations in (4.2.10).*

$$\kappa_0 = \frac{\ln \tau_0^*}{\ln L - \ln \zeta\varepsilon} + rh, \quad \textit{and} \quad \kappa_1 = \frac{\ln \tau_1^*}{\ln L - \ln \zeta\varepsilon} + rh, \tag{4.3.7}$$

where

$$r = \frac{(L - \zeta\varepsilon)}{\ln L - \ln \zeta\varepsilon}$$

Proof. Recall Wald's approximations are

$$\tau_0^* = \frac{\xi}{1-\varphi}, \text{ and } \tau_1^* = \frac{1-\xi}{\varphi}.$$

Equation (4.3.7) follows since

$$\begin{aligned}\mathbb{L}(y) \geq \tau_1 &\equiv \ln \mathbb{L}(y) \geq \ln \tau_1^* \\ &\equiv -nh\left(L - \zeta\varepsilon\right) + S(Y) \ln\left(\frac{L}{\zeta\varepsilon}\right) \geq \ln \tau_1^* \\ &\equiv S(Y) \geq \frac{\ln \tau_1^* + nh\left(L - \zeta\varepsilon\right)}{\ln L - \ln \zeta\varepsilon} = \kappa_1\end{aligned}$$

Deriving κ_0 is similar. □

Taking into account excess over boundaries yields more accurate decision boundaries. These can be determined from the distribution of the log-likelihood ratio statistic [65, 103]. Define the test statistic $S_n = \ln \mathbb{L}(y) = \sum_{i=1}^{n} Z_i$. Based on the log-likelihood ratio defined in (4.3.2), the random variables $\{Z_i\}$ are iid and computed as

$$Z_i = Y_i \ln \frac{L}{\zeta\varepsilon} + h\left(L - \zeta\varepsilon\right) \tag{4.3.8}$$

Evaluating decision boundaries requires the Operating Characteristic (OC) function while evaluating the expected stopping time depends on the Average Sample Number (ASN) function. These functions depend on the distribution of observations and are defined next; also, see [44, 121] for more details. Let θ be the parameter of the distribution function in the hypothesis testing problem. In our detection problem, θ would be the rate of the Poisson process, in particular the rate of packet flows with unknown source IP addresses.

Definition 11. *The OC function $\mathcal{Q}(\theta)$ is the probability that the test stops and the final decision is H_0. More specifically,*

$$\mathcal{Q}(\theta) = \mathbb{P}\left\{S_n \leq \kappa_0 | \theta\right\} \tag{4.3.9}$$

Recall $\zeta\varepsilon$ and L are the average numbers of distinct unknown source IP addresses under H_0 and H_1, respectively. Moreover, φ and ξ are the desired levels of type I and type II error probabilities. Then, solving $\mathcal{Q}(\theta = \zeta\varepsilon) = 1 - \varphi$ and $\mathcal{Q}(\theta = L) = \xi$ yields the upper and lower decision boundaries, i.e.κ_0 and κ_1 [44, 121].

Definition 12. *The expected number of samples $\mathbb{E}\left[n; \theta\right]$ function is*

$$\mathbb{E}\left[n; \theta\right] = \sum_{n=1}^{n} n\, p(n; \theta), \tag{4.3.10}$$

where

$$p(n;\theta) = \begin{cases} \mathbb{P}\{S_n \leq \kappa_0|\theta\} + \mathbb{P}\{S_n \geq \kappa_1|\theta\} & \text{if } n = 1 \\ \mathbb{P}\{\kappa_0 < S_1 < \kappa_1, \ldots, \kappa_0 < S_{n-1} < \kappa_1, S_n \leq \kappa_0|\theta\} + \mathbb{P}\{S_n \geq \kappa_1|\theta\} & \\ +\mathbb{P}\{\kappa_0 < S_1 < \kappa_1, \ldots, \kappa_0 < S_{n-1} < \kappa_1, S_n \geq \kappa_1|\theta\} + \mathbb{P}\{S_n \geq \kappa_1|\theta\} & \text{if } n > 1 \end{cases}$$

For iid $\{Z_i\}$, Corollary 1 implies that $\mathbb{E}[n;\theta]$ can be written as

$$\mathbb{E}[n;\theta] = \frac{\mathcal{Q}(\theta)\mathbb{E}[S_n|S_n \leq \ln\tau_0;\theta] + (1-\mathcal{Q}(\theta))\mathbb{E}[S_n|S_n \geq \ln\tau_1;\theta]}{\mathbb{E}[Z_i|\theta]}, \tag{4.3.11}$$

If $\mathbb{E}[S_i|\theta] = 0$, then

$$\mathbb{E}[n;\theta] = \frac{\mathcal{Q}(\theta)\mathbb{E}\left[S_n^2|S_n \leq \ln\tau_0;\theta\right] + (1-\mathcal{Q}(\theta))\mathbb{E}\left[S_n^2|S_n \geq \ln\tau_1;\theta\right]}{\mathbb{E}\left[Z_i^2|\theta\right]}, \tag{4.3.12}$$

If $\mathcal{T}$ is the stopping time then the final decision depends on the test statistic $S_{\mathcal{T}} = \sum_{i=1}^{\mathcal{T}} Z_i$. Given that both $\mathcal{T}$ and Z_i are random variables, Corollary 1 indicates that $\mathbb{E}[S_{\mathcal{T}}] = \mathbb{E}[\mathcal{T}]\mathbb{E}[Z_i]$. Using Wald's approximations τ_0^* and τ_1^* and (4.3.11), the expected values of $\mathcal{T}$, under the null and alternative hypothesis, respectively are

$$\begin{aligned} \mathbb{E}[\mathcal{T}|H_0] &= \frac{\varphi\ln\tau_1^* + (1-\varphi)\ln\tau_0^*}{\mathbb{E}[Z_i|H_0]} \\ \mathbb{E}[\mathcal{T}|H_1] &= \frac{(1-\xi)\ln\tau_1^* + \xi\ln\tau_0^*}{\mathbb{E}[Z_i|H_1]} \end{aligned} \tag{4.3.13}$$

(4.3.13) assumes no excess over boundaries.

Excess Over One Boundary

Assume the excess scenario two occurs, in particular $\tau_0 - \mathbb{L}(y) \approx 0$ and $\mathbb{L}(y) - \tau_0 > 0$ (see (4.2.9)). Recall $\mathcal{S}(\kappa_0, \kappa_1)$ denotes a sequential hypothesis test with decision boundaries κ_0 and κ_1. Since $\tau_0 - \mathbb{L}(y) \approx 0$, the lower decision boundary κ_0 consequently remains unchanged and is computed as in (4.3.7). We address the case of computing κ_1, the upper decision boundary.

Assume the error probabilities limits satisfy $\varphi + \xi < 1$. Condition $\varphi + \xi < 1$ is easily met since the error probabilities frequently satisfy $\varphi, \xi \in [0.01, 0.05]$ [121]. Ghosh [44] derives the OC function as

$$\mathcal{Q}(\theta) = \frac{e^{\theta A/s}\sum_{j=0}^{B-1}\frac{(-1)^j}{j!}\left[(B-j)\frac{\theta}{s}e^{-\theta/s}\right]^j}{\sum_{j=0}^{B-A-1}\frac{(-1)^j}{j!}\left[(B-A-j)\frac{\theta}{s}e^{-\theta/s}\right]^j}, \tag{4.3.14}$$

where

$$A = \frac{\kappa_0}{\ln(L/\zeta\varepsilon)}, \;\; B = \frac{\kappa_1}{\ln(L/\zeta\varepsilon)}, \;\; s = \frac{L - \zeta\varepsilon}{\ln(L/\zeta\varepsilon)}$$

Furthermore, it is shown that the upper decision boundary satisfies

$$\ln \frac{1 - \mathcal{Q}^*(L)}{1 - \mathcal{Q}^*(\zeta\varepsilon)} - \ln \frac{L}{\zeta\varepsilon} \leq \tau_1^* \leq \ln \frac{1 - \mathcal{Q}^*(L)}{1 - \mathcal{Q}^*(\zeta\varepsilon)}, \tag{4.3.15}$$

where τ_1^* is given by (4.2.10). However, solving $\mathcal{Q}(\theta = \zeta\varepsilon) = 1 - \varphi$ and $\mathcal{Q}(\theta = L) = \xi$ for the exact decision boundary κ_1 is tedious. Kazakos and Papantoni-Kazakos [65] propose alternative approximations as shown in Proposition 10.

Proposition 10. *Define* $p = \frac{L}{\zeta\varepsilon} - 1$. *The upper and lower decision boundaries are*

$$\begin{aligned} e^{-\kappa_0} &= \frac{\varphi}{1 - \xi} + \frac{h(p)}{2(1 - \xi)}, \text{ and} \\ e^{\kappa_1} &= \frac{\xi}{1 - \varphi} + \frac{b(p)}{2(1 - \varphi)}, \end{aligned} \tag{4.3.16}$$

respectively, with

$$\begin{aligned} b(p) &= \frac{h(p) - 2\varphi\xi p}{1 + p}, \text{ and} \\ h(p) &= p(\varphi\xi + 1 - \xi) + 1 - \xi - \varphi - [p(\varphi\xi + 1 - \xi) + 1 - \xi - \varphi]^2 + (4p\varphi^2\xi)^{1/2}. \end{aligned}$$

4.4 Time-Discrete Sequential Test for Homogeneous Poisson with Cluster Size

As opposed to the previous section, we now consider the full marked Poisson process (T_j, K_j, B_j) and monitor the total number of data packets with unknown source IP addresses, regardless of duplicates. Recall K_j is the flow cluster size, i.e., it represents the total number of data packets in flow j. K_j has the mixture distribution described in Section 2.1. Observations are collected based on the time-discrete sampling method, i.e., for every successive, non overlapping time intervals of length h. By definition, $\{Y_i = \sum K_j\}$ are iid compound Poisson random variables and the binary hypothesis testing becomes:

$$\begin{cases} H_0 : Y_i \sim \text{Compound Poisson intensity } \zeta\varepsilon \\ H_1 : Y_i \sim \text{Compound Poisson intensity } L,\ L > \zeta\varepsilon \end{cases} \tag{4.4.1}$$

Note that the distribution of cluster size K_j has different parameters under both hypothesis. The likelihood ratio can be written as

$$\mathbb{L}(y) = \prod_{i=1}^{n} \frac{\mathbb{P}\{Y_i = y_i | H_1\}}{\mathbb{P}\{Y_i = y_i | H_0\}} \tag{4.4.2}$$

Compound random variables do not have a closed-form pdf so that we can deduce an optimal sufficient statistic from (4.4.2). In the sequential test, the likelihood ratio should then be replaced by an appropriate test statistic. The test statistic must be continuously sufficient so that a decision can be made at any time [44]. We assume that all sufficient statistics are functions of the observation data i.e. $\mathbb{S}(Y) = \sum_{i=1}^{n} g(Y_i)$, where $g : \mathbb{R} \to \mathbb{R}$. This is justifiable since all observations $\{Y_i\}$ are iid and carry the same weight, therefore they should be treated equally. The test statistic simply accumulates the scores from observations. In this section, we choose g to be the identity function, the square function Y_i^2 or the log log Y_i function are other potential test statistics. We develop the SPRT by computing and comparing $\mathbb{S}(Y) = \sum_{i=1}^{n} Y_i$ to predetermined thresholds. Using Theorem 8, Wald's sequential test described in (4.2.5) can be reformulated in terms of the sufficient statistic as

$$\begin{cases} H_0 & \text{if } \mathbb{S}(Y) \leq \tilde{\kappa}_0 \\ H_1 & \text{if } \mathbb{S}(Y) \geq \tilde{\kappa}_1 \end{cases} \tag{4.4.3}$$

Else, if $\tilde{\kappa}_0 < \mathbb{S}(Y) < \tilde{\kappa}_1$, update $\mathbb{S}(Y)$ with the new sample Y_{n+1}.

The following section considers that $\{K_j\}$ are pure Pareto random variables and designs a sequential test using $\mathbb{S}(Y)$ as the test statistic, with new decision boundaries. The mixture distribution case is addressed later.

4.4.1 Pareto Cluster Size

By definition, $\mathbb{S}(Y)$ is a sum of n iid compound poisson random variables since

$$\mathbb{S}(Y) = \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \sum_{j=1}^{\mathcal{N}(h)} K_{ij} \ , \tag{4.4.4}$$

where $\mathcal{N}(h)$ is Poisson random variable in the time interval h. Mikosch [80] proves that $\mathbb{S}(Y)$ is also a compound Poisson random variable. Assume cluster size are distributed according to a pure Pareto distribution with tail index α_k and scale parameter c_k, for $k = 0, 1$ denoting the hypothesis H_0 or H_1. Computing the thresholds in (4.3.6) solely depend on the distribution function of the test statistic $\mathbb{S}(Y)$ and the desired level of type I error probability. The proposed test is thus uniformly most powerful among all tests using the same statistic and the same limits on error probabilities. The asymptotic distribution of the test statistic $\mathbb{S}(Y)$ is given next.

Pdf Test Statistic

Y_i is a sum of $\mathcal{N}(h)$ iid Pareto random variables. Given H_0, $\mathcal{N}(h)$ is a Poisson random variable with mean $h\zeta\varepsilon$. Theorem 11 extends results from the previous chapter to show that if $0 < \alpha < 2$, the Pareto distribution has infinite variance and the limiting distribution of $\mathbb{S}(Y)$ as $n \to \infty$ is a stable distribution with the same tail index. When the tail index is greater than 2, $f_{\mathbb{S}}$, the pdf of $\mathbb{S}(Y)$, is asymptotically normal and can be approximated using the saddlepoint technique (see Section 3.2.3).

Theorem 11. *Define $\widetilde{\mathbb{S}}(Y)$ as the normalized test statistic. If $0 < \alpha < 2$ and $nh\zeta\varepsilon \to \infty$ then*

$$\widetilde{\mathbb{S}}(Y) \sim \mathcal{S}_\alpha \left(\left[\Gamma(1-\alpha) \cos\left(\frac{\pi\alpha}{2}\right) \right]^{1/\alpha}, 1, 0 \right) \tag{4.4.5}$$

with characteristic function [108, 110]

$$\psi_{\widetilde{\mathbb{S}}(Y)}(\theta) = \exp\left\{ -|\theta|^\alpha \Gamma(1-\alpha) \cos\left(\frac{\pi\alpha}{2}\right) \left[1 - i \ (sign\,\theta) \tan\left(\frac{\pi\alpha}{2}\right) \right] \right\} \tag{4.4.6}$$

and pdf

$$f_{\widetilde{\mathbb{S}}(Y)}(y) = \frac{1}{\alpha\pi} \int_0^\infty e^{-a\theta} \cos\left(\theta^{1/\alpha} y + b\theta \right) \theta^{(1/\alpha)-1} d\theta \ , \tag{4.4.7}$$

where $a = -c^\alpha \alpha\Gamma(-\alpha)\cos(\alpha\pi/2)$ and $b = c^\alpha \alpha\Gamma(-\alpha)\sin(\alpha\pi/2)$

Proof. The case of $0 < \alpha < 1$ and $1 < \alpha < 2$ must be treated separately since the normalization is different, the reason being that a mean exists in the latter case, while it is infinite in the former case. Assume $0 < \alpha < 1$ and define the normalized as

$$\widetilde{K}_j = \frac{K_j}{(n\zeta\varepsilon h)^{1/\alpha}} \ , \tag{4.4.8}$$

where K_j is a Pareto random variable, $K_j \sim \text{Par}(\alpha, 1)$, $0 < \alpha < 1$. The corresponding normalized sufficient statistic is then $\widetilde{\mathbb{S}}(Y) = \sum_{j=1}^{\tilde{n}} \sum_{j=1}^{\mathcal{N}(h)} \widetilde{K}_{ij}$. Note that $\widetilde{\mathbb{S}}(Y)$ has a one-to-one correspondence with $\mathbb{S}(Y)$ and subsequently is a sufficient statistic. Recall from Lemma 4, $\widetilde{K}_j$ is also Pareto distributed, $\widetilde{K}_j \sim \text{Par}\left(\alpha, \frac{1}{(\zeta\varepsilon h)^{1/\alpha}}\right)$. By independence of the $\widetilde{K}_j$'s,

$$\begin{aligned}
\psi_{\widetilde{\mathbb{S}}(Y)}(\theta) &= \left(\psi_{\widetilde{Y}_i}(\theta) \right)^n \\
&= \left(\exp\left\{ \zeta\varepsilon h \left[\psi_K \left(\frac{\theta}{(\zeta\varepsilon h)^{1/\alpha}} \right) - 1 \right] \right\} \right)^n \\
&= n\zeta\varepsilon h \left[\int_{\frac{1}{(n\zeta\varepsilon h)^{1/\alpha}}}^\infty e^{i\theta x} \frac{\alpha \left(\frac{1}{(n\zeta\varepsilon h)^{1/\alpha}} \right)^\alpha}{x^{1+\alpha}} \, dx - 1 \right]
\end{aligned}$$

$$\psi_{\widetilde{\mathbb{S}}(Y)}(\theta) = \exp\left\{ \int_{\frac{1}{(n\zeta\varepsilon h)^{1/\alpha}}}^{\infty} \frac{\alpha}{x^{1+\alpha}} \left(e^{i\theta x} - 1 \right) \, dx \right\}$$

As $n\zeta\varepsilon h \to \infty$, $\frac{1}{(n\zeta\varepsilon h)^{1/\alpha}} \to 0$ and the rest of the proof follows as in Theorem 6.

Now assume K_j is purely Pareto distributed with tail index $1 < \alpha < 2$. The proof can be found in [108]. Define $\tilde{n} = n * \mathbb{E}[\mathcal{N}(h)] = n * \zeta\varepsilon h$ and approximate $\mathbb{S}(Y)$ as the sum of $\tilde{n}$ iid Pareto random variables. Note that if we choose the observation window length h such that $h < 1/\zeta\varepsilon$, then at most $\mathcal{N}(h) = 1$ and $\tilde{n} \sim n$. Therefore, the sufficient statistic is approximated as $\mathbb{S}(Y) = \sum_{j=1}^{\tilde{n}} K_j$. Again, the normalized sufficient statistic is

$$\widetilde{\mathbb{S}}(Y) = \frac{1}{\tilde{n}^{1/\alpha}} \left(\sum_{i=1}^{\tilde{n}} K_i - \tilde{n}\mathbb{E}[K_i] \right) .$$

Recall for any i, the mean of the Pareto distribution with $1 < \alpha < 2$ is

$$\mu = \mathbb{E}[K_i] = \frac{\alpha c}{\alpha - 1},$$

and the pdf,

$$f_K(k) = \frac{\alpha c^\alpha}{k^{\alpha+1}} \mathbb{1}\{k \geq c > 0\}$$

The Laplace transform of $f_K(y)$ is

$$\phi_K(\theta) = \alpha c^\alpha \int_c^\infty \frac{\exp\{-\theta k\}}{k^{\alpha+1}} d\theta = \alpha (\, c\,\theta)^\alpha \Gamma(-\alpha, c\,\theta) \ , \ \Re(\theta) > 0 \tag{4.4.9}$$

Given the expansion of the incomplete gamma function,

$$\Gamma(a, x) = \Gamma(a) - x^a \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{x^n}{a+n} \ ,$$

$\phi_K(\theta)$ can be written as

$$\begin{aligned} \phi_K(\theta) &= \alpha(c\,\theta)^\alpha \left(\Gamma(-\alpha) - (c\theta)^{-\alpha} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{(c\theta)^n}{-\alpha+n} \right) \\ &= 1 + \alpha(c\,\theta)^\alpha \Gamma(-\alpha) - \frac{\alpha}{\alpha-1}(c\,\theta) + \frac{\alpha}{2(\alpha-2)}(c\,\theta)^2 + \dots \end{aligned}$$

Center and normalize the K_i's as follows

$$\hat{K}_i = \frac{K_i - \mu}{\tilde{n}^{1/\alpha}} \tag{4.4.10}$$

$\hat{K}_i$'s are no longer restricted to positive values and we can write

$$\psi_{K_i}(\theta) = \int_{-\infty}^{\infty} e^{i\theta k} f_{K_i}(k)dk = \phi_{K_i}(-i\theta) \ ,$$

Using the characteristic function property $\psi_{aX+b}(\theta) = e^{i\theta b}\psi_X(a\theta)$, we get

$$\psi_{\hat{K}_i}(\theta) = \exp\left\{-\frac{i\,\theta\,\mu}{\tilde{n}^{1/\alpha}}\right\}\phi_{K_i}\left(-\frac{i\,\theta}{\tilde{n}^{1/\alpha}}\right)$$

The exponential and incomplete gamma functions are expanded near $\theta = 0$ such that

$$\psi_{\hat{K}_i}(\theta) = \left[1 - \frac{i\theta\mu}{\tilde{n}^{1/\alpha}} - \frac{1}{2}\left(\frac{\theta\mu}{\tilde{n}^{1/\alpha}}\right)^2 + \ldots\right]\left[1 + \frac{\alpha}{\alpha-1}\frac{i\,c\,\theta}{\tilde{n}^{1/\alpha}} + \alpha\Gamma(-\alpha)\left(\frac{-i\,c\,\theta}{\tilde{n}^{1/\alpha}}\right)^{\alpha} + \ldots\right]$$

Then

$$\begin{aligned}
\psi_{\widetilde{\mathbb{S}}(Y)}(\theta) &= \lim_{n\to\infty}\left(\psi_{\hat{K}_i}(\theta)\right)^{\tilde{n}} \\
&= \lim_{n\to\infty}\left(\exp\left\{\ln\psi_{\hat{K}_i}(\theta)\right\}\right)^{\tilde{n}} \\
&= \lim_{n\to\infty}\exp\left\{\tilde{n}\ln\left[\exp\left\{-\frac{i\,\theta\,\mu}{\tilde{n}^{1/\alpha}}\right\}\phi_{K_i}\left(-\frac{i\,\theta}{\tilde{n}^{1/\alpha}}\right)\right]\right\} \\
&= \lim_{n\to\infty}\exp\left\{\tilde{n}\times\left(-\frac{i\,\theta\,\mu}{\tilde{n}^{1/\alpha}}\right) + \tilde{n}\ln\left[1 + \frac{\alpha}{\alpha-1}\frac{i\,c\,\theta}{\tilde{n}^{1/\alpha}} + \alpha\Gamma(-\alpha)\left(\frac{-i\,c\,\theta}{\tilde{n}^{1/\alpha}}\right)^{\alpha} + \ldots\right]\right\}
\end{aligned}$$

Around zero, the natural logarithm admits the Taylor expansion $\ln(1+x) \simeq x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots$. Thus, for $\tilde{n}$ large,

$$\psi_{\widetilde{\mathbb{S}}(Y)}(\theta) = \lim_{n\to\infty}\exp\left\{-i\,\theta\,\mu\,\tilde{n}^{1-1/\alpha} + \frac{\alpha\,i\,c\,\theta}{\alpha-1}\tilde{n}^{1-1/\alpha} + \alpha\Gamma(-\alpha)\,(-i\,c\,\theta)^{\alpha} + \ldots\right\}$$

Recall the definition of the mean μ, then for $\tilde{n}$ large,

$$\begin{aligned}
\psi_{\widetilde{\mathbb{S}}(Y)}(\theta) &= \exp\left\{\alpha\Gamma(-\alpha)\left(\frac{c\theta}{i}\right)^{\alpha}\right\} \\
&= \exp\left\{\alpha c^{\alpha}\Gamma(-\alpha)\cos\left(\alpha\frac{\pi}{2}\right)|\theta|^{\alpha}\left[1 - i\,\mathrm{sign}(\theta)\tan\left(\alpha\frac{\pi}{2}\right)\right]\right\} \qquad (4.4.11)
\end{aligned}$$

Comparing $\psi_{\widetilde{\mathbb{S}}(Y)}(\theta)$ to the general characteristic function of a stable random variable, which for $1 < \alpha < 2$ is, $\exp\left\{i\delta\theta - a|\theta|^{\alpha}\left[1 + i\beta\,\mathrm{sign}(\theta)\tan\left(\frac{\pi\alpha}{2}\right)\right]\right\}$, we identify $\delta = 0$, $\beta = -1$, and $a = -c^{\alpha}\alpha\Gamma(-\alpha)\cos(\alpha\pi/2)$. The inverse Fourier transform of $\psi_{\widetilde{\mathbb{S}}(Y)}(\theta)$ gives (4.4.7). □

For given levels of decision boundaries, the next paragraph outlines the procedure to approximate

the type I and type II error probabilities, i.e. φ and ξ.

Approximated Error Probabilities

The sequential test is not defined in terms of likelihood ratio, hence we must approximate new decision boundaries using the test statistic $\mathbb{S}(Y) = \sum_{i=1}^{n} Y_i$. Poor [103] explains the procedure, which requires that the moment generating function of the test statistic $g(Y_i)$ satisfies $\mathcal{M}_k(\hat{t}_k) = 1$ for $k = 0, 1$ denoting the hypothesis H_0 or H_1. Additional assumptions include $\mathbb{P}\{g(Y_i) < 0\} > 0$ and $\mathbb{P}\{g(Y_i) > 0\} > 0$, for all $i = 1, 2, \ldots, n$. Setting g to be the identity function violates the assumptions $\mathbb{P}\{g(Y_i) < 0\} > 0$ since the random variables $\{Y_i\}$ are positive by definition. This issue can be resolved by simply adjusting g and subtracting the mean of Y_i. Tartakovsky et al. [116] propose a similar adjustment for the non-parametric sequential change-point problem. Let the adjusted test statistic be

$$g(Y_i) = Y_i - \mu_i^0 - \epsilon\hat{\mu}_i^1, \tag{4.4.12}$$

where μ^0 is the common and known mean of Y_i under H_0 and is computed as

$$\mu^0 = \mathbb{E}[Y_i|H_0] = \zeta\varepsilon t\frac{\alpha c}{\alpha - 1} \tag{4.4.13}$$

The last equality follows from Wald's Identity (see Corollary 1). The parameter $\hat{\mu}_{n-1}^1$ is the sample mean under H_1 estimated from the $n-1$ previous observations. The parameter $\epsilon \in (0, 1)$ must be determined based on historical data. The decision boundaries can now be estimated as shown in Proposition 11.

Proposition 11. *Let $k = 0, 1$ denote the hypothesis H_0 or H_1.*
Assume the excess over boundaries (4.2.9) are negligible then the error probabilities satisfy

$$\begin{aligned} \varphi &\simeq \left(1 - e^{\hat{t}_0\tilde{\kappa}_0}\right) / \left(e^{\hat{t}_0\tilde{\kappa}_1} - e^{\hat{t}_0\tilde{\kappa}_0}\right) \\ \xi &\simeq \left(1 - e^{\hat{t}_1\tilde{\kappa}_1}\right) / \left(e^{\hat{t}_1\tilde{\kappa}_0} - e^{\hat{t}_1\tilde{\kappa}_1}\right) \end{aligned} \tag{4.4.14}$$

where $\hat{t}_0$ and $\hat{t}_1$ are solutions to

$$-\hat{t}_k(\mu_i^0 + \epsilon\hat{\mu}_i^1) + \lambda_k\left[\alpha_k(-c_k\hat{t}_k)\Gamma(-\alpha_k, -c_k\hat{t}_k) - 1\right] = 0. \tag{4.4.15}$$

The expected samples sizes given H_0 and H_1 respectively are

$$\begin{aligned} \mathbb{E}[\mathcal{T}|H_0] &= -\frac{\varphi\ln\tilde{\kappa}_1 + (1-\varphi)\ln\tilde{\kappa}_0}{\epsilon\hat{\mu}_i^1} \\ \mathbb{E}[\mathcal{T}|H_1] &= \frac{(1-\xi)\ln\tilde{\kappa}_1 + \xi\ln\tilde{\kappa}_0}{-\mu_i^0 - \epsilon\hat{\mu}_i^1 + \mathbb{E}[Y_i|H_1]} \end{aligned} \tag{4.4.16}$$

Proof. Let λ_k denote the rate of the compound Poisson process, given hypothesis H_k, $k = 0, 1$. The mgf of the random variable, $g(Y_i)$ $i = 1, 2, \ldots, n$, defined in (4.4.12) is

$$\begin{aligned}\mathcal{M}_k(t) &= \mathbb{E}\left[e^{tg(Y_i)}|H_k\right] \\ &= \exp\left\{-t(\mu_i^0 + \epsilon\hat{\mu}_i^1)\right\}\exp\left\{\lambda_k\left[\alpha_k(-c_k t)\Gamma(-\alpha_k, -c_k t) - 1\right]\right\}\end{aligned}$$

Derive $\hat{t}_k$ by numerically solving the following equation

$$\mathcal{M}_k(\hat{t}_k) = 1 \equiv -\hat{t}_k(\mu_i^0 + \epsilon\hat{\mu}_i^1) + \lambda_k\left[\alpha_k(-c_k\hat{t}_k)\Gamma(-\alpha_k, -c_k\hat{t}_k) - 1\right] = 0$$

At $\hat{t}_k$, Wald's Fundamental Identity implies that

$$\mathbb{E}\left[e^{t\mathbb{S}(Y)}|H_k\right] = 1$$

Ignoring the excess over boundaries implies that

$$\mathbb{S}(Y) = \begin{cases}\tilde{\kappa}_0 & \text{with probability } 1-\varphi \\ \tilde{\kappa}_1 & \text{with probability } \varphi\end{cases}$$

under H_0 and similarly under H_1

$$\mathbb{S}(Y) = \begin{cases}\tilde{\kappa}_0 & \text{with probability } \xi \\ \tilde{\kappa}_1 & \text{with probability } 1-\xi\end{cases}$$

From these equations, we deduce that

$$\begin{aligned}(1-\varphi)e^{\hat{t}_0\tilde{\kappa}_0} + \varphi e^{\hat{t}_0\tilde{\kappa}_1} &\simeq 1 \\ \xi e^{\hat{t}_1\tilde{\kappa}_0} + (1-\xi)e^{\hat{t}_1\tilde{\kappa}_1} &\simeq 1\end{aligned}$$

which yields (4.4.14) [103].
The expected sample size equations (4.4.16) follow from (4.3.13) with the new decision boundaries and

$$\begin{aligned}\mathbb{E}\left[g(Y_i)|H_0\right] &= \mathbb{E}\left[Y_i - \mu_i^0 - \epsilon\hat{\mu}_i^1|H_0\right] = -\epsilon\hat{\mu}_i^1 \\ \mathbb{E}\left[g(Y_i)|H_1\right] &= \mathbb{E}\left[Y_i - \mu_i^0 - \epsilon\hat{\mu}_i^1|H_1\right] = -\mu_i^0 - \epsilon\hat{\mu}_i^1 + \mathbb{E}\left[Y_i|H_1\right]\end{aligned}$$

□

4.4.2 Mixture Cluster Size

Assume cluster sizes have a mixture distribution as described in (2.1.3). Results from the previous section apply. The distribution of the test statistic $\mathbb{S}(Y)$ in (4.4.4) can be approximated by a stable distribution. For large values of the pdf, Propositions 5 and 6 shows that the Pareto component with the heaviest tail dominates. Therefore Theorem 11 applies. Detection error probabilities can be estimated with Equation (11). The mean under H_0 given by (4.4.13) becomes

$$\mu^0 = \mathbb{E}\left[Y_i|H_0\right] = \zeta\varepsilon t\left[q\sum_{i=1}^{\lfloor c\rfloor-1} x_i\mathbb{P}\left\{K=x_i\right\} + q_1\frac{\alpha_1 c_1}{\alpha_1-1} + q_2\frac{\alpha_2 c_2}{\alpha_2-1}\right] \tag{4.4.17}$$

4.5 Sequential HT for Homogeneous Poisson with Cluster Size

Given that each flow i as a cluster of size K_i, we can apply the "standard" sampling method, which consists of observing one flow at the time. This contrasts from the previous section with the time-discrete sampling approach. Assume packet flows are independent.

4.5.1 Pareto Cluster Size

Given one single flow i with unknown source IP address, the total number of data packets with unknown IP address is equivalent to the flow cluster size K_i. In this section observations are taken directly to be the flow cluster sizes, $Y_i = K_i$, $i = 1, 2, \ldots, n$. Cluster sizes are assumed to be Pareto random variables, where the parameters vary according to the tested hypotheses H_0 and H_1. This follows since realistically attack packet flows may have a different cluster size distribution than legitimate packet flows, which modifies the distribution under H_1. The binary hypothesis test corresponding to the DDoS detection problem can be written as

$$\begin{cases} H_0 : Y_i \sim \text{Par}(\alpha_0, c_0) \\ H_1 : Y_i \sim \text{Par}(\alpha_1, c_1) \end{cases} \tag{4.5.1}$$

Note that observations $\{Y_i\}$ are no longer compound random variables. Therefore, the technique of sampling one flow at the time simplifies computations because unlike compound random variables, the Pareto distribution has a closed form pdf. Next, we formulate the sequential HT for solving the binary test in (4.5.1). Essentially, we present the modified Wald's boundaries and compute the average number of collected samples.

Lemma 4 indicates that under H_0, $\frac{Y_i}{c_0} \sim \text{Par}(\alpha_0, 1)$. Similarly, $\frac{Y_i}{c_1} \sim \text{Par}(\alpha_0, 1)$, given H_1. Given

n flows, the joint likelihood ratio is

$$\mathbb{L}(y) = \prod_{i=1}^{n} \frac{f_{Y_i|H_1}(y_i)}{f_{Y_i|H_0}(y_i)} = \prod_{i=1}^{n} \frac{\alpha_1}{\alpha_0} y_i^{\alpha_0 - \alpha_1}. \tag{4.5.2}$$

Applying the log function yields

$$\ln \mathbb{L}(y) = \sum_{i=1}^{n} \left[\ln \left(\frac{\alpha_1 c_1^{\alpha_1}}{\alpha_0 c_0^{\alpha_0}} \right) + (\alpha_0 - \alpha_1) \ln y_i \right] \tag{4.5.3}$$

Define $S_n = \sum_{i=1}^{n} Z_i$, where

$$Z_i = \ln \left(\frac{\alpha_1 c_1^{\alpha_1}}{\alpha_0 c_0^{\alpha_0}} \right) + (\alpha_0 - \alpha_1) \ln y_i. \tag{4.5.4}$$

Lemma 13 shows that the random variable $\{\ln Y_i\}$ in (4.5.4) are distributed according to an exponential distribution.

Lemma 13. *If $U \sim Par(\alpha, c)$ then $V = \ln\left(\frac{U}{c}\right) \sim Exp(\alpha)$.*

Proof. The result follows since

$$\mathbb{P}\{V < v\} = \mathbb{P}\left\{ \ln\left(\frac{U}{c}\right) < v \right\} = \mathbb{P}\{U < ce^v\} = 1 - \left(\frac{c}{ce^v}\right) = 1 - e^{-\alpha v}$$

□

Therefore $\ln Y_i \sim \text{Exp}(\alpha_0)$ and $\ln Y_i \sim \text{Exp}(\alpha_1)$ given H_0 and H_1, respectively. Note that $S(Y) = \sum_{i=1}^{n} \ln Y_i$ is a sufficient statistic for α and $\sum_{i=1}^{n} \ln Y_i \sim \text{Gamma}(\alpha, n)$. We now deduce the upper and lower decision boundaries. Recall φ and ξ are the type I and II error probabilities, respectively.

Proposition 12. *Assume the excess over boundaries (4.2.9) are negligible. The sequential ratio test can be formulated in terms of the sufficient statistic $S(Y) = \sum_{i=1}^{n} \ln Y_i$ as in (4.3.6). The corresponding decision boundaries are defined in terms of Wald's approximations in (4.2.10).*

$$\kappa_0 = \frac{\ln \tau_0^*}{\alpha_1 - \alpha_0} + rn, \quad \textit{and} \quad \kappa_1 = \frac{\ln \tau_1^*}{\ln L - \ln \zeta\varepsilon} + rn, \tag{4.5.5}$$

where

$$r = \frac{1}{\alpha_1 - \alpha_0} \times \ln \left(\frac{\alpha_1 c_1^{\alpha_1}}{\alpha_0 c_0^{\alpha_0}} \right)$$

The proof is similar to that of Proposition 9.

Proposition 13. *Considering no excess over boundaries, the detection error probabilities φ and ξ must satisfy*

$$\varphi \simeq \left(1 - e^{\hat{t}_0 \tilde{\kappa}_0}\right) / \left(e^{\hat{t}_0 \tilde{\kappa}_1} - e^{\hat{t}_0 \tilde{\kappa}_0}\right) \tag{4.5.6}$$

and

$$\xi \simeq \left(1 - e^{\hat{t}_1 \tilde{\kappa}_1}\right) / \left(e^{\hat{t}_1 \tilde{\kappa}_0} - e^{\hat{t}_1 \tilde{\kappa}_1}\right) \tag{4.5.7}$$

where $\hat{t}_0 = 1$ and $\hat{t}_1 = -1$.

Proof. Let $k = 0, 1$ denoting the hypothesis H_0 or H_1 then the mgf of z_i is

$$\begin{aligned} \mathcal{M}_{Z,k}(t) &= \exp\left\{t \ln\left(\frac{\alpha_1}{\alpha_0}\right)\right\} \times \mathcal{M}_{\ln Y,k}((\alpha_0 - \alpha_1)t) \\ &= \exp\left\{t \ln\left(\frac{\alpha_1}{\alpha_0}\right)\right\} \times \left(1 - \frac{(\alpha_0 - \alpha_1)t}{\alpha_k}\right)^{-1} \end{aligned}$$

We want to find the solution $\hat{t}_k$ such that $\mathcal{M}_{Z,k}(\hat{t}_k) = 1$. In other words, solve

$$\hat{t}_k \ln\left(\frac{\alpha_1}{\alpha_0}\right) = \ln\left(1 - \frac{(\alpha_0 - \alpha_1)\hat{t}_k}{\alpha_k}\right)$$

The equation above yields $\hat{t}_0 = 1$ and $\hat{t}_1 = -1$.
If we neglect the excess over boundaries then S_n is a discrete random variable which, given H_0, takes the following values

$$S_n = \begin{cases} \ln \tau_0 & \text{with probability } 1 - \varphi \\ \ln \tau_1 & \text{with probability } \varphi \end{cases}$$

Similarly, under H_1,

$$S_n = \begin{cases} \ln \tau_0 & \text{with probability } \xi \\ \ln \tau_1 & \text{with probability } 1 - \xi \end{cases}$$

By Wald's Fundamental Identity,

$$\mathbb{E}\left[e^{tS_n} \left(\mathcal{M}_k(t)\right)^{-n} | H_k\right] = 1 \tag{4.5.8}$$

□

Similarly, Wald's Fundamental Identity can be applied to approximate the expected sample size of the test (4.5.1), using the random variables z_i.

The decision boundaries presented above corresponds to the case of negligible excess over boundaries. Assume at the stopping time, the likelihood ratio exceeds one of the decision boundaries. This case is addressed in the following section.

Excess Over One Boundary

Exploiting the relationship between Pareto and exponential random variables (see Lemma 13), the binary hypothesis problem in (4.5.1) reduces to comparing the parameters of two exponential distributions. It can be reformulated as

$$\begin{cases} H_0 : \hat{Y}_i = \ln \frac{Y_i}{c_0} \sim \text{Expo}(\alpha_0) \\ H_1 : \hat{Y}_i = \ln \frac{Y_i}{c_1} \sim \text{Expo}(\alpha_1) \end{cases} \tag{4.5.9}$$

Note that hypothesis testing is invariant under the same transformation. In practice, the same transformation should be applied to observations. On one hand, we can assume the scale parameters satisfy $c_0 = c_1 = c$. Alternatively, c can be set as $c = \max\{c_0, c_1\}$.

The case of hypothesis testing with exponential means has been studied earlier. Given preassigned detection error probabilities φ and ξ, Albert [1] determined that the decision boundaries must satisfy

$$\begin{aligned} \kappa_0 &= \ln \frac{\xi}{1-\varphi} - \ln \frac{\alpha_1}{\alpha_0} \\ \kappa_1 &= \ln \frac{1-\xi}{\varphi} - (m+1) \ln \frac{\alpha_1}{\alpha_0} + \ln M_m(s'; \alpha_0), \end{aligned} \tag{4.5.10}$$

where for $0 \le x \le s'$

$$\begin{aligned} M_m(x;\alpha) &= \frac{e^{(1+h')x} G_m\left(-\kappa_1' + (m+1)s'\right) - G_m\left(-\kappa_1' + (m+1)s' - x\right)}{e^{x} G_m\left(-\kappa_1' + (m+1)s'\right) - G_m\left(-\kappa_1' + (m+1)s' - x\right)}, \\ G_m(t) &= \sum_{j=0}^{m} \frac{(-1)^j}{j!} e^{-js'} (t + \kappa_1' - js')^j, \\ \kappa_1' &= \frac{\kappa_1^* \alpha}{\alpha_1 - \alpha_0}, \\ s' &= \frac{\alpha \ln(\alpha_1/\alpha_0)}{\alpha_1 - \alpha_0}. \end{aligned}$$

κ_1^* is Wald's approximations defined in (4.2.10). Equation (4.5.10) indicates that Wald's approximation is accurate when α_1/α_0 is close to one, that is when the exponential distribution parameters satisfy $\alpha_1 \approx \alpha_0$.

Computing the parameter m in the threshold function (4.5.10) is a trial and error procedure, which might be inconvenient. Based on Wald's Fundamental Identity (see Theorem 10), the OC function can be approximated as follows.

Proposition 14.

$$1 - \mathcal{Q}(\alpha) = \left(1 - e^{\hat{t}_0 \tilde{\kappa}_0}\right) \left(e^{\hat{t}_k \kappa_1} \frac{\alpha}{\alpha - t_k} - e^{\hat{t}_0 \kappa_0}\right)^{-1} \tag{4.5.11}$$

Proof. The log-likelihood ratio corresponding to the binary hypothesis problem in (4.5.9) is

$$\ln \mathbb{L}(y) = \ln \prod_{i=1}^{n} \frac{\alpha_1 e^{-\alpha_1 y}}{\alpha_0 e^{-\alpha_0 y}} = \sum_{i=1}^{n} \left[\ln \left(\frac{\alpha_1}{\alpha_0} \right) + (\alpha_0 - \alpha_1) y_i \right] \tag{4.5.12}$$

Similarly let $\hat{S}_n = \sum_{i=1}^{n} \hat{Z}_i$, where

$$\hat{Z}_i = \ln \left(\frac{\alpha_1}{\alpha_0} \right) + (\alpha_0 - \alpha_1) y_i. \tag{4.5.13}$$

Let $k = 0, 1$ and α_k be the parameter of the exponential distribution, given hypothesis H_k. The mgf of the random variable $\hat{Z}_i$ is computed as

$$\mathcal{M}_{\hat{Z}}(t, \alpha) \triangleq \mathbb{E}\left[e^{tZ} | \alpha \right] = \exp \left\{ t \ln \left(\frac{\alpha_1}{\alpha_0} \right) \right\} \times \left(1 - \frac{(\alpha_0 - \alpha_1) t}{\alpha} \right)^{-1}. \tag{4.5.14}$$

We want to find the solution $\hat{t}_k$ such that $\mathcal{M}_{\hat{Z}}(\hat{t}_k, \alpha_k) = 1$. In other words, solve

$$\hat{t}_k \ln \left(\frac{\alpha_1}{\alpha_0} \right) = \ln \left(1 - \frac{(\alpha_0 - \alpha_1) \hat{t}_k}{\alpha_k} \right)$$

The equation above yields $\hat{t}_0 = 1$ and $\hat{t}_1 = -1$. By Wald's Fundamental Identity,

$$\mathbb{E}\left[e^{tS_n} \left(\mathcal{M}_k(t) \right)^{-n} | H_k \right] = 1 \tag{4.5.15}$$

Define the stopping time $\mathcal{T} = \inf \{ n : S_n > \kappa_1 \}$. Then

$$\mathbb{E}\left[e^{t_k S_{\mathcal{T}}} | H_k \right] = 1, \tag{4.5.16}$$

with $k = 0, 1$, $\hat{t}_0 = 1$ and $\hat{t}_1 = -1$.

As mentioned earlier, at the stopping time, the likelihood ratio $\mathbb{L}(y)$ crosses the decision boundaries in three possible scenarios: no excess, excess over one decision boundary, and finally, excess over both decision boundaries. Assume the second scenario, that is let $\tau_0 - \mathbb{L}(y) \approx 0$ and $\mathbb{L}(y) - \tau_1 > 0$. Then, under H_0 and under H_1, the test statistic S_n respectively takes values

$$S_n = \begin{cases} \ln \tau_0 & \text{with probability } 1 - \varphi \\ \ln \tau_1 + R & \text{with probability } \varphi \end{cases} \tag{4.5.17}$$

$$S_n = \begin{cases} \ln \tau_0 & \text{with probability } \xi \\ \ln \tau_1 + R & \text{with probability } 1 - \xi \end{cases}$$

Let $\kappa_1 = \ln \tau_1$, then $R = S_n - \kappa_1$ denotes the excess over boundary, which is a random variable

with unknown distribution function. Recall $S_n = \sum_{i=1}^{n} Z_i$. Let μ_Z be the common finite mean of the random variables $\{Z_i\}$ described in (4.5.13). Then Corollary 1 shows that $\mathbb{E}[S_{\mathcal{T}}] = \mu_Z \mathbb{E}[\mathcal{T}]$. Or equivalently $\mu_Z \mathbb{E}[\mathcal{T}] = \mathbb{E}[R] + \kappa_1$. Wald [121] provides an upper bound to $\sup_{\tau_1} \mathbb{E}[R]$ as

$$\sup_{\tau_1} \mathbb{E}[R] \leq \sup_{r \geq 0} \mathbb{E}[Z_i - r | Z_i > r] \tag{4.5.18}$$

Lorden [73] emphasizes that the bound is exact for the exponential distribution, in which case the supremum is attained when $r = 0$. Therefore,

$$\mathbb{E}[R|H_k] \leq \mathbb{E}[Z_i|H_k] \leq \frac{1}{\alpha_k}(\alpha_0 - \alpha_1) \tag{4.5.19}$$

In terms of the OC function, φ and ξ, the type I and II error probabilities, can be written as $\mathcal{Q}(\alpha_0) = 1 - \varphi$ and $\mathcal{Q}(\alpha_1) = \xi$. Equations (4.5.16) and (4.5.17) imply that

$$\begin{aligned} (1-\varphi)e^{t_0\kappa_0} + \varphi \mathbb{E}\left[e^{t_0(\kappa_1 + R)}\right] &= 1 \\ \xi e^{t_1\kappa_0} + (1-\xi)\mathbb{E}\left[e^{t_1(\kappa_1+R)}\right] &= 1 \end{aligned} \tag{4.5.20}$$

From the equation, deduce the complementary OC function as

$$1 - \mathcal{Q}(\alpha) = \left(1 - e^{\hat{t}_0 \tilde{\kappa}_0}\right)\left(e^{\hat{t}_k \kappa_1}\mathbb{E}\left[e^{t_k R}|\alpha\right] - e^{\hat{t}_0\kappa_0}\right)^{-1} \tag{4.5.21}$$

$$\xi \simeq \left(1 - e^{\hat{t}_1\tilde{\kappa}_1}\right)\left(e^{\hat{t}_1\tilde{\kappa}_0} - e^{\hat{t}_1\tilde{\kappa}_1}\right)^{-1} \tag{4.5.22}$$

The excess over boundary random variable is non-negative and is exponentially distributed due to the memoryless property of the exponential distribution. Therefore,

$$1 \leq \mathbb{E}\left[e^{t_k R}|\alpha\right] = \frac{\alpha}{\alpha - t_k} \tag{4.5.23}$$

□

4.5.2 Mixture Cluster Size

If cluster sizes $\{K_j\}$ are distributed according to a mixture distribution, then the likelihood ratio has an explicit expression and the SPRT applies with Wald's approximations representing the decision boundaries (see (4.2.10)). The binary hypothesis testing is the same as (4.5.1), except the random variables Y_i have a mixture distribution.

Consider a mixture distribution with m components and q_i as the mixture parameters. The

likelihood ratio is

$$\mathbb{L}(y) = \prod_{i=1}^{n} \sum_{j=1}^{m} \frac{q_j f_{j|H_1}(y)}{q_j f_{j|H_0}(y)} \tag{4.5.24}$$

In particular, consider a mixture of one discrete and two Pareto distributions as introduced in Section 2.1 then

$$\mathbb{L}(y) = \prod_{i=1}^{n} \frac{\left[\hat{q} \sum_{k=1}^{\lfloor c_1 - 1 \rfloor} \hat{p}_k \mathbb{1}\left\{y = y_k\right\} + q_3 \frac{\alpha_3 c_3^{\alpha_3}}{y^{\alpha_3+1}} + q_4 \frac{\alpha_4 c_4^{\alpha_4}}{y^{\alpha_4+1}} | H_1\right]}{\left[q \sum_{k=1}^{\lfloor c_1 - 1 \rfloor} p_k \mathbb{1}\left\{y = y_k\right\} + q_1 \frac{\alpha_1 c_1^{\alpha_1}}{y^{\alpha_1+1}} + q_2 \frac{\alpha_2 c_2^{\alpha_2}}{y^{\alpha_2+1}} | H_0\right]} \tag{4.5.25}$$

The sequential test is described in (4.2.5) where $\tau_0 = \frac{\xi}{1-\varphi}$ and $\tau_1 = \frac{1-\xi}{\varphi}$.

4.6 Summary

In this chapter, sequential detection models were presented. By allowing the sample size to vary, detection performances can improve. Test statistics were computed from the likelihood ratio or as a real function of the observations. The alternative hypothesis which suggests a DDoS attack is taking place, was formulated as a composite hypothesis testing so that UMP tests were required. Note that the models presented here apply when the packet flows arrive according to an inhomogeneous Poisson process. The SPRT can be truncated appropriately.

CHAPTER 5

EMPIRICAL DDOS DETECTION MODEL

This chapter departs from the previous ones in that no theoretical model is developed for network traffic. As we have seen, modeling Internet traffic is very challenging due to the complex and dynamic nature of the network. Non-parametric detection models can be more indicated then. The detection model presented here is non-parametric and is developed using empirical real network traces. Previous models can potentially lead to larger detection errors since they combined both estimation and detection theory. The non-parametric model then has the advantage of being more realistic since estimation errors are eliminated.

The empirical detection system monitors two traffic statistics simultaneously to test the hypothesis that only legitimate traffic is observed (H_0) against the hypothesis that a DDoS attack is taking place (H_1). In addition to the percentage of unknown source IP addresses, a second statistic, defined as the mean to standard deviation ratio of packet sizes, is introduced. Given that no closed form pdf exist, the joint likelihood ratio is estimated using a probability grid. Simulations are conducted using the MIT network traces.

Two models are developed. The first is equivalent to a supervised learning classification. The observation space, i.e. the probability grid, is partitioned into the rejection and acceptance region [59], using a Neyman-Pearson criterion. The binary decision between two hypothesis relies on the likelihood ratio and where it classifies the observed sample in the grid. The second detection model is a non-parametric sequential change-point model [116, 135].

5.1 Data Packet Sampling Procedure and Traffic Statistics

The following section discusses observation windows, how data packet are sampled and specify how traffic statistics are computed.

5.1.1 Measuring Traffic Statistics

Two traffic statistics are collected at discrete intervals called observation windows. Each observation window Δ_i, $i = 1, 2, \ldots, M$, is measured in terms of number of data packets rather than time intervals. Data packet sampling is addressed in a subsequent section. As in the previous chapters,

the first monitored statistic is the percentage of unknown source IP addresses. Assume all attack data packets have an unknown source IP address. Even though all spoofed source IP addresses are new to SIL, the identification of attack packets is not straightforward as legitimate data packets also have unknown source IP addresses. Let A_i be the number of attack data packets and U_i the total number of legitimate data packets with unknown source IP addresses in an observation window $Delta_i$, respectively. Assume Δ_i contains m_i data packets, and define the percentage of new IP addresses as

$$N_i = \frac{U_i + A_i}{m_i}. \tag{5.1.1}$$

N_i is a discrete random variable defined on the set $\mathcal{Y}_N = \{0, \frac{1}{m_L}, \frac{2}{m_L}, \ldots, \frac{m_L-1}{m_L}, \frac{1}{m_U}, \ldots, \frac{m_U-1}{m_U}, 1\}$. In addition to differentiating legitimate from attack traffic, N_i can determine whether data packets are part of a flash crowd or from a flooding DDoS attack. This is possible since most source IP addresses from a flash crowd appear in SIL [63, 102].

The second network traffic statistic to be monitored is the dispersion of data packet sizes. Authentic Internet users have different needs therefore legitimate traffic presents a large range of packet sizes. On the other hand, attack data are machined generated and are usually of the same size. For example, the SYN and ICMP flooding attacks have 40 and1500 byte data packets, respectively [29]. Moreover, Lan et al. [69] studied 90 DDoS, all attacks used 60 byte size packets. Unless, otherwise specified, assume attack data packets all have the same size. This characteristic can be exploited to identify potential threats to the target network but once again the identification process is marred by the incessant arrival of legitimate data packets. In their attempt to discriminate attack from legitimate data packets, Du and Abe [29] analyze the entropy of the distribution of data packet sizes. Although packet size entropy can effectively measure the dispersion of a random variable, an attacker can be clever and bypass this approach. For example, consider an attacker who send data packets of sizes 40, 41, 42, 43... bytes. The resulting entropy is large, indicating legitimate traffic conditions while the network is in fact under attack. Instead, we propose to compute the mean to standard deviation ratio of packet size for each observation window. Let Z_k be the size of packet k, where $k \in \Delta_i$, $k = 1, 2, \ldots, m_i$. For each observation window Δ_i of length m_i ($i = 1, 2, \ldots, M$), define the ratio as

$$X_i = \frac{\bar{Z}_i}{\sigma_i} \ , \tag{5.1.2}$$

where $\bar{Z}_i$ is the arithmetic average of packet sizes in Δ_i, i.e.,

$$\bar{Z}_i = \frac{1}{m_i} \sum_{k=1}^{m_i} Z_k \ ,$$

and σ_i the corresponding unbiased estimator of the standard deviation of packet sizes

$$\sigma_i = \sqrt{\frac{1}{m_i - 1} \sum_{k=1}^{m_i} \left(Z_k - \bar{Z}_i\right)^2} \ . \tag{5.1.3}$$

Ratio X_i is normalized between zero to compute the second traffic statistic as follows

$$S_i = \frac{X_i}{X_i + 1} = \frac{\bar{Z}_i}{\bar{Z}_i + \sigma_i} \ . \tag{5.1.4}$$

S_i can effectively measure the dispersion of packet sizes and also overcomes the limitation of the entropy approach proposed in [29]. Note that both traffic statistics are normalized such that, for any i, $0 \leq N_i, S_i \leq 1$. In all subsequent sections, assume the random variables N_i and S_i are independent. However note that successive readings (i.e. N_{i-1} and N_i or S_{i-1} and S_i) can be correlated over time [6, 126]. The two heterogeneous statistics allow monitoring different aspects of Internet traffic at once to effectively distinguish legitimate from attack data. Detection results are more reliable than using only one statistic.

5.1.2 Packet Sampling

Let m_i be the size of observation window Δ_i. This indicates that traffic statistics are computed for every batch of m_i data packets that arrive at the control router. If m_i is small computations are fast but resulting statistics can be ambiguous. Whereas large m_i's require more calculation and a longer data collection process [118]. Weak victims particularly don't survive long delays of computation. For instance, if the average data packet rate is 10 pps then collecting 100 packets require waiting at least 10s. However, some DDoS attacks last less than 10s [101, 29]. The necessity arises to define a lower (m_L) and an upper bounds (m_U) on observation window lengths. Consequently, $\{m_i\}$ are iid random variables with range $[m_L, m_U]$, and observation windows can have uneven sizes. One can think of an alternative option to define observation window lengths as time intervals. However, due to the dynamic nature of Internet traffic (non-homogeneous arrival process), this option is inefficient. For example, simulations with the MIT traces show that a one second time interval can contain anywhere between 0 to 662 data packets. Next, we explain how to set the limits m_L and m_U.

Our approach to finding the boundaries consists of sampling data packets until $\sigma_i > 0$, where σ_i is the standard deviation defined in (5.1.3). This is similar to applying a Wald's sequential analysis with stopping criteria as $\sigma_i > 0$ [121]. Figure 5.1 illustrates the distribution of samples sizes resulting from using the criteria $\sigma_i > 0$, $m_U = \infty$ and $m_L = 0$. For clarity, the graph has been truncated after $m > 100$.

Figure 5.1 demonstrates that with probability $\simeq 0.45$, two consecutive legitimate data packets

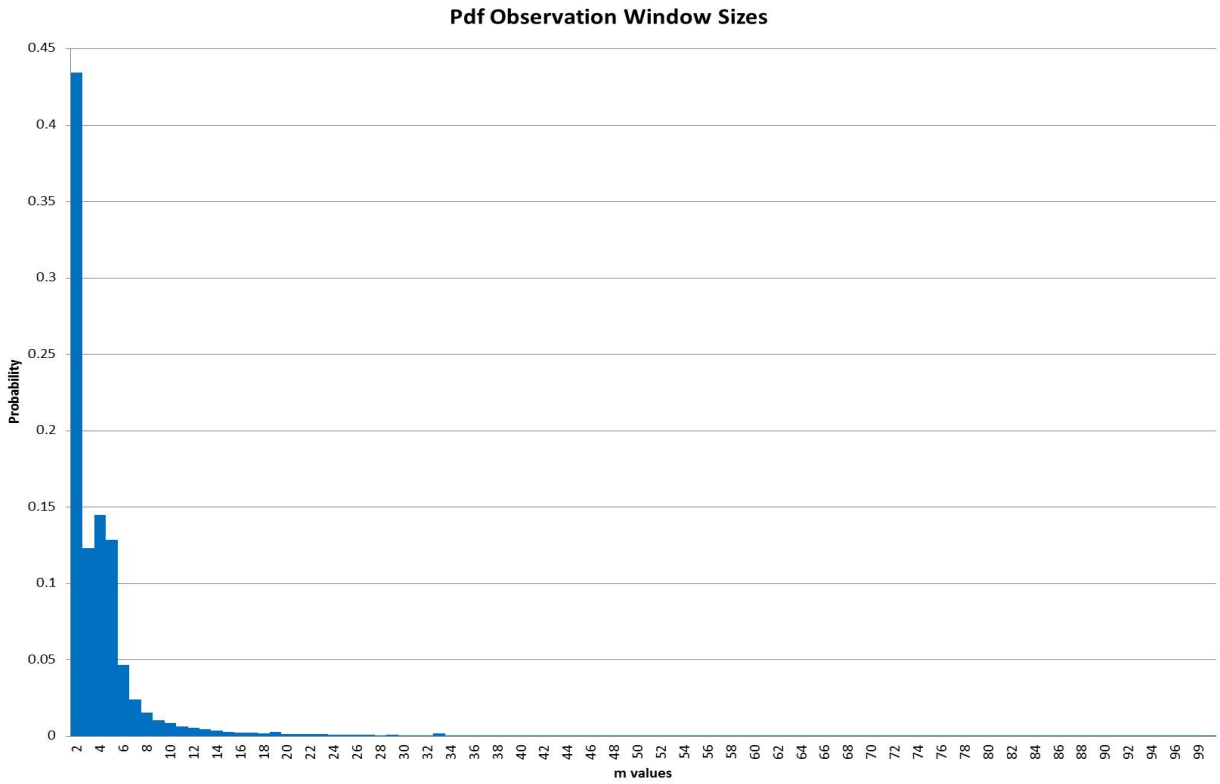


Figure 5.1: Sampling Size (m) Distribution for W1-M

have different sizes. Even though it is a very rare event (probability 1/130425), it can happen that as many as 1173 consecutive legitimate data packets have the same size. This emphasizes the need of an upper bound m_U on observation window lengths. Additional numerical results are given in Table 5.1 which expresses for each dataset, the percentage of observations with size less or equal to 2 , 5, 10, 20, 50 and 60 packets. Observe that about 99% of observation windows have size less than 50 packets and on average 97.28% of samples have size less than 20. In other words, setting $m_U = 50$ guarantees that about 99% of observation windows satisfy the condition $\sigma_i > 0$.

Table 5.1: Percentiles Observation Window Size

m_i	W1-M	W1-Tu	W1-W	W1-Th	W1-F
2	43.42	30.93	48.09	37.97	39.76
5	83.1	80.25	84.24	82.25	79.08
10	93.66	93.22	94.67	93.49	92.24
20	97.19	97.27	98.07	96.55	96.73
50	99.33	99.19	99.65	99.31	99.16
60	99.51	99.4	99.76	99.49	99.38

Before setting the final sampling parameters, we also measure the impact of varying m_U and m_L on traffic statistics. For the W1-M dataset, Figure 5.2 illustrates the probability density of the statistics N and S, respectively. In both graphs, the red area depicts the probability distribution when data is sampled with parameters $m_L = 0$ and $m_U = \infty$ while the blue area represents the sampling case with $m_L = 20$ and $m_U = \infty$. Tables 5.2 and 5.3 provide additional numerical results and show quantiles for both traffic statistics.

Table 5.2: N Distribution Statistics

Statistic	$m_L = 0$, $m_U = \infty$	$m_L = 10$, $m_U = \infty$	$m_L = 20$, $m_U = \infty$
Average	0.152	0.149	0.15
Median	0	0	0
25th Percentile	0	0	0
75th Percentile	0	0.1	0.15
95th Percentile	1	0.923	0.9

Table 5.3: S Distribution Statistics

Statistic	$m_L = 0$, $m_U = \infty$	$m_L = 10$, $m_U = \infty$	$m_L = 20$, $m_U = \infty$
Average	0.69	0.652	0.635
Median	0.78	0.61	0.561
25th Percentile	0.446	0.429	0.425
75th Percentile	0.871	0.9	0.897
95th Percentile	0.972	0.956	0.9514

In general, quantiles decrease as the lower bound m_L increases. The impact is less noticeable for the percentage of unknown source IP address statistic. Based on these experiments, we set the maximum observation window size to be $m_U = 50$ so that only large tail values are eliminated. The lower bound is set to $m_L = 20$ and observation windows lengths are random variables taking integer values between 20 and 50. Note that the overall average packet rate for the MIT datasets is approximately 17.98 pps. Therefore, collecting 50 and 20 data packets require an average waiting time of 2.78s and 1.11s, respectively. We estimate that these are reasonable time intervals. Recall the training phase uses historical data only, computation delays and consequently m_U are irrelevant, however a lower bound m_L is necessary. To calibrate the model as close as possible to real-life conditions, training and testing both use $m_L = 20$ and $m_U = 50$. Having an upper bound on observation window sizes also helps identify when there is a surge of data packets with the same size suggesting an attack is taking place. The phenomenon manifests by several consecutive observation windows having the same size m_U.

The next section present the first detection model, which is formulated as a binary detection model. Similar to supervised learning, the detection mechanism involves two steps. The training

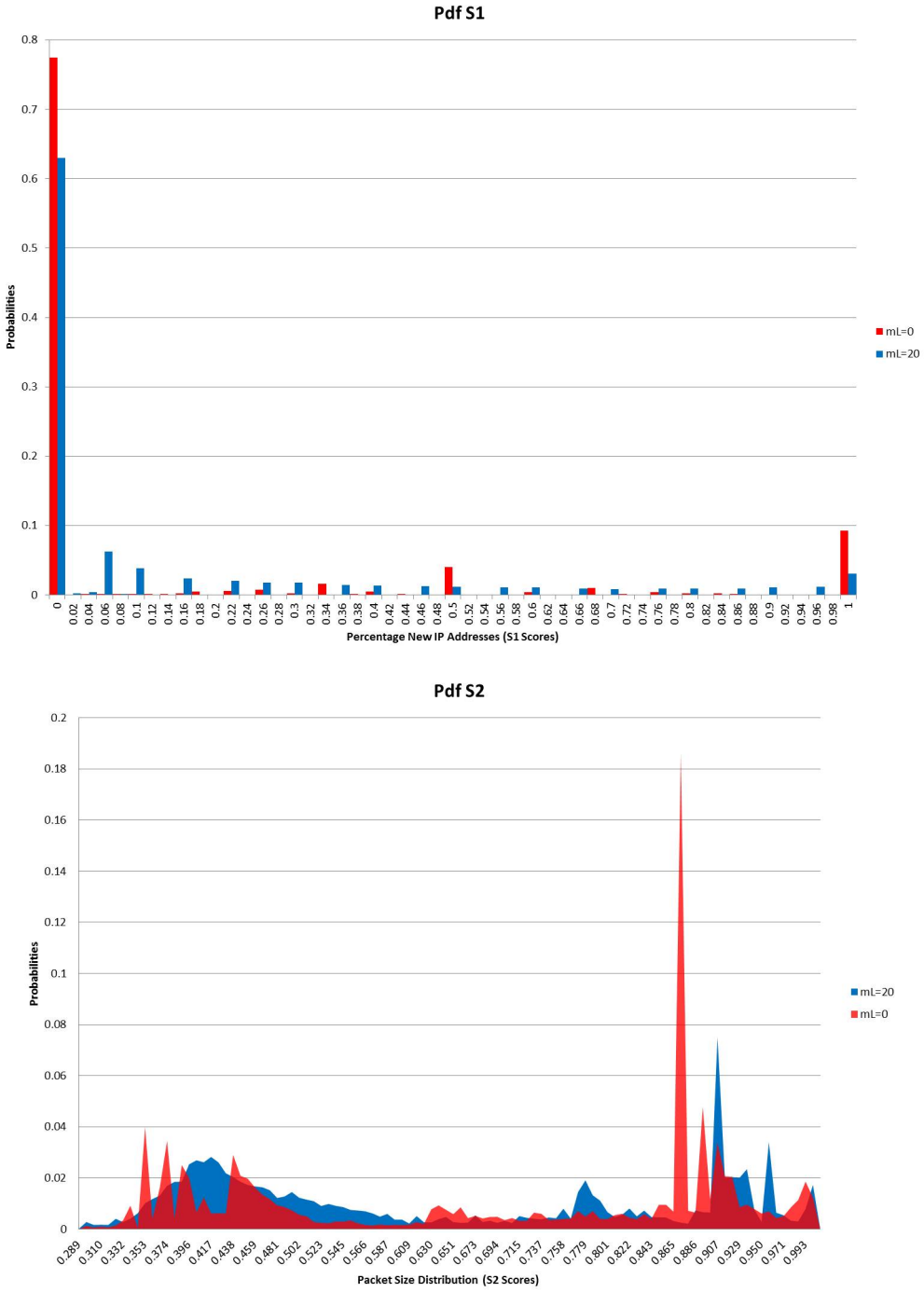


Figure 5.2: N-S Probability Distribution

phase consists of developing and calibrating the detection system using historical data whereas the testing procedure verifies that the DDoS detection structure functions properly under real conditions. If necessary, the detection system should be re-calibrated using the feedback from testing. Training dataset are exclusively composed of legitimate data packets whereas testing datasets contain a mixture of legitimate and attack data packets.

5.2 Supervised Learning Detection Model

The first steps in the detection process include packet sampling and statistic collection as explained earlier. Once traffic statistics are collected, the joint empirical likelihood ratio must be evaluated as part of the training phase. Estimating the joint likelihood ratio requires a 2D probability grid. The procedure to design the grid is explained next.

5.2.1 Training Phase: Designing the Probability Grid

Each dimension of the probability grid corresponds to a traffic statistic. In other words, the grid represents the observation space in the binary detection problem. Recall, N_i represents the percentage of unknown source IP addresses in observation window Δ_i. It is a discrete random variable with sample space $\mathcal{Y}_N = \{0, \frac{1}{m_L}, \frac{2}{m_L}, \ldots, \frac{m_L-1}{m_L}, \frac{1}{m_U}, \frac{2}{m_U}, \ldots, \frac{m_U-1}{m_U}, 1\}$ whereas S_i, the packet size statistic, is a continuous random variables with sample space $\mathcal{Y}_S = [0, 1]$. The observation space therefore is

$$\mathcal{Y} = \{0, \frac{1}{m_L}, \frac{2}{m_L}, \ldots, \frac{m_L-1}{m_L}, \frac{1}{m_U}, \frac{2}{m_U}, \ldots, \frac{m_U-1}{m_U}, 1\} \times [0, 1]. \tag{5.2.1}$$

Without loss of generality, let the horizontal axis represent $\mathcal{Y}_N$ and the vertical axis be $\mathcal{Y}_S$. To evaluate the likelihood ratio, we discretize the observations space $\mathcal{Y}$, that is the 2D grid is partitioned into cells e_j's ($j = 1, 2, \ldots, C$), where $C = C_1 \times C_2$. Let $C1, C_2 \in \mathbb{N}$ be the total number of coordinates on the primary and secondary axis, respectively. Both parameters must be provided by the user. They controls the granularity of each axis and ultimately the granularity of the grid. Let Vectors $\vec{\mathcal{U}}$ and $\vec{\mathcal{V}}$ describe the coordinates of grid cells. Cell coordinates are generated by inverse transform of F_N and F_S, the empirical cumulative distribution functions of random variables N and S. Steps to build the grid can be summarized as follows

1. Sample a dataset containing only *legitimate* data packets into observation windows Δ_i's using the lower and upper bounds $m_L = 20$ and $m_U = 50$.

2. Compute the two traffic scoress, N_i and S_i for each Δ_i, $i = 1, 2, \ldots, M$, using (5.1.1) and (5.1.4).

3. Evaluate the empirical cdfs F_N and F_S using a cumulative histogram.

4. Apply an inverse transform method to invert F_N and compute the horizontal coordinates of each grid cell e_j.

5. Conditioning on the horizontal coordinates, invert F_S to estimate the vertical coordinates of each grid cell e_j.

The last two steps above imply that horizontal cell coordinates correspond to percentiles of the random variable N while vertical cell coordinates are conditional percentiles of the random variable S. Next, the inverse transform method is described in more details. First, for all ω ($0 \leq \omega \leq 1$), define the cdf range of the random variable Λ, with $\Lambda = N, S$ as

$$R_\Lambda = \max F_\Lambda(\omega) - \min F_\Lambda(\omega). \tag{5.2.2}$$

Given $C1$, define the vector $\vec{\mathcal{U}} = [\mathcal{U}_{t_1}, \mathcal{U}_{t_2}, \mathcal{U}_{t_3}, \ldots, \mathcal{U}_{t_{C1}}]$ as the primary axis coordinates. For all ω and $j = 1, 2, \ldots, C1$, t_j's are computed as

$$t_j = \begin{cases} \min F_N(\omega), & j = 1 \\ 1, & j = C1 \\ t_{j-1} + \frac{R_N}{C1}, & 2 \leq j < C1. \end{cases} \tag{5.2.3}$$

Since $t_j = P\left\{N \leq \mathcal{U}_{t_j}\right\}$ for all $j = 1, 2, \ldots, C1$, the primary coordinate axis must satisfy the following equations

$$\mathcal{U}_{t_j} = \begin{cases} \min N_i, & j = 1 \\ \max N_i, & j = C1 \\ F_N^{-1}(t_j), & 2 \leq j < C1 \end{cases} \tag{5.2.4}$$

If $\mathcal{U}_{t_1} \neq 0$, define $\mathcal{U}_{t_0} = 0$ and insert it as the starting point of vector $\vec{\mathcal{U}}$. Also, if $\mathcal{U}_{t_{C1}} \neq 1$, define $\mathcal{U}_{t_{C1+1}} = 1$ and add it as the ending point of $\vec{\mathcal{U}}$. This adjustment is necessary since by definition, $0 \leq N \leq 1$; it guarantees that observation windows with scores $N < \mathcal{U}_{t_1}$ or $N > \mathcal{U}_{t_{C1}}$ will be properly assigned to a grid cell. Consequently the number of coordinates on the primary axis may vary from $C1$ to $C1 + 2$.

Define $\vec{\mathcal{V}}^j = \left[\mathcal{V}^j_{h_1}, \mathcal{V}^j_{h_2}, \mathcal{V}^j_{h_3}, \ldots, \mathcal{V}^j_{h_{C2}}\right]$ as the secondary axis coordinates. The same procedure is applied to compute the secondary coordinates, only this time they are conditioned on the primary axis coordinates, i.e. on the intervals defined by $\vec{\mathcal{U}}$. For each j, $j = 1, 2, \ldots, C1$, let

$$\mathcal{V}^j_{h_i} = \begin{cases} \min\left\{S_i \,|\, \mathcal{U}_{t_{j-1}} \leq N < \mathcal{U}_{t_j}\right\}, & i = 1 \\ \max\left\{S_i \,|\, \mathcal{U}_{t_{j-1}} \leq N < \mathcal{U}_{t_j}\right\}, & i = C2 \\ F^{-1}_{S\,|\,\mathcal{U}_{t_{j-1}} \leq N < \mathcal{U}_{t_j}}(h_i), & 2 \leq i < C2 \end{cases} \tag{5.2.5}$$

The variables h_i's are computed as described in (5.2.3), using F_S instead of F_N. Like $\vec{\mathcal{U}}$, if necessary the starting and ending points of $\vec{\mathcal{V}}$ are modified to be $\mathcal{V}_{t_0} = 0$ and $\mathcal{V}_{t_{C1+1}} = 1$, respectively. Axis coordinates must be unique otherwise zero width column/row intervals will be created. In practice if $\mathcal{U}_{t_i} = \mathcal{U}_{t_j} = \mathcal{U}$ for $i \neq j$ then only $\mathcal{U}_{t_i} = \mathcal{U}$ is represented in the $\vec{\mathcal{U}}$.

The estimation procedure results in a grid partitioned into a total of $C = C1 \times C2$ non-overlapping rectangles. Due to the inverse transform method and conditioning, grid cells may have uneven proportions. Any cell e_j is uniquely described by two intervals: $\left[\mathcal{U}_{t_{j-1}}, \mathcal{U}_{t_j}\right] \times \left[\mathcal{V}^j_{h_{i-1}}, \mathcal{V}^j_{h_i}\right]$, where $\mathcal{U}_{t_{j-1}} < \mathcal{U}_{t_j}$ and $\mathcal{V}^j_{h_{i-1}} < \mathcal{V}^j_{h_i}$ Note that the constant C_1 and C_2 are not necessarily the same, therefore the grid is not always square. The larger $C1$ and $C2$, the more precise the grid is. Since the detection model monitors only two statistics, rectangular-shaped cells are appropriate. As the number of random variables increases, the dimension of the grid increases and more sophisticated shapes can be used. Although we illustrated the procedure by choosing the horizontal as the primary axis, either axis, vertical or horizontal, can be designated as primary. The shape of the resulting grid might change depending on this choice, however, the likelihood ratio estimation is unaffected.

Once the probability grid is obtained, the next step consists of estimating the joint likelihood ratio corresponding to each grid cell e_j, $j = 1, 2, \ldots, C$. Although the probability grid is built using only legitimate data packets, estimating the joint likelihood ratio requires both legitimate and attack data packets. The statistics, as described in Section 5.1.1, may not be enough to discriminate the two types of data packets. The next section describes the procedure to improve traffic statistics through the use of a weighting system. Joint likelihood ratios can then be computed using the new weighted statistic.

5.2.2 Training Phase: Weighted Statistic and Likelihood Ratios Estimation

The weight function applies to the packet size statistic, i.e. S_i, and is defined in terms of the number of attack data packets contained in each observation window. Let $\tau_i = \frac{A_i}{m_i}$ be the percentage of attack data packets contained in the observation window Δ_i and let $\mathfrak{T}$ denote the maximum threat level allowed per observation window. There are two types of observation windows, as defined below.

Definition 13. *An observation window Δ_i is a threat if $\tau_i > \mathfrak{T}$ and a non-threat if $\tau_i \leq \mathfrak{T}$.*

Recall in most DDoS flooding attacks, data packets are not malformed (as opposed to logic DDoS attacks). Depending on the capacities of the target network, "few" attack data packets can be processed without crashing the target server. In that case, $\mathfrak{T}$ can be a positive constant. Otherwise, $\mathfrak{T} = 0$, implying that an observation window is considered as a threat (and should be removed) if it contains at least one attack data.

By definition, as τ_i increases, i.e. $\tau_i \to 1$, the corresponding observation window Δ_i should yield traffic statistics close to one, i.e. $N_i, S_i \to 1$. Measurements from non-threat observation windows must remain close to zero. The traffic statistic formulas given in (5.1.1) and (5.1.4) do not always reflect that property. To enforce it, we propose to conditionally weight the packet size statistic S. Assume for now that the percentage of unknown source IP addresses (N) remains identical to (5.1.1).

Let $\bar{S}_i$ be the weighted and normalized ratio of mean to standard deviation of packets sizes in Δ_i.

$$\bar{S}_i = \frac{\bar{Z}_i}{\bar{Z}_i + \kappa_i \times \sigma_i} \ , \tag{5.2.6}$$

where Z_i is the arithmetic average of packet sizes in Δ_i. The weights in (5.2.6) are computed as follows

$$\kappa_i = \begin{cases} \bar{X} & \text{if } A_i = 0 \\ \frac{1}{\tau_i} = \frac{m_i}{A_i} & \text{if } A_i > 0, \end{cases} \tag{5.2.7}$$

where $\bar{X}$ is the average of the mean to standard deviation random variable, X_i (see Formula (5.1.2)). When an observation window Δ_i does not contain any attack data packets ($A_i = 0$), the weighting function (5.2.7) causes the packet size statistic to decrease and be close to zero, as desired. Otherwise, S_i increases proportionally to the number of attack data packets contained in Δ_i. The more attack data packets Δ_i contains, the closer $\bar{S}_i$ is to one. Note that the training phase uses only historical data, therefore A_i and subsequently τ_i can be evaluated exactly. Figure 5.3 illustrates the effect of the weight function for legitimate data packet only. The first graph depict the distribution of the packet size random variable without any weighting system. All observation windows yields a packet size statistic greater than 0.2889 despite the fact that they do not contain any attack data packets. Using the weight $\kappa_i = 10$ causes the distribution to be concentrated towards zero as shown in the second graph. The third graph illustrates the weighted statistic with $\kappa_i = 35$, which corresponding to the 95% percentile of the random variable X_i described in (5.1.2).

We conclude that the weighted packet size statistic performs better than (5.1.4) to discriminate legitimate from attack data packets. Subsequently, the joint likelihood ratios can be estimated more accurately. The binary hypothesis problem consists of two hypothesis: H_0 is the hypothesis that only legitimate traffic is observed whereas H_1 denotes the hypothesis that an attack is taking place. Define R_i as the aggregated score of each observation window Δ_i, $i = 1, 2, \ldots, M$. Assuming the random variables N_i and R_i are independent, the pdfs satisfy $f_R(y) = f_N(y) \times f_S(y)$. For each grid cell e_j, $j = 1, 2, \ldots, C$, the joint likelihood ratio is defined as

$$\mathbb{L}(y) = \frac{f_{R|H_1}(y)}{f_{R|H_0}(y)} \tag{5.2.8}$$

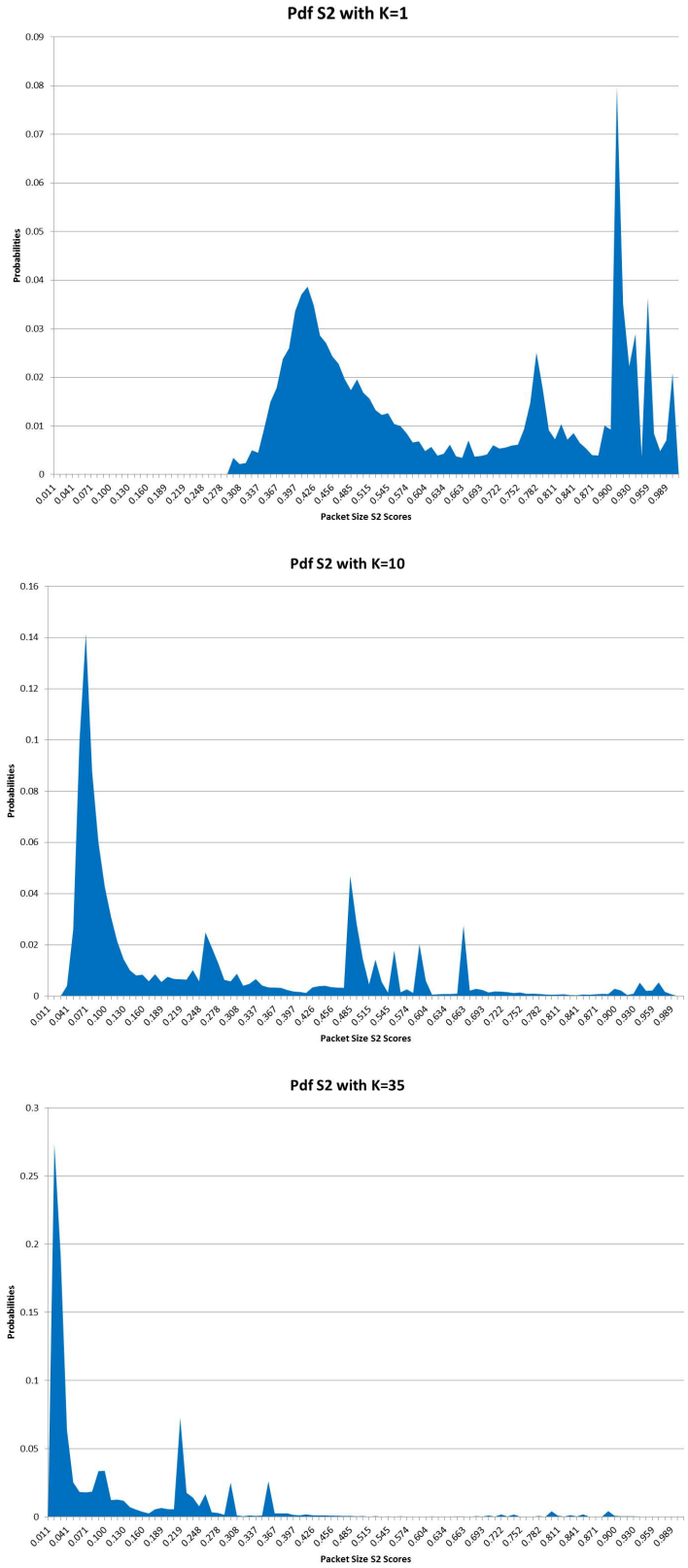


Figure 5.3: Weighted Packet Size Statistic

The procedure to estimate the pdfs $f_{R|H_1}(y)$ and $f_{R|H_0}(y)$ can be summarized as follows.

1. Sample a dataset containing both legitimate and attack data packets into observation windows Δ_i's with parameters $m_L = 20$ and $m_U = 50$.

2. Compute the two traffic statistics, N_i and S_i for each Δ_i, $i = 1, 2, \ldots, M$, using (5.1.1) and (5.2.6), respectively.

3. Classify each Δ_i into the probability grid using N_i and S_i as the horizontal and vertical coordinates, respectively.

4. Estimate the empirical joint pdf $f_{R|H_1}(y)$ as the percentage of threat observation windows that falls into a particular cell e_j. Similarly, $f_{R|H_0}(y)$ corresponds to the percentage of non-threat observation windows in e_j [59].

5. For each grid cell e_j, compute the joint likelihood ratio using Formula (5.2.8).

Note that any observation window Δ_i can be classified into one (and only one) cell e_j. Furthermore, one attack dataset cannot be representative of every possible DDoS attack scenario. We suggest replicating the procedure described above with various attack parameters to improve the robustness of the detection system.

The last step in the training process is to partition the probability grid in two mutually exclusive classes, the acceptance and rejection regions. This is achieved by performing multiple Neyman-Pearson tests based on likelihood ratios estimated from the empirical pdf of the random variables $\{N_i\}$ and $\{S_i\}$ (see Formulas (5.1.1) and (5.2.6)).

5.2.3 Training Phase: Defining the Acceptance and Rejection Region

The main step in the detection procedure consists of partitioning the observation space into two mutually exclusive subsets, the rejection $\mathcal{R}$ and acceptance $\mathcal{A}$ regions. An observation window Δ_i is said to be a threat if $R_i \in \mathcal{R}$. On the contrary, label Δ_i as non-threat if $R_i \in \mathcal{A}$. Recall R_i denotes the aggregated sensor response for Δ_i, obtained by combining the two statistics S_i and N_i. The binary detection problem comprises the set of states $\mathcal{X} = \{0, 1\}$, and the observation space $\mathcal{Y} = \mathcal{Y}_N \times \mathcal{Y}_S$, described in (5.2.1). Section 5.2.1 presents the procedure to discretize the observations space $\mathcal{Y}$ into cells using an inverse transform method. The likelihood ratio is evaluated for each of these cells. The Neyman-Pearson criteria suggests rejecting the null hypothesis if the likelihood ratio exceeds a threshold. The threshold η between $\mathcal{R}$ and $\mathcal{A}$ must be set such that the false positive rate is minimized while the false negative rate remains under a fixed threshold δ. Given the rejection region $\mathcal{R} = \{Y_i \,|\, \mathbb{L}(y) \geq \eta\}$, the type I and type II error probabilities can be

defined as

$$P_{FA} = \int_{\mathcal{R}} f_{Y|H_0}(y)dy \tag{5.2.9}$$

$$P_{FC} = = 1 - \int_{\mathcal{R}} f_{Y|H_1}(y)dy \tag{5.2.10}$$

Since no closed form expression exists for the pdf of R_i, the threshold η must be numerically evaluated . Recall the vector describing the primary axis coordinates is $\vec{\mathcal{U}} = [\mathcal{U}_{t_1}, \mathcal{U}_{t_2}, \mathcal{U}_{t_3}, \ldots, \mathcal{U}_{t_{C1}}]$. Given $S_i \in (\mathcal{U}_{c-1}, \mathcal{U}_c]$, the Neyman-Pearson threshold is computed by numerically solving the following optimization model:

$$\begin{aligned} \text{maximize} \quad & P\{\mathbb{L}(y) < \eta_c \,|\, y \in \mathcal{A}\} \\ \text{subject to} \quad & P\{L(y) < \eta_c \,|\, y \in \mathcal{R}\} \leq \delta \\ & 0 \leq \eta_c \leq 1 \end{aligned} \tag{5.2.11}$$

The objective function consists of maximizing the "true negative" rate subject to the constraint that the false negative rate remains under the threshold δ. Since both probabilities are increasing functions of η_c,

$$\eta_c^* = \max\left\{\eta_c' : P\{L(y) < \eta_c\} \leq \delta\right\} \tag{5.2.12}$$

Let $\mathcal{R}_c$ and $\mathcal{A}_c$ denote the rejection and acceptance regions when S_i falls into the interval $(\mathcal{U}_{c-1}, \mathcal{U}_c]$. The decision variable η_c in (5.2.11) represents the threshold between $\mathcal{R}_c$ and $\mathcal{A}_c$. The optimization model in (5.2.11) is solved for each interval $(\mathcal{U}_{c-1}, \mathcal{U}_c]$. The thresholds $\{\eta_c\}$ may vary depending on the interval $(\mathcal{U}_{c-1}, \mathcal{U}_c]$ where S_i falls. In fact, as $\S_i \to 1$, η_c gets smaller since the probability of a threat gets higher. Note that the likelihood ratio in (5.2.11) is computed based on pdf of the percentage of unknown source IP addresses. Consider an observation window $Delta_i$ containing m_i data packets. Assume data packets arrive according to the marked Poisson described in Section 2.1, and let A_i be the number of attack data packets and U_i the total number of legitimate data packets with unknown source IP addresses in $Delta_i$, respectively. Also define Q_i as the number of packet flows with unknown source IP addresses in Δ_i. If collecting m_i data packets requires t_i time units, then the independent increments and thinning properties of the Poisson process allows us to deduce that Q_i is a Poisson random variable with rate $\zeta\varepsilon t_i$ and $L = (\zeta\varepsilon + \lambda_\varsigma)t_i$, under H_0 and H_1, respectively. Proposition shows how to compute the likelihood ratio in (5.2.11).

Proposition 15. *Set the maximum threat level as $\mathfrak{T} = 0$ (see Definition 13). The likelihood ratio of N_i, the random variable corresponding to the percentage of unknown source IP addresses, can be computed as*

$$\mathbb{L}(y) = \frac{\sum_{a=1}^{m_i} \mathbb{P}(U_i = ym_i - a)\mathbb{P}(A_i = a)}{1 - e^{-\lambda_\varsigma t_i}} \times \frac{e^{-\lambda_\varsigma t_i}}{\mathbb{P}(U_i = ym_i)}, \tag{5.2.13}$$

where

$$\begin{aligned}\mathbb{P}\{U_i = u\} &= \sum_{q=1}^{u} \frac{(\zeta\varepsilon t_i)^q e^{-\zeta\varepsilon t_i}}{q!} \times P\left\{\sum_{j=1}^{q} K_j = u - q\right\} \\ \mathbb{P}\{A_i = a\} &= \sum_{q=1}^{a} \frac{(\lambda\varsigma t_i)^q e^{-\lambda\varsigma t_i}}{q!} \times P\left\{\sum_{j=1}^{q} K_j = a - q\right\}.\end{aligned} \tag{5.2.14}$$

K_j are iid mixture random variable and for any j,

$$\begin{aligned}\mathbb{P}\{K_j = k\} &= q\left(\sum_{i=1}^{\lfloor c_1\rfloor - 1} p_i \mathbb{1}\{k = x_i\}\right)\mathbb{1}\{1 \le k \le \lfloor c_1\rfloor - 1\} + q_1 \frac{\alpha_1 c_1^{\alpha_1}}{k^{\alpha_1+1}}\mathbb{1}\{k \ge \lfloor c_1\rfloor\} \\ &+ q_2 \frac{\alpha_2 c_2^{\alpha_2}}{k^{\alpha_2+1}}\mathbb{1}\{k \ge \lfloor c_2\rfloor\}\end{aligned}$$

Proof. Recall τ_i is the assessed threat level of the observation window δ_i. Let $\mathfrak{T} = 0$, then the likelihood ratio is then

$$\begin{aligned}\mathbb{L}(y) &= \frac{\mathbb{P}\{N_i = y|\tau_i > \mathfrak{T}\}}{\mathbb{P}\{N_i = y|\tau_i \le \mathfrak{T}\}} = \frac{\mathbb{P}\{U_i + A_i = ym_i|\tau_i > \mathfrak{T}\}}{\mathbb{P}\{U_i + A_i = ym_i|\tau_i \le \mathfrak{T}\}} \\ &\overset{(a)}{=} \frac{\mathbb{P}\{U_i + A_i = ym_i|A_i > 0\}}{\mathbb{P}\{U_i + A_i = ym_i|A_i = 0\}} \\ &= \frac{\mathbb{P}\{U_i + A_i = ym_i, A_i > 0\}}{\mathbb{P}\{U_i + A_i = ym_i, A_i = 0\}} \times \frac{\mathbb{P}\{A_i = 0\}}{\mathbb{P}\{A_i > 0\}} \\ &\overset{(b)}{=} \frac{\sum_{a=1}^{m_i} \mathbb{P}(U_i = ym_i - a)\mathbb{P}(A_i = a)}{1 - e^{-\lambda\varsigma t_i}} \times \frac{e^{-\lambda\varsigma t_i}}{\mathbb{P}(U_i = ym_i)}\end{aligned}$$

Equality (a) follows since τ_i is proportional to A_i, the number of attack data packets contained in δ_i, and $A_i > 0$. Equality (b) holds since the number of attack packet flows with unknown source IP addresses is a Poisson random variable with rate $\lambda\varsigma t_i$ and $A_i = 0$ implies that there are no arrival of attack flow during t_i. Given that U_i and A_i are compound random variables then Formula (5.2.14) follow and Formula (5.2.15) is deduced from the definition of the distribution function of cluster sizes (see Section 2.1). □

5.2.4 Testing Phase

The testing phase, as the name indicates, consists of using a dataset representative of the target network being under attack to evaluate the detection system. All techniques from the training phase, including the data sampling, traffic statistics estimation, and classification into the 2D grid, are reprised here except for the weighting method in Section 5.2.2. In fact, the training phase uses

historical data that have attack packets clearly identified. However, that is not the case in the testing phase since real-life attack data packets are not labeled as such. Following this reasoning, τ_i the assessed threat value of Δ_i is unknown until m_i data packets have been collected. Therefore, adjustments in the weighting function of the packet size statistic (S_i) become necessary.

Let $\tilde{S}_i$ denote the packet size statistic in the testing phase. $\tilde{S}_i$ is defined as in (5.2.6) with weight $\tilde{\kappa}_i$. Define γ_i as the number of *distinct* source IP addresses that fall into observation window Δ_i, then

$$\tilde{\kappa}_i = \begin{cases} \bar{X} & \text{if } \gamma_i < H \\ \frac{1}{\gamma_i} & \text{else,} \end{cases} \tag{5.2.15}$$

where H is a positive constant. If $\gamma_i < H$, we assume the corresponding observation window is more likely a non-threat. Therefore, the weight $\tilde{\kappa}_i$ is applied to decrease the value of S_i so that $S_i \to 0$. Else, if $\gamma_i \geq H$, S_i is increased proportionally to γ_i.

Setting the parameter H in (5.2.15) depends on the probability $P\{A_i \geq 1 \,|\, \gamma_i \geq H\}$, which is the conditional probability that an observation window Δ_i contains at least an attack data packet given that it has at least H distinct unknown source IP addresses. H must be set such that the probability $P\{A_i \geq 1 \,|\, \gamma_i \geq H\}$ is maximized. Let ν and λ denote the total number of spoofed IP addresses and the intensity of the DDoS attack, respectively. As ν and λ increase, the more likely it is that a threat observation window contains more than H distinct unknown source IP addresses. In the following, assume all attack packet flows have unknown source IP addresses. Proposition 16 gives the expression of the conditional probability $P\{A_i \geq 1 \,|\, \gamma_i \geq H\}$.

Proposition 16. *Consider an observation window Δ_i that contains m_i data packets. Let γ_i^L and γ_i^A be the number of distinct unknown source IP addresses from legitimate and attack traffic, respectively.*

$$\mathbb{P}\{A_i \geq 1 \,|\, \gamma_i \geq H\} = 1 - \frac{1 - e^{-\zeta\varepsilon t_i} - \sum_{x=1}^{H-1} P\left\{\gamma_i^L = x\right\}}{1 - e^{-\zeta\varepsilon t_i} * e^{-\lambda t_i} - \sum_{h=1}^{H-1} P\{\gamma_i = h\}}\,, \tag{5.2.16}$$

where

$$\begin{aligned} \mathbb{P}\left\{\gamma_i^L = x\right\} &= \frac{n!}{(n-x)!\, n^x}\left(\frac{(\zeta\varepsilon t_i)^x\, e^{-\zeta\varepsilon t_i}}{x!} + \sum_{u=x+1}^{m_i} \frac{(u-1)!}{(x-1)!\,(u-x)!\,x!}\right. \\ &\times \left.\frac{(\lambda_L \beta t_i)^u\, e^{-\lambda_L \beta t_i}}{u!}\right) \end{aligned} \tag{5.2.17}$$

$$\begin{aligned}\mathbb{P}\{\gamma_i = h\} &= \sum_{y=0}^{m_i}\sum_{x=0}^{h}\left(\sum_{z=0}^{y}\mathbb{P}\left\{\gamma_i^L = h-x \,|\, u_i = z\right\}\frac{(\zeta\varepsilon t_i)^z\, e^{-\zeta\varepsilon t_i}}{z!}\frac{(\lambda t_i)^{y-z}\, e^{-\lambda t_i}}{(y-z)!}\right)\\ &\times \left(\sum_{w=0}^{y}\mathbb{P}\left\{\gamma_i^A = x \,|\, a_i = w\right\}\frac{(\lambda t_i)^w\, e^{-\lambda t_i}}{w!}\frac{(\zeta\varepsilon t_i)^{y-w}\, e^{-\zeta\varepsilon t_i}}{(y-w)!}\right) \qquad (5.2.18)\end{aligned}$$

$$\mathbb{P}\left\{\gamma_i^A = x \,|\, A_i = w\right\} = \begin{cases} 0 & \text{if } x > w \\ \frac{\nu!}{(\nu-w)!\,\nu^w} & \text{if } x = w \\ \frac{(w-1)!}{(x-1)!\,(w-x)!\,x!} \times \frac{\nu!}{(\nu-w)!\,\nu^w} & \text{else} \end{cases} \qquad (5.2.19)$$

Note that $\mathbb{P}\left\{\gamma_i^L = h-x \,|\, U_i = z\right\}$ in (5.2.18) is computed as in (5.2.19) with $\nu = n$, $w = z$ and $x = h - x$. Moreover, $\mathbb{P}\left\{\gamma_i^L = x\right\} = 0$ for all $x > n$. Also $\mathbb{P}\left\{\gamma_i^A = z\right\} = 0$ for all $z > \nu$. Therefore, threshold H must satisfy the conditions $H \leq \nu - 2$ and $H \leq n - 1$.

Proof.

$$\begin{aligned}\mathbb{P}\{A_i \geq 1 \,|\, \gamma_i \geq H\} &= 1 - \mathbb{P}\{A_i = 0 \,|\, \gamma_i \geq H\} = 1 - \frac{\mathbb{P}\{A_i = 0 \cap \gamma_i \geq H\}}{\mathbb{P}\{\gamma_i \geq H\}}\\ &= 1 - \frac{\mathbb{P}\left\{\gamma_i^L \geq H\right\}}{\mathbb{P}\{\gamma_i \geq H\}} = 1 - \frac{1 - \sum_{x=0}^{H-1}\mathbb{P}\left\{\gamma_i^L = x\right\}}{1 - \sum_{h=0}^{H-1}\mathbb{P}\{\gamma_i = h\}}\\ &= 1 - \frac{1 - e^{-\zeta\varepsilon t_i} - \sum_{x=1}^{H-1}\mathbb{P}\left\{\gamma_i^L = x\right\}}{1 - e^{-\zeta\varepsilon t_i} \times e^{-\lambda t_i} - \sum_{h=1}^{H-1}\mathbb{P}\{\gamma_i = h\}}\end{aligned}$$

The third equality holds since by definition $\gamma_i = \gamma_i^A + \gamma_i^L$ and $\gamma_i^A = 0$ if $A_i = 0$. Moreover,
$\mathbb{P}\left\{\gamma_i^L = 0\right\} = \mathbb{P}\{U_i = 0\} = e^{-\zeta\varepsilon t_i}$, and
$\mathbb{P}\{\gamma_i = 0\} = \mathbb{P}\{U_i = 0\} \times \mathbb{P}\{A_i = 0\} = e^{-\zeta\varepsilon t_i} \times e^{-\lambda t_i}$
We now give the proofs for (5.2.17) and (5.2.18).

$$\begin{aligned}\mathbb{P}\left\{\gamma_i^L = x\right\} &= \mathbb{P}\left\{\gamma_i^L = x \cap u_i < x\right\} + P\left\{\gamma_i^L = x \cap u_i = x\right\} + P\left\{\gamma_i^L = x \cap u_i > x\right\}\\ &= 0 + \mathbb{P}\left\{\gamma_i^L = x \,|\, u_i = x\right\}\mathbb{P}\{u_i = x\} + \mathbb{P}\left\{\gamma_i^L = x \,|\, u_i > x\right\}\mathbb{P}\{u_i > x\}\\ &= \mathbb{P}\left\{\gamma_i^L = x \,|\, u_i = x\right\}\mathbb{P}\{u_i = x\} + \sum_{u=x+1}^{m_i}\mathbb{P}\left\{\gamma_i^L = x \,|\, u_i = u\right\}\mathbb{P}\{U_i = u\}\end{aligned}$$

$\mathbb{P}\left\{\gamma_i^L = x \,|\, U_i = x\right\}$ is the probability that Δ_i contains U_i legitimate data packets with unknown source IP addresses and all u_i packets have distinct source IP addresses. $\mathbb{P}\left\{\gamma_i^L = x \,|\, U_i = u\right\}$ is the probability that Δ_i contains $U_i = u$ legitimate data packets. These u data packets can be separated

into x subsets where each subset x contains at least one data packet. There are $\binom{u-1}{u-x}$ subsets (each with at least 1 element) of a set containing u elements. All x subsets must have distinct source IP addresses. Selecting x distinct source IP addresses, from the set of legitimate unknown IP, is similar to the "birthday problem" with equally likely days; the probability is $\frac{n!}{(n-x)!\,n^x}$. Moreover, $P\left\{\gamma_i^L = x\right\} = 0$ if $x > n$. Equation (5.2.17) then follows by the thinning property of Poisson processes and also the fact that source IP addresses are independent from legitimate packet arrival times. The proof of $\mathbb{P}\left\{\gamma_i^A = x \,|\, A_i = w\right\}$ is similar to that of $\mathbb{P}\left\{\gamma_i^L = x\right\}$ only with ν instead of n. Also the Poisson distribution rate is λt_i instead of $\zeta\varepsilon t_i$. The proof of (5.2.18) is as follows:

$$\begin{aligned}
\mathbb{P}\{\gamma_i = h\} &= \sum_{y=0}^{m_i} P\{\gamma_i = h \,|\, U_i + A_i = y\}\, \mathbb{P}\{U_i + a_i = y\} \\
&= \sum_{y=0}^{m_i}\sum_{x=0}^{yi} \mathbb{P}\left\{\gamma_i^A + \gamma_i^L = h \,|\, \gamma_i^A = x, U_i + A_i = y\right\} \mathbb{P}\left\{\gamma_i^A = x \,|\, U_i + A_i = y\right\} P\{U_i + A_i = y\} \\
&= \sum_{y=0}^{m_i}\sum_{x=0}^{yi} \mathbb{P}\left\{\gamma_i^L = h - x \,|\, U_i + A_i = y\right\} P\left\{\gamma_i^A = x \,|\, U_i + A_i = y\right\} \mathbb{P}\{U_i + A_i = y\} \\
&= \sum_{y=0}^{m_i}\sum_{x=0}^{h} \left(\sum_{z=0}^{y} P\left\{\gamma_i^L = h - x \,|\, U_i = z\right\} \frac{(\zeta\varepsilon t_i)^z\, e^{-\zeta\varepsilon t_i}}{z!} \frac{(\lambda t_i)^{y-z}\, e^{-\lambda t_i}}{(y-z)!} \right) \\
&\times \left(\sum_{w=0}^{y} \mathbb{P}\left\{\gamma_i^A = x \,|\, A_i = w\right\} \frac{(\lambda t_i)^w\, e^{-\lambda t_i}}{w!} \frac{(\zeta\varepsilon t_i)^{y-w}\, e^{-\zeta\varepsilon t_i}}{(y-w)!} \right)
\end{aligned}$$

The last equality follows by independence of the legitimate and attack packet flows processes. □

From Proposition 16 we can deduce the expected number of distinct unknown source IP addresses in an observation window Δ_i, given that it contains at least one attack data packet. This assumes the maximum threat level is $\mathfrak{T} = 0$.

Corollary 2. *Consider an observation window Δ_i that contains m_i data packets. If collecting m_i data packets requires t_i time units, then*

$$\mathbb{E}\left[\gamma_i \,|\, A_i \geq 1\right] = \sum_{x=1}^{m_i} x \, \frac{\sum_{y=1}^{x} \mathbb{P}\left\{\gamma_i^L = x - y\right\} \mathbb{P}\left\{\gamma_i^A = y\right\}}{\sum_{j=1}^{m_i} \frac{(\lambda t_i)^j\, e^{-\lambda t_i}}{j!}}, \tag{5.2.20}$$

where

$$\mathbb{P}\left\{\gamma_i^A = z\right\} = \frac{\nu!}{(\nu - z)!\,\nu^z} \left(\frac{(\lambda t_i)^z\, e^{-\lambda t_i}}{z!} + \sum_{a=z+1}^{m_i} \frac{(a-1)!}{(z-1)!\,(a-z)!\,z!} \times \frac{(\lambda t_i)^a\, e^{-\lambda t_i}}{a!} \right) \tag{5.2.21}$$

The probability $\mathbb{P}\left\{\gamma_i^L = x - y\right\}$ in Equation (5.2.20) is computed as in (5.2.17).

Proof.

$$
\begin{aligned}
E\{\gamma_i \mid A_i \geq 1\} &= \sum_{x=1}^{m_i} x\, \mathbb{P}\{\gamma_i \mid A_i \geq 1\} \\
&= \sum_{x=1}^{m_i} x\, \frac{\mathbb{P}\{\gamma_i \cap a_i \geq 1\}}{\mathbb{P}\{a_i \geq 1\}} = \sum_{x=1}^{m_i} x\, \frac{\mathbb{P}\left\{\gamma_i^A + \gamma_i^L = x, \gamma_i^A \geq 1\right\}}{\mathbb{P}\{a_i \geq 1\}} \\
&= \sum_{x=1}^{m_i} x\, \frac{\sum_{y=1}^{x} \mathbb{P}\left\{\gamma_i^L = x - y\right\} \mathbb{P}\left\{\gamma_i^A = y\right\}}{\sum_{j=1}^{m_i} \frac{(\lambda t_i)^j\, e^{-\lambda t_i}}{j!}}
\end{aligned}
$$

□

Assume $\lambda > \lambda_L$. As λ increases and , $\mathbb{E}\{\gamma_i \mid A_i \geq 1\} \to \nu$.

The following section presents some numerical results based on the MIT dataset. Unless otherwise specified, the maximum threat level (see Definition 13) is set to $\mathfrak{T} = 0$.

5.3 Simulation Model and Computational Results

As mentioned earlier, legitimate data packets are real network traces collected from the MIT lincoln laboratory network. Attack traffic is simulated, allowing us to vary parameters such as the rate, the duration, and the start time. The next section describes how training/testing datasets are obtained.

5.3.1 Training/ Testing Datasets

Constant rate DDoS are the most common [133, 82], therefore attack data packets are generated according to an homogeneous Poisson process with rate λ. Let d denote the duration in time units. The spoofed source IP library ($SSIL$) is artificially designed and each attack data packet is assigned an IP address at random, following a uniform distribution [18, 35]. Using a standard uniform distribution means the master attacker doesn't have a preference over the spoofed IP addresses he has created. Unless otherwise specified, our experiments use $|\nu| = 8$; Du and Abe [29] set $|\nu| > 50$. Note that the larger ν, the easier the detection problem, especially if we monitor the percentage of distinct unknown source IP addresses. Let Π denote the proportion of attack data packet with respect to the total number of data packets. Π can be controlled through the attack parameters d and λ. Shon and Moon [113] set $\Pi = 1 - 1.5\%$ while Carl et al. [18] let $\Pi = 20\%$. Du and Abe [29] state that DDoS attacks are effectively detected if $\Pi = 40\%$. In all following simulations, Π ranges from 2 to 35%. Some examples of training datasets are given in Table 5.4.

Table 5.4: Training Data with DDoS Data Packets

Dataset	Start Time (s)	Duration d (s)	Rate λ (pps)	Packet Size (bytes)	Π (%)
W1-M-DDoS-1	6	3600	10	40	5.30
W1-M-DDoS-2	70000	1800	160	1500	30.62
W1-M-DDoS-3	17200	300	500	576	18.72
W1-Th-DDoS-2	3500	1800	10	40	1.96

Start time indicates when the first attack packet arrives at the control router. Given that the arrival rate of legitimate data packets varies with time (see Figure 2.3), the attack start time is an important parameter. In fact, a DDoS attack is more easily detected if there are less legitimate traffic that arrive at the control router at the same time.

5.3.2 Probability Grid

Given the W1-M dataset, Figure 5.4 illustrates a probability grid with cdf steps set to $C1 = C2 = 10$ and sampling parameters of $m_L = 20$ and $m_U = 50$. Recall, W1-M contains only legitimate data packets. The likelihood ratio is estimated using a combination of legitimate data packets from W1-M and attack data packets W1-M-DDoS-2; see Table 5.4 for the DDoS parameters. Attack and legitimate datasets are combined based on the arrival time of data packets. In Figure 5.4, the

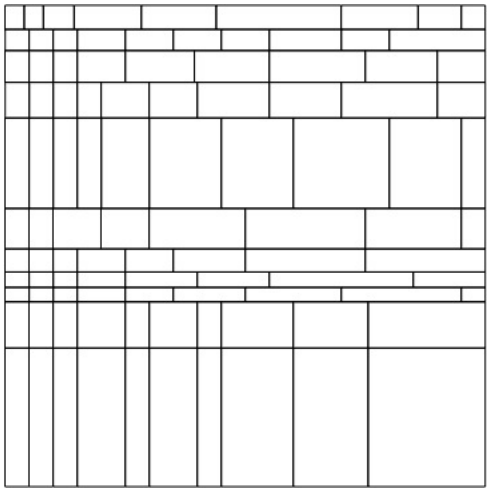


Figure 5.4: Probability Grid with Acceptance and Rejection Regions

maximum threat level is $\mathfrak{T} = 0$, meaning that observation windows are considered to be a threat if they contain at least one attack data packet. It is important to emphasize that a single dataset cannot possibly represent all cases of attacks and determine for certain which grid cells belong to the acceptance/rejection regions. Replicating the likelihood ratio estimation with different attack

datasets is an efficient way of overcoming this limitation. Besides all DDoS parameters, there are other factors that can influence the shape of the acceptance and rejection regions. These include the sampling parameters m_L and m_U, the cdf steps $C1$ and $C2$, and the maximum threat level $\mathfrak{T}$.

5.3.3 Effect of Upper Bound on Packet Sampling

Simulations showed that truncating the data collection process using the upper bound m_U induces values such that legitimate data behave like attack data packets. In other words, given that an observation window Δ_i of size m_i contains m_i legitimate data packets and yields statistics $N_i = S_i = 1$, the conditional probability that $m_i = m_U$ is one. Table 5.5 summarizes the probability of event $\{N = 1, S = 1\}$ for various lower and upper bounds, m_L and m_U. Data are shown for the W1-M dataset.

Table 5.5: Probability of Event $\{N = 1, S = 1\}$

m_L	m_U	$P\{N = 1, S = 1\}$
5	20	0.0035
10	20	0.0055
20	20	0.0084
20	35	0.0031
20	50	0.0015

Entries in Table 5.5 indicate that for a fixed upper bound m_U, the probability $P\{N = 1, S = 1\}$ increases as the lower bound m_L increases, . Moreover, the larger m_U, the lower $P\{N = 1, S = 1\}$. The probability distribution function of event $\{N = 1, S = 1\}$ could not be formally characterized. Distribution fitting with the geometric (G) and normal (N) distributions did not yield any conclusive results, as shown in Table 5.6. A chi-square goodness-of-fit test is performed and the null hypothesis is rejected at 5% significance level for all but one.

Table 5.6: P-Values

Sampling	Distribution	Chi Stat	P-value
$m_L = 10$, $m_U = 20$	$G(0.0055)$	67.76	1.9345e-15
$m_L = 10$, $m_U = 20$	$N(175.16, 454.03)$	86.43	1.4506e-20
$m_L = 20$, $m_U = 50$	$G(0.0015)$	3.2525	0.1967
$m_L = 20$, $m_U = 50$	$N(640.87, 1231.33)$	9.8351	0.0017

Even though the event $\{N = 1, S = 1\}$ occurs with a very small probability, it is still an undesirable result. This can prevented by modifying the weight function on the packet size random variable S_i. Since all data packets have the same size then $\sigma_i = 0$, an adjustment, other than

(5.2.6), is necessary. We propose

$$\widetilde{S}_i = \frac{\bar{Z}_i}{\bar{Z}_i + \bar{X}} \ . \tag{5.3.1}$$

In the next section, we present detection results along with error probabilities. We also compare the proposed detection model to some existing mechanisms.

5.3.4 Model Evaluation: FP and FC rates

The objective of this section is to prove the effectiveness of our detection system. Six testing data, presented in Table 5.7, are created using W3-F as legitimate traffic. Columns 2-5 specify the DDoS attack parameters. For each dataset, the last column gives the proportion of attack data packets with respect to the total number of data packets.

Table 5.7: Testing Data with DDoS

Dataset	Start Time (s)	Duration d (s)	Rate λ (pps)	Packet Size (bytes)	Π (%)
W3-F-DDoS-1	50	1800	10	576	2.16
W3-F-DDoS-2	50	1800	160	576	15.82
W3-F-DDoS-3	50	1800	500	576	35.57
W3-F-DDoS-4	13200	1800	500	576	25.77
W3-F-DDoS-5	50	1800	500	60	15.49
W3-F-DDoS-6	50	300	500	576	2.8

For testing, two probability grids are considered. Grid W1-M-DDoS-1 is trained with a low attack rate (λ = 10pps) and Grid W1-M-DDoS-2 with a medium attack rate (λ = 160pps). Table 5.8 summarizes the probabilities of false negatives and false positives. Results expressed in percentage. The former is estimated as the ratio of number of threat observation windows in the acceptance region $\mathcal{A}$ with respect to the total number of threat observation windows. The second number indicates the false positives rate, given by the ratio of number of non-threat observation windows in the attack region to the total number of non-threat observation windows. All datasets are sampled with parameters $m_L = 20$ and $m_U = 50$ and the packet size statistic is weighted with parameter $H = 1$ (see Formula 5.2.15). $H = 1$ implies that any observation window Δ_i with one or more unknown source IP addresses is considered as a threat.

Although Grid W1-M-DDoS-1 performs well with a 100% detection rate, the false alarm rates are prohibitively high for all testing datasets. Entries in Table 5.8 demonstrate the importance of a well-trained model, which can be achieved by replication of the training phase with different attack parameters. In fact, a DDoS detection system trained with medium rates DDoS, namely Grid W1-M-DDoS-2, effectively detects medium and high rate DDoS with great FP and FN rates while its performance is less desirable with low DDoS rate data (W3-F-DDoS-1). Depending on

Table 5.8: FP-FN Rates with $H = 1$

Dataset — Grid	Grid W1-M-DDoS-1	Grid W1-M-DDoS-2
W3-F-DDoS-1	{0.0; 48.47}	{12.14; 25.99}
W3-F-DDoS-2	{0.0; 48.7}	{0.031; 26.02}
W3-F-DDoS-3	{0.0; 48.7}	{0.0; 26.01}
W3-F-DDoS-4	{0.0; 49.51}	{0.66; 26.92}
W3-F-DDoS-5	{0.0; 48.7}	{0.016; 26.02}
W3-F-DDoS-6	{0.0; 48.6}	{0.0; 25.93}

the application, a 12% false negative rate can be deemed acceptable. On the other hand, a DDoS detection system trained with low rates DDoS, in this case Grid W1-M-DDoS-1, results in a larger attack region. Consequently, it correctly identifies threats from low, medium and high DDoS attacks however, FP rates are very high.

Another experiment is performed with the parameter $H = 3$, meaning observation windows are considered as a threat if they contain at least three distinct unknown source IP addresses. Table 5.9 shows detection performances results from which we conclude that the false positive rates are much lower when H is large.

Table 5.9: FP-FN Rates with $H = 3$

Dataset — Grid	Grid W1-M-DDoS-1	Grid W1-M-DDoS-2
W3-F-DDoS-1	{14.5; 18.37}	{15.14; 3.7}
W3-F-DDoS-2	{0.11; 16.25}	{0.12; 2.49}
W3-F-DDoS-3	{0.0; 18.38}	{0.0; 3.7}

Observe that the false negative rates for Grid W1-M-DDoS-1 have increased. To illustrate why, we plot, in Figure 5.5, the probability mass functions (pmf) of the number of distinct unknown source IP addresses (γ) in observation windows. The first and second graphs show the pmfs for non-threat and threat observation windows, respectively. The means of the distribution in Figure 5.5 are given in Table 5.10.

Table 5.10: γ_i's Expectated Values

Dataset	Mean NT	Mean T
W3-F-DDoS-1	0.41935	6.11619
W3-F-DDoS-2	0.41969	7.7956
W3-F-DDoS-3	0.41972	7.923

While the number of distinct unknown source IP addresses are very similar for non-threat observation windows, it is not the case for threat observation windows. The rate of the DDoS attack

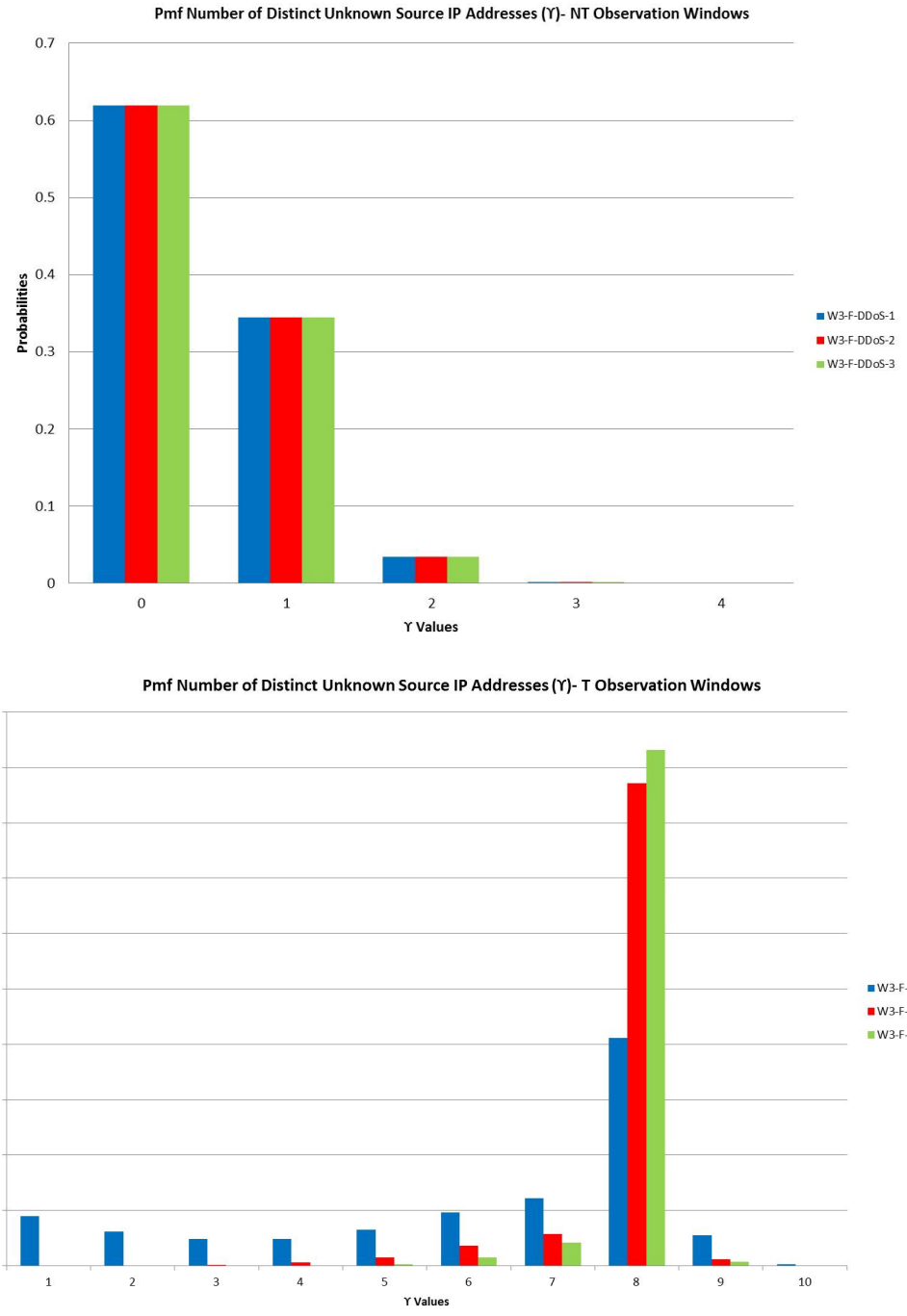


Figure 5.5: Pmf Distinct Unknown Source IP Addresses

clearly has a significant impact on the density functions. As expected, non-threat observation windows contain very few unknown source IP addresses regardless of the DDoS rate λ. On the contrary, as the attack rate λ increases, threat observation windows tend to contain more unknown source IP addresses. Note that the number of distinct unknown source IP addresses in threat observation windows also depends on ν, the number of spoofed IP addresses. Figure 5.5 illustrate simulation results for $\nu = 8$ and 10% of legitimate source IP addresses are unknown.

In summary, large values of H perform well if the DDoS attack has a medium to high rate because $\{\gamma_i\}$, the distinct unknown source IP addresses random variables, take larger values. This emphasizes again that DDoS detection is harder if the attack rate is lower than the legitimate traffic rate. Overall, Grid W1-M-DDoS-2 provides better FP and FN rates compared to Grid W1-M-DDoS-1. Decision-makers should design probability grids with attack/clear regions based on historical attack data but also by taking into consideration their legitimate traffic rate. Clearly, designing DDoS detection systems requires a trade-off between the false positive and negative rates.

Comparing to Other Detection Systems

As emphasized by Mirkovic et al. [83], comparing detection models is not an obvious task mainly due to the disparity of testing conditions. Even testing datasets greatly vary from one research study to another. Below we attempt to compare the performances of our detection models against those presented in [101, 29]. In the following experiments, the number of spoofed source IP addresses is $\nu = 15$, the attack starts at $t = 50s$ and lasts $d = 300$s, and 2% of legitimate source IP addresses are unknown. Peng et al. [101] uses the Auckland dataset as legitimate traffic while we use the 1999 DARPA dataset, the latter is a most realistic dataset because data packets have different sizes. The attack in their experiments has rate 500 kbps. We reproduce the attack rate as a combination of number of packets per second rate and packet sizes as shown in Table 5.11. When $\nu = 15$, their detection system yields 90% (see Table III in [101]) with a detection delay of $127.3s$, while our detection model achieves a 100% detection rate (DR) and a 3% false alarm rate (FR).

Table 5.11: Comparing with IP Filtering [101]

Legitimate	Rate	Size	DR (%) [101]	DR (%)	FP (%)
Auckland	500 kbps	–	90	100	–
DARPA W3-F	870 pps	576	90	100	2.8
DARPA W3-F	335 pps	1500	90	100	2.8
DARPA W3-F	10 pps	576	90	99	2.8

The higher the DDoS attack rate, with respect to legitimate traffic rate, the easier the detection. The last entry line in Table 5.11 shows that even when the attack rate is much lower than 500 kbps, i.e. 10*576=5760 bps, our method can achieve 99% detection rate. Note that the start

time is also an important factor in the detection performance; the reason being the inhomogeneous packet arrival rate.

In comparing our work with the entropy method by Du and Abe [29], we use 50 spoofed source IP addresses. In their case $\nu \geq 50$. In both experiments, all attack data packets are 1500 bytes long, and attack traffics are Poisson processes. Our attack starts at $t = 50$s and 10% of legitimate data packets have unknown source IP addresses. The rate and duration of the DDoS attack are given in the first and second columns of Table 5.12, respectively. Table 5.12 proves that our method can detect very short DDoS attacks.

Table 5.12: Comparing with IP Filtering [29]

Legitimate	Rate (pps)	Duration (s)	DR (%) [29]	DR (%)
DARPA	100	2	90	100
DARPA W3-F	200	1	90	100
DARPA W3-F	300	0.667	95	100
DARPA W3-F	400	0.5	100	100

Tartakovsky et al. [116] propose a sequential change point method where data packets are partitioned by time intervals and size. Although they emphasize that their MNA-CUSUM algorithm does not better than the log-likelihood ratio based CUSUM, it still yields good detection results. We created a testing dataset using the W3-F as legitimate traffic combined with attack data packets that constituted 20% of the total traffic. The testing conditions for the MNA-CUSUM algorithm include sampling data packets each 1s time interval, with three bins each, i.e. packets were classified in bin 1, 2 or 3 if their size was less than 576 bytes, between 576 and 1500 bytes, and finally larger than 1500 bytes. Three type of data packets were monitored, TCP, UDP or ICMP. The historical mean for TCP with size under 576 bytes was estimated as 58.97. The DDoS attack parameters were taken from [116] as $d = 60s$, $\lambda = 3177pps$, $size = 60bytes$. Because the attack rate is so high, our detection model yield a 100% detection rate and MNA-CUSUM had no detection delay, meaning the attack was detected right at the start. For a better evaluation, further simulations should be run with a much lower attack rate.

5.4 Summary

In this chapter, we presented an efficient mechanism for DDoS detection. Contrary to Peng et al. [101] who compare the percentage of unique new source IP addresses to a threshold and build a random sequence, our method finds the percentage of unknown source IP addresses for each predefined observation window. The second statistic is computed as the mean to standard deviation ratio of packet sizes for the same observation window. The two heterogeneous statistics allow

monitoring different aspects of Internet traffic at once to effectively distinguish legitimate from attack data. Detection results are more reliable than using only one statistic.

The advantage of the supervised learning detection model is that the acceptance and rejection are built and maintained off-line. Therefore, new data packets can be observed and classified as a threat or non-threat immediately. In case of malicious data packets, significant damages to the target network can be avoided with early detection.

CHAPTER 6

CONCLUSION

The last chapter provides a brief overview of the research presented in this thesis. Potential directions for future research are also outlined.

6.1 Summary

Using an anomaly-based approach this thesis presents three novel techniques (parametric and empirical) to detect DDoS attacks. We particularly address the detection of flooding attacks at the victim end. The models are shown to be efficient as the detection error probabilities are minimized. The first two systems are parametric that is they assume network traffic can be modeled as a marked Poisson process. Legitimate and attack packet flows arrive at the control router independently of each other, according to a Poisson process. The flow cluster size, that is the number of data packets contained in the packet flow, is a mixture random variable.

The first detection model is formulated as a fixed sample size binary hypothesis testing. The decision making is based on the Neyman-Pearson criteria. The second parametric model is a sequential probability ratio test, that is the sample size is a random variable. The acceptance/rejection boundaries are deduced based on Wald's Fundamental Identity. Although it allows the use of efficient theoretical results, developing parametric models tend to be more challenging given that they address both estimation and detection issues. Moreover, parametric distributions may fail to capture the complex and dynamic nature of the Internet, therefore a third non-parametric detection model is proposed. The likelihood ratio is estimated using a probability grid and Neyman-Pearson threshold is numerically estimated. Our detection systems are designed to detect stealthy attacks including where IP spoofing is used to make attack data packets seem to originate from a known source.

The work presented in this thesis is relevant to several areas other than network security. In fact, there are evidence of heavy tail random variables in the following areas: household income, claim size in insurance, daily log-returns in financial markets, and geophysics. Therefore, the mathematical tools presented here can be useful in all these areas.

6.2 Possible Research Extensions

The work presented here can be extended in several possible ways. DDoS attacks are getting more and more elaborate, therefore detection models must become even more efficient. The first possible extension is to develop additional traffic statistics and eventually extend the detection mechanism to integrate more than two statistic. In particular, we can extend the marked Poisson process presented in Section in order to include the average packet size for packet flows. Monitoring the average packet rate for packet flows [35] or evaluating how often data packets with same header appear at the control router can be also efficient in detecting DDoS attacks.

For networks with substantial traffic, the marked Poisson distribution can be extended to model network traffic at the session level. As session is defined as a set of packet flows. Hohn et al. [55] suggest the log-normal distribution to model session sizes in number of data packets. Theoretical results from detection can then be applied.

Furthermore, the evaluation procedure for our detection models can be improved by testing their behavior in the presence of a flash crowd. The latter is defined as a surge of requests from legitimate users, thus may have a behavior similar to a DDoS attacks.

APPENDIX

Fast Fourier Transform

For any interval $[a, b]$, define the left (LPR), mid-point (MPR), and right (RPR) quadrature rules as

$$\begin{aligned} \text{LPR} : \int_b^a f(x)dx &\simeq f(a) \times (b-a) \\ \text{MPR} : \int_b^a f(x)dx &\simeq f(\frac{a+b}{2}) \times (b-a) \\ \text{RPR} : \int_b^a f(x)dx &\simeq f(\frac{b}{2}) \times (b-a) \end{aligned} \tag{.0.1}$$

Ujevic and Roberts [120] propose the Optimal 3-point rule (O3R) as a linear combination of LPR, MPR and RPR and proves that it provides a better approximation than the classical Simpson rule.

$$\text{O3R} : \int_a^b f(x)dx = \frac{1}{4}(b-a)\left[f(a) + 2f\left(\frac{a+b}{2}\right) + f(b)\right] \tag{.0.2}$$

For $j = 0, 1, \ldots, N$, define the steps as

$$\begin{aligned} u_j &= \frac{1}{b-a}\left(j - \frac{N}{2}\right) \\ m_j &= \frac{1}{2}\left(u_j + u_{j+1}\right) = \frac{1}{b-a}\left(j - \frac{N-1}{2}\right) \end{aligned} \tag{.0.3}$$

Furthermore, define the variables

$$\begin{aligned} L_j &= \frac{2\pi}{b-a}\left(j - \frac{N}{2}\right) \\ M_j &= \frac{2\pi}{b-a}\left(j - \frac{N-1}{2}\right) \\ R_j &= \frac{2\pi}{b-a}\left(j - \frac{N-2}{2}\right) \end{aligned} \tag{.0.4}$$

Applying LPR, MPR and RPR to (2.5.1) yields

$$\begin{aligned}
\text{LPR}: \frac{1}{b-a}\sum_{j=0}^{N-1} e^{-2\pi i u_j x_k}\psi_X(2\pi u_j) &= \frac{1}{b-a}\sum_{j=0}^{N-1} e^{-2\pi i\left(\frac{a}{h}+k\right)\left(j-\frac{N}{2}\right)\frac{1}{N}}\psi_{\widetilde{Y}}\left(L_j\right) \qquad (.0.5)\\
\text{MPR}: \frac{1}{b-a}\sum_{j=0}^{N-1} e^{-2\pi i m_j x_k}\psi_X(2\pi m_j) &= \frac{1}{b-a}\sum_{j=0}^{N-1} e^{-2\pi i\left(\frac{a}{h}+k\right)\left(j-\frac{N-1}{2}\right)\frac{1}{N}}\psi_{\widetilde{Y}}\left(M_j\right)\\
\text{RPR}: \frac{1}{b-a}\sum_{j=0}^{N-1} e^{-2\pi i u_{j+1} x_k}\psi_X(2\pi u_{j+1}) &= \frac{1}{b-a}\sum_{j=0}^{N-1} e^{-2\pi i\left(\frac{a}{h}+k\right)\left(j-\frac{N-2}{2}\right)\frac{1}{N}}\psi_{\widetilde{Y}}\left(R_j\right)
\end{aligned}$$

Since $e^{i\pi} = -1$ then,

$$e^{-2\pi i\left(\frac{a}{h}+k\right)\left(j-\frac{N-\epsilon}{2}\right)\frac{1}{N}} = (-1)^{\left(\frac{a}{b-a}+\frac{k}{N}\right)(N-\epsilon)}\cdot(-1)^{\left(\frac{2a}{b-a}\right)j}\cdot e^{-2\pi i k\frac{j}{N}}$$

The k_{th} component of the DFT vectors associated to $y^L = \left(y_0^L,\ldots,y_{N-1}^L\right)$, $y^M = \left(y_0^M,\ldots,y_{N-1}^M\right)$ and $y^R = \left(y_0^R,\ldots,y_{N-1}^R\right)$ are

$$\begin{aligned}
\text{LPR}: f_k^L &= \frac{1}{b-a}\cdot(-1)^{\left(\frac{a}{b-a}+\frac{k}{N}\right)N}\cdot\sum_{j=0}^{N-1}(-1)^{\left(\frac{2a}{b-a}\right)j}\cdot\psi_{\widetilde{Y}}\left(L_j\right)\cdot e^{-2\pi i k\frac{j}{N}} \qquad (.0.6)\\
\text{MPR}: f_k^M &= \frac{1}{b-a}\cdot(-1)^{\left(\frac{a}{b-a}+\frac{k}{N}\right)(N-1)}\cdot\sum_{j=0}^{N-1}(-1)^{\left(\frac{2a}{b-a}\right)j}\cdot\psi_{\widetilde{Y}}\left(M_j\right)\cdot e^{-2\pi i k\frac{j}{N}}\\
\text{RPR}: f_k^R &= \frac{1}{b-a}\cdot(-1)^{\left(\frac{a}{b-a}+\frac{k}{N}\right)(N-2)}\cdot\sum_{j=0}^{N-1}(-1)^{\left(\frac{2a}{b-a}\right)j}\cdot\psi_{\widetilde{Y}}\left(R_j\right)\cdot e^{-2\pi i k\frac{j}{N}}\ .
\end{aligned}$$

That is,

$$\begin{aligned}
\text{LPR}: f_k^L &= C_k^L\times DFT(y^L)_k \qquad (.0.7)\\
\text{MPR}: f_k^M &= C_k^M\times DFT(y^M)_k\\
\text{RPR}: f_k^R &= C_k^M\times DFT(y^R)_k\ ,
\end{aligned}$$

where

$$C_k^L = \frac{1}{b-a}(-1)^{\left(\frac{a}{b-a}+\frac{k}{N}\right)N}\ ,\ C_k^M = \frac{1}{b-a}(-1)^{\left(\frac{a}{b-a}+\frac{k}{N}\right)(N-1)}\ ,\ C_k^R = \frac{1}{b-a}(-1)^{\left(\frac{a}{b-a}+\frac{k}{N}\right)(N-2)}$$

and

$$y_j^L = (-1)^{\left(\frac{2a}{b-a}\right)j}\psi_X\left(L_j\right)\ ,\ y_j^M = (-1)^{\left(\frac{2a}{b-a}\right)j}\psi_X\left(M_j\right)\ ,\ y_j^R = (-1)^{\left(\frac{2a}{b-a}\right)j}\psi_X\left(R_j\right)$$

REFERENCES

[1] G. E. Albert. Accurate sequential tests on the mean of an exponential distribution. *The Annals of Mathematical Statistics*, 27(2):pp. 460–470, 1956.

[2] P. Amore. Asymptotic and exact series representations for the incomplete gamma function, Jan 2005.

[3] M. F. Arlitt and C. L. Williamson. Web server workload characterization: the search for invariants. *SIGMETRICS Perform. Eval. Rev.*, 24(1):126–137, May 1996.

[4] S. Asmussen. *Applied Probability and Queues*. 2003.

[5] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, pages 71–82, 2002.

[6] P. Barford, S. Jha, and V. Yegneswara. Fusion and filtering in distributed intrusion detection systems. In *In Proceedings of the 42nd Annual Allerton Conference on Communication, Control and Computing*, 2004.

[7] T. Bass. Multisensor data fusion for next generation distributed intrusion detection systems. In *In Proceedings of the IRIS National Symposium on Sensor and Data Fusion*, pages 24–27, 1999.

[8] M. Becchi. From poisson processes to self-similarity: a survey of network traffic models.

[9] I. A. Belov. On the computation of the probability density function of α-stable distributions. In *Proceedings of the* 10^{th} *international conference Mathematical Modeling and Analysis 2005 and CMA M2*, pages 333–341, 2005.

[10] U. Blahak. Efficient approximation of the incomplete gamma function for use in cloud model applications. *Geoscientific Model Development*, 3(2):329–336, 2010.

[11] J. Błażejowski. Several expansion formulae of the complementary incomplete gamma function and possible applications. *Thermochimica Acta*, 48(1-2):125 – 135, 1981.

[12] M. Blum. On the Sums of Independently Distributed Pareto Variates. *Siam Journal on Applied Mathematics*, 19, 1970.

[13] K. Böcker and C. Klüppelberg. Operation var: a closed-form approximation, 2005.

[14] J. Boudec. *Performance Evaluation of Computer and Communication Systems*. 2010.

[15] R. Braga, Braga, E. Mota, Mota, and A. Passito, Passito. Lightweight ddos flooding attack detection using nox/openflow. In *Proceedings of the 2010 IEEE 35th Conference on Local Computer Networks*, pages 408–415, 2010.

[16] L. E. Brennan, I. S. Reed, and W. Sollfrey. A comparison of average-likelihood and maximum-likelihood ratio tests for detecting radar targets of unknown Doppler frequency. *IEEE Transactions on Information Theory*, 14:104–110, 1968.

[17] S. Broda and M. Paolella. Saddlepoint approximations: A review and some new applications. In *Handbook of Computational Statistics*, pages 953–983. 2012.

[18] G. Carl, G. Kesidis, R. R. Brooks, and S. Rai. Denial-of-service attack-detection techniques. *IEEE Internet Computing*, 10(1):82–89, Jan. 2006.

[19] J. M. Chambers, C. L. Mallows, and B. W. Stuck. A method for simulating stable random variables. *Journal of the American Statistical Association*, 71(354):pp. 340–344, 1976.

[20] V. Chatzigiannakis, G. Androulidakis, K. Pelechrinis, S. Papavassiliou, and V. Maglaris. Data fusion algorithms for network anomaly detection: classification and evaluation. In *Proceedings of the Third International Conference on Networking and Services*, pages 50–60, 2007.

[21] Y. Chen and K. Hwang. Collaborative detection and filtering of shrew ddos attacks using spectral analysis. *J. Parallel Distrib. Comput.*, 66(9):1137–1151, Sept. 2006.

[22] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, Nov. 2009.

[23] M. E. Crovella and M. S. Taqqu. Estimating the heavy tail index from scaling properties. *Method. Comput. Appl. Prob.*, 1(1):55–79, July 1999.

[24] C. Cunha, A. Bestavros, and M. Crovella. Characteristics of www client-based traces. Technical report, 1995.

[25] H. E. Daniels. Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, 25(4):631–650, 1954.

[26] A. C. Davison and D. Kuonen. Saddlepoint approximations for bootstrap distributions. *Computer Science and Statistics*, 28:248–253, 1998.

[27] A. C. Davison and D. Mastropietro. Saddlepoint approximation for mixture models. *Biometrika*, 96(2):479–486, 2009.

[28] C. Douligeris and A. Mitrokotsa. Ddos attacks and defense mechanisms: classification and state-of-the-art. *Comput. Netw.*, 44(5):643–666, Apr. 2004.

[29] P. Du and S. Abe. Ip packet size entropy-based scheme for detection of dos/ddos attacks. *IEICE - Trans. Inf. Syst.*, E91-D(5):1274–1281, May 2008.

[30] R. Durrett. *Probability: theory and examples*. 1996.

[31] G. S. Easton and E. Ronchetti. General saddlepoint approximations with applications to l statistics. *Journal of the American Statistical Association*, 81(394):pp. 420–430, 1986.

[32] P. Embrechts and M. Frei. Panjer recursion versus FFT for compound distributions. *Mathematical Methods of Operations Research*, 69(3):497–508, July 2009.

[33] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events: For Insurance and Finance*. 1997.

[34] G. Faÿ, B. González-Arévalo, T. Mikosch, and G. Samorodnitsky. Modeling teletraffic arrivals by a poisson cluster process. *Queueing Syst. Theory Appl.*, 54(2):121–140, Oct. 2006.

[35] L. Feinstein, D. Schnackenberg, R. Balupari, and D. Kindred. Statistical approaches to ddos attack detection and response. In *Proceedings of the DARPA Information Survivability Conference and Exposition*, pages 303–314, 2003.

[36] W. Feller. *An introduction to probability theory and its applications*. 1971.

[37] T. Ferguson. *Mathematical statistics: a decision theoretic approach*. 1967.

[38] H. Fofack and J. P. Nolan. Tail behavior, modes and other characteristics of stable distributions. *Extremes*, 2(1):39–58, 1999.

[39] H. Furrer, Z. Michna, and A. Weron. Stable levy motion approximation in collective risk theory. *Insurance: Mathematics and Economics*, 20(2):97 – 114, 1997.

[40] X. Gabaix. Power laws in economics and finance. Working Paper 14299, National Bureau of Economic Research, September 2008.

[41] R. Gatto. A saddlepoint approximation to the distribution of inhomogeneous discounted compound poisson processes. *Methodology and Computing in Applied Probability*, 12(3):533–551, 2010.

[42] W. Gautschi. A computational procedure for incomplete gamma functions. *ACM Trans. Math. Softw.*, 5(4):466–481, Dec. 1979.

[43] W. Gautschi, F. Harris, and N. Temme. Expansions of the exponential integral in incomplete gamma functions. *Applied Mathematics Letters*, 16(7):1095 – 1099, 2003.

[44] B. Ghosh. *Sequential tests of statistical hypotheses*. 1970.

[45] J. Gil-Pelaez. Note on the inversion theorem. *Biometrika*, 38(3-4):481–482, 1951.

[46] C. S. Gillespie and E. Renshaw. An improved saddlepoint approximation. *Mathematical Biosciences*, 208(2):359 – 374, 2007.

[47] P. Glasserman and K.-K. Kim. Saddlepoint approximations for affine jump-diffusion models. *Journal of Economic Dynamics and Control*, 33(1):15 – 36, 2009.

[48] B. González-Arévalo and J. Roy. Simulating a poisson cluster process for internet traffic packet arrivals. *Comput. Commun.*, 33(5):612–618, Mar. 2010.

[49] C. Goutis and G. Casella. Explaining the saddlepoint approximation. *The American Statistician*, 53(3):pp. 216–224, 1999.

[50] M. Grabchak and G. Samorodnitsky. Do financial returns have finite or infinite variance? a paradox and an explanation. *Quantitative Finance*, 10(8):883–893, 2010.

[51] I. S. Gradshteyn and I. M. Ryzhik, 2007.

[52] G. Hagstroem, K. Remarks on pareto distributions. *Scandinavian Actuarial Journal*, 1960 (1-2):59–71, 1960.

[53] H. Heffes and D. Lucantoni. A markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE J.Sel. A. Commun.*, 4(6): 856–868, Sept. 2006.

[54] B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

[55] N. Hohn, D. Veitch, and P. Abry. Cluster processes: a natural language for network traffic. *Signal Processing, IEEE Transactions on*, 51(8):2229–2244, 2003.

[56] X. Huang, C. W. Oosterlee, and H. van der Weide. Higher order saddlepoint approximations in the vasicek portfolio credit loss model. *Journal of Computational Finance*, 11(1):93–113, 2007.

[57] W. Hürlimann. Improved fft approximations of probability functions based on modified quadrature rules. *International Mathematical Forum*, 8(17):829–840, 2013.

[58] K. Hwang, M. Cai, Y. Chen, and M. Qin. Hybrid intrusion detection with weighted signature generation over anomalous internet episodes. *IEEE Trans. Dependable Secur. Comput.*, 4(1): 41–55, Jan. 2007.

[59] S. Jacobson, J. Kobza, and A. Easterling. A detection theoretic approach to modeling aviation security problems using the knapsack problem. *IIE Transactions*, 33:747–759, 2001.

[60] R. Jain and S. Routhier. Packet trains–measurements and a new model for computer network traffic. *IEEE J.Sel. A. Commun.*, 4(6):986–995, Sept. 2006.

[61] J. Jensen. *Saddlepoint Approximations*. 1995.

[62] N. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions*. Number v. 2. 1995.

[63] J. Jung, B. Krishnamurthy, and M. Rabinovich. Flash crowds and denial of service attacks: characterization and implications for cdns and web sites. In *Proceedings of the 11th international conference on World Wide Web*, pages 293–304, 2002.

[64] C. Kahl and R. Lord. Fourier inversion methods in finance, 2010.

[65] D. Kazakos and P. Papantoni-Kazakos. Sequential detection between poisson processes (corresp.). *Information Theory, IEEE Transactions on*, 26(1):116–120, 1980.

[66] J. Keilson. On the asymptotic behaviour of queues. *Journal of the Royal Statistical Society. Series B (Methodological)*, 25(2):pp. 464–476, 1963.

[67] J. Klafter and I. Sokolov. *First Steps in Random Walks: From Tools to Applications*. New York, 2011.

[68] C. Klüppelberg and T. Mikosch. Large deviations of Heavy-Tailed random sums with applications in insurance and finance. *Journal of Applied Probability*, 34(2), 1997.

[69] K. Lan, A. Hussain, and D. Dutta. Effect of malicious traffic on the network, 2003.

[70] P. Lévy. *Calcul des probabilités*. 1925.

[71] O. Lieberman. On the approximation of saddlepoint expansions in statistics. *Econometric Theory*, 10(5):pp. 900–916, 1994.

[72] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das. The 1999 darpa off-line intrusion detection evaluation. *Comput. Netw.*, 34(4):579–595, Oct. 2000.

[73] G. Lorden. On excess over the boundary. *The Annals of Mathematical Statistics*, 41(2):pp. 520–527, 1970.

[74] G. Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):pp. 1897–1908, 1971.

[75] R. Lugannani and S. Rice. Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability*, 12(2):475–490, 1980.

[76] B. Mandelbrot. Self-similar error clusters in communication systems and the concept of conditional stationarity. *Communication Technology, IEEE Transactions on*, 13(1):71–90, 1965.

[77] B. Mandelbrot and T. J. W. R. C. Y. Heights). *The Variation of Certain Speculative Prices*. 1962.

[78] M. M. Meerschaert and H.-P. Scheffler. Portfolio modeling with heavy tailed random vectors. In *Handbook of Heavy-Tailed Distributions in Finance*, pages 595–640, 2003.

[79] T. Mikosch. Regular variation, subexponentiality and their applications in probability theory, 1999.

[80] T. Mikosch. *Non-Life Insurance Mathematics: An Introduction With Stochastic Processes*. Number v. 13. 2004.

[81] J. Mirkovic. D-ward: Source-end defense against distributed denial-of-service attacks, August 2003.

[82] J. Mirkovic and P. Reiher. A taxonomy of ddos attack and ddos defense mechanisms. *SIGCOMM Comput. Commun. Rev.*, 34(2):39–53, Apr. 2004.

[83] J. Mirkovic, S. Fahmy, P. Reiher, and R. K. Thomas. How to test dos defenses. In *Proceedings of the 2009 Cybersecurity Applications & Technology Conference for Homeland Security*, pages 103–117, 2009.

[84] J. Mirkovic, A. Hussain, S. Fahmy, P. Reiher, and R. Thomas. Accurately measuring denial of service in simulation and testbed experiments. *Dependable and Secure Computing, IEEE Transactions on*, 6(2):81 –95, april-june 2009.

[85] S. Mittnik, T. Doganoglu, and D. Chenyao. Computing the probability density function of the stable paretian distribution. *Mathematical and Computer Modelling*, 29(1012):235 – 240, 1999.

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

[87] M. Moscadelli. The modelling of operational risk: Experience with the analysis of the data collected by the basel committee, 2004.

[88] P. Moulin and V. V. Veeravalli. Detection and estimation, lectures notes, 2013.

[89] S. Nadarajah and S. Kotz. On the laplace transform of the pareto distribution. *Queueing Systems*, 54(4):243–244, 2006.

[90] S. V. Nagaev. Large deviations of sums of independent random variables. *The Annals of Probability*, 7(5):pp. 745–789, 1979.

[91] J. Nazario. Ddos attack evolution. *Network Security*, 2008(7):7 – 10, 2008.

[92] J. Nolan. *Stable Distributions: Models for Heavy-Tailed Data*. 2007.

[93] J. P. Nolan. Numerical calculation of stable densities and distribution functions. *Communications in Statistics Stochastic Models*, 13(4):759–774, 1997.

[94] J. P. Nolan. An algorithm for evaluating stable densities in zolotarev's (m) parameterization. *Math. Comput. Model.*, 29(10-12):229–233, May 1999.

[95] R. Novak. Statistical signal processing, lectures notes, 2010.

[96] P. Ohman-Strickland and G. Casella. Approximate and estimated saddlepoint approximations. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 30(1):pp. 97–108, 2002.

[97] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors. *NIST Handbook of Mathematical Functions*. Cambridge University Press, New York, NY, 2010.

[98] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1-2):100–115, 1954.

[99] V. Paxson. Empirically derived analytic models of wide-area tcp connections. *IEEE/ACM Trans. Netw.*, 2(4):316–336, Aug. 1994.

[100] V. Paxson and S. Floyd. Wide area traffic: the failure of poisson modeling. *IEEE/ACM Trans. Netw.*, 3(3):226–244, June 1995.

[101] T. Peng, C. Leckie, and K. Ramamohanarao. Detecting distributed denial of service attacks using source ip address monitoring. In *In Proceedings of the Third International IFIP-TC6 Networking Conference (Networking 2004*, pages 771–782, 2002.

[102] T. Peng, C. Leckie, and K. Ramamohanarao. Protection from distributed denial of service attack using history-based ip filtering. pages 482–486, 2003.

[103] H. V. Poor. An introduction to signal detection and estimation, 1988.

[104] S. Prowell, R. Kraus, and M. Borkin. *Seven Deadliest Network Attacks*. 2010.

[105] C. M. Ramsay. The distribution of sums of certain i.i.d. pareto variates. *Communications in Statistics - Theory and Methods*, 35(3):395–405, 2006.

[106] C. M. Ramsay. The distribution of compound sums of pareto distributed losses. *Scandinavian Actuarial Journal*, 2009(1):27–37, 2009.

[107] S. Resnick and H. Rootzn. Self-similar communication models and very heavy tails. *The Annals of Applied Probability*, 10(3):pp. 753–778, Aug. 2000.

[108] B. Roehner and P. Winiwarter. Aggregation of independent paretian random variables. *Advances in Applied Probability*, 17(2):pp. 465–469, 1985.

[109] S. M. Ross. Chapter 5 - the exponential distribution and the poisson process, 2008.

[110] G. Samorodnitsky and M. Taqqu. *Stable non-Gaussian random processes: stochastic models with infinitive variance*. 1994.

[111] P. Schaller and G. Temnov. Efficient and precise computation of convolutions: Applying fft to heavy tailed distributions. *Computaional Methods in Applied Mathematics*, 8(2):187–200, 2008.

[112] L. Shinder and M. Cross. Chapter 10 - understanding network intrusions and attacks, 2008.

[113] T. Shon and J. Moon. A hybrid machine learning approach to network anomaly detection. *Inf. Sci.*, 177(18):3799–3821, 2007.

[114] D. L. Snyder and M. I. Miller. *Chapter 4 - Compound Poisson-Processes*. Second edition edition, 1991.

[115] A. Tartakovsky. Quickest detection of targets in multiple-resolution-element systems: Sequential detection versus non-sequential detection. *Proceedings of the Workshop on Estimation, Tracking, and Fusion*, pages 326–350, 2001.

[116] A. Tartakovsky, B. Rozovskii, R. Blazek, and H. Kim. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *Signal Processing, IEEE Transactions on*, 54(9):3372–3382, 2006.

[117] N. M. Temme. Uniform asymptotic expansions of the incomplete gamma functions and the incomplete beta function. *Mathematics of Computation*, 29(132):pp. 1109–1114, 1975.

[118] T. Thapngam, S. Yu, W. Zhou, and G. Beliakov. Discriminating ddos attack traffic from flash crowd through packet arrival patterns. In *Computer Communications Workshops, 2011 IEEE Conference on*, pages 952 –957, 2011.

[119] V. Uchaikin and V. Zolotarev. *Chance and Stability: Stable Distributions and Their Applications*. 1999.

[120] N. Ujevic and A. J. Roberts. A corrected quadrature formula and applications. *ANZIAM J.*, 45(E):E41–E56, Feb. 2004.

[121] A. Wald. *Sequential Analysis*. 1997.

[122] L. A. Waller, B. W. Turnbull, and J. M. Hardin. Obtaining distribution functions by numerical inversion of characteristic functions with applications. *The American Statistician*, 49(4):pp. 346–350, 1995.

[123] H. Wang, D. Zhang, and K. G. Shin. Detecting syn flooding attacks. In *In Proceedings of the IEEE Infocom*, pages 1530–1539, 2002.

[124] J. Wang, R.-W. Phan, J. Whitley, and D. Parish. Ddos attacks traffic and flash crowds traffic simulation with a hardware test center platform. In *Internet Security (WorldCIS), 2011 World Congress on*, pages 15 –20, feb. 2011.

[125] S. Wang. General saddlepoint approximations in the bootstrap. *Statistics & Probability Letters*, 13(1):61 – 66, 1992.

[126] R. A. Wasniowski. Multi-sensor agent-based intrusion detection system. In *Proceedings of the 2nd annual conference on Information security curriculum development*, pages 100–103, 2005.

[127] R. Weron. Levy-stable distributions revisited: tail index ¿ 2 does not exclude the levy-stable regime. 2001.

[128] W. Willinger, V. Paxson, and M. S. Taqqu. A practical guide to heavy tails. chapter Self-similarity and heavy tails: structural modeling of network traffic, pages 27–53. 1998.

[129] M. Wilson. A historical view of network traffic models.

[130] S. Winitzki. Computing the incomplete gamma function to arbitrary precision. In *Proceedings of the 2003 international conference on Computational science and its applications: PartI*, pages 790–798, 2003.

[131] A. T. A. Wood, J. G. Booth, and R. W. Butler. Saddlepoint approximations to the cdf of some statistics with nonnormal limit distributions. *Journal of the American Statistical Association*, 88(422):pp. 680–686, 1993.

[132] J. Yang, T. Hurd, and X. Zhan. Saddlepoint approximation method for pricing CDOs. *Journal of Computational Finance*, 10(1):1 –20, 2006.

[133] J. Yuan and K. Mills. Monitoring the macroscopic effect of ddos flooding attacks. *IEEE Trans. Dependable Secur. Comput.*, 2(4):324–335, Oct. 2005.

[134] I. Zaliapin, Y. Kagan, and F. Schoenberg. Approximating the distribution of pareto sums. *pure and applied geophysics*, 162(6-7):1187–1228, 2005.

[135] S. Zheng, T. Jiang, and J. Baras. Performance comparison of two sequential change detection algorithms on detection of in-band wormholes. In *Information Sciences and Systems, 2009. CISS 2009. 43rd Annual Conference on*, pages 270–275, 2009.