Chapter 2

Analyzing Network Traffic for Malicious Hacker Activity

Problem Presented By: Surrey Kim (Random Knowledge Inc.)

Mentors: Hongwei Long and Weiguang Shi (University of Alberta); Randall Pyke (University College of the Fraser Valley); Lang Wu (University of British Columbia)

Student Participants: Stanislava Peker (Concordia University); Benjamin Chan, Radu Haiduc and, Andrei Maxim (Cornell University); Vilen Abramov (Kent State University); Bo Zeng (Purdue University); Pengpeng Wang (Simon Fraser University); Robert Liao, Yury Petrachenko, Yulia Romaniuk, Mengzhe Wang, Zhian Wang and Mohammad Ali Yassaei (University of Alberta); Gabriel Mititica, Shijun Song and Xuekui Zhang (University of British Columbia); Song Li and Tzvetalin Vassilev (University of Saskatchewan); Nancy Azer, Mahin Salmani and Jiaping Zhu (University of Victoria)

Report prepared by: Randall Pyke (`pyke@math.toronto.edu`)

2.1 Problem Statement

Network security is still at its infancy. Existing intrusion detection and prevention solutions lack accuracy, broad attack coverage, speed, performance, and scalability. They do not provide reliable protection to today's vital networks.

Random Knowledge Inc.'s approach to intrusion detection is to apply Mathematically Optimal Detection that outperforms other methods, including pattern matching, neural networks and statistical techniques. This detection system, Portscan Detection System (PDS), detects and localizes traffic patterns consistent with possibly stealthy forms of attacks from within hoards of legitimate traffic. With the network's packet traffic stream being its input, PDS relies on high fidelity models for normal traffic from which it can critically judge the legitimacy of any substream of packet traffic.

In this modelling workshop, we try to characterize normal traffic which involves:

- Defining all the different types of connection sessions.

- Verification of a Poisson measure model for the incoming connection sessions, i.e. if the connection session types are labelled $1, \ldots, n$, determining if $N(A \times (0, t])$ is Poisson distributed for any subset A of $\{1, \ldots, n\}$, where N is the Poisson measure.
- Determining the rates for $N(A \times (0, t])$ or equivalently its mean measure if the session generation indeed conforms reasonably to the Poisson measure model, otherwise suggesting other suitable models.
- Verification for self-similar processes and heavy tailed distributions within connection sessions (for example the transmission time), and the estimation of its parameters.

Hitherto, there has been much study of traffic characterization that focuses on the implications for improved network performance. Random Knowledge's approach is the study of traffic characterization for the implications of detecting malicious hacker activity.

2.2 Glossary

- Normal traffic: Traffic generated by legitimate users.
- Mark: The IP address and port number pair. Each mark is a server with a specific service.
- Session: The analog of a phone call between a client (source IP) and a mark (the HTTP server). It starts when the first packet from a client arrives at a mark and ends if the silent (no packet) time is longer than a preset time-out value. Note: One session could include transmission of multiple packets.
- Interarrival time: Time between the arrivals of two consecutive sessions.

2.3 Introduction

2.3.1 What is the Problem?

Since the Internet came into life in the 1970s, it has been growing more than 100% every year. On the other hand, the solutions to detecting network intrusion are far outpaced. The economic impact of malicious attacks in lost revenue to a single e-commerce company can vary from 66 thousand up to 53 million US dollars [1]. At the same time, there is no effective mathematical model widely available to distinguish anomaly network behaviours such as port scanning, system exploring, virus and worm propagation from normal traffic.

PDS proposed by Random Knowledge Inc., detects and localizes traffic patterns consistent with attacks hidden within large amounts of legitimate traffic. With the network's packet traffic stream being its input, PDS relies on high fidelity models for normal traffic from which it can critically judge the legitimacy of any substream of packet traffic. Because of the reliability on an accurate baseline model for normal network traffic, in this workshop, we concentrate on modelling normal network traffic with a Poisson process.

2.3.2 Data Description

The dataset is a record of network traffic of the University of Auckland, New Zealand in March 2001. The arrival times are recorded to the nearest nanosecond and are collected with IP addresses and port numbers. Each combination of IP and port denotes a network service provided by a specific server of the University. The pair of IP address 5122 and port 80, for example, denotes a server with IP 5122 providing HTTP web service. In this workshop, we analyzed the traffics of IP addresses 5122, 5226 and 5264 and services ftp (ports 20, 21), telnet (port 23) and HTTP (port 80).

The record also contains both inbound and outbound network traffic of the University. For our purpose of modelling normal traffic from outside, only the inbound traffic is relevant and considered. Note that a relatively small portion, about 7% of half a day, of the traffic data of IP 5264 is missing. Data analysis of this server was conducted on the reduced dataset.

2.3.3 Goals

The primary motivation of this work is to justify the modelling of normal session traffic with a nonhomogeneous Poisson process. We justify this by noting that i) service usages of different clients are expected to be independent events and ii) the number of packets having arrived by time t is a counting process $\{N(t), t \geq 0\}$ whereby the number of events in an interval of length t is $E(N(t)) = \lambda(t)t$, $\lambda > 0$. The actual recorded session traffic will be used to model the normal network traffic which may include some malicious network behaviour.

One method used to identify vulnerable ports of a network service system is to send a sequence of probing packets to all available ports over a relatively short period of time. This reconnaissance behaviour identifies which ports of a network are open and which services have been made available. In the traditional network traffic model using packets, port scanning takes up a tiny portion of the traffic and is difficult to detect. By grouping the packets of each session together a probing session will violate the assumption of independence of the arrival times across the ports of the network. This violation of independence allows one to identify of this type of malicious behaviour much more efficiently.

To justify the usage of the Poisson process model we note that the sessions representing different service requests are independent events. However it is known that the arrival rate can be considered constant only over a relatively short (five minutes) interval. Extensions beyond this short interval do not model modern servers very well. A rather comprehensive critique of the use of a single homogeneous Poisson process in modelling network arrivals can be found in [6]. Rather than a single Poisson process, we assume that the overall session traffic, namely session arrivals, can be modelled as a sequence of homogeneous Poisson processes. Alternatively this can be thought of as a nonhomogeneous Poisson process where the arrival rate $\lambda(t)$ is piecewise constant.

We can break up the model characterization into the following tasks:

1. Analyze the arrival patterns of normal traffic of different service types (marks) of connection sessions. (A mark is defined as the IP address and port pair. It can be also thought of as a specific service running on a server.)

2. Test if the arrivals are nonhomogeneous Poisson. If they are indeed nonhomogeneous Poisson, then estimate the relevant parameters of the models such as mean and standard deviation, also maximize time interval within which the arrival rate is constant.

3. Otherwise, suggest other suitable model and estimating its parameters.

2.4 Methodology

In this section we will describe the statistical framework for testing the nonhomogeneous Poisson model. Because of the close relationship between the exponential and Poisson distributions (detailed below) we test if the session arrivals are Poisson distributed by checking that the interarrival times are i) exponentially distributed and ii) independent. In the following we begin by briefly reviewing the definitions of the Poisson and exponential distributions and their relationship. To test for a non-homogeneous Poisson distribution both a Goodness-of-Fit test and a test for independence are then described. More sophisticated nonparametric techniques for estimating the model parameters of a nonhomogeneous Poisson model are described in detail in [4].

2.4.1 Poisson and Exponential Distributions

A Poisson distribution is typically used to model the number of events happening per time interval, such as the number of customers arriving at a store per hour, or the number of visits per minute to an internet site. A random variable (r.v.) N that takes values $0, 1, 2, ...$ has λ-Poisson distribution if

$$P(N = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots . \tag{2.1}$$

The mean and variance of the Poisson distribution both equal λ.

A continuous random variable X is said to have an exponential distribution with parameter λ, ($\lambda > 0$), if its probability distribution is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

Equivalently the cumulative distribution function $F(x) = 1 - e^{-\lambda x}$, $x \geq 0$; $F(x) = 0$, $x < 0$. The mean and variance of the exponential distribution are $1/\lambda$ and $1/\lambda^2$ respectively.

Recall that a counting process $\{N(t), t \geq 0\}$ is said to be a Poisson process having rate λ, $\lambda > 0$, provided

1. $N(0) = 0$.

2. The process has independent increments.

3. The number of events in an interval of length t is Poisson distributed with mean λt.

A nonhomogeneous Poisson process allows λ to be a function of time and as a result, allows for the possibility that events may be more likely to occur at certain times during the day.

Consider a Poisson process and let T_1 denote the time of the first event. In addition let T_n denote the time elapsed between the nth and $(n-1)$st events. To determine the distribution of the sequence of interarrival times $\{T_n\}$ note that the event $(T_1 > t)$ occurs only if no events of the Poisson process occur over the interval $[0, t]$. As a result,

$$P(T_1 > t) = P(N(t) = 0) = e^{-\lambda t}$$

so that T_1 is exponentially distributed with mean $1/\lambda$. The distribution of T_2 is obtained by computing

$$\begin{aligned} P(T_2 > t | T_1 = s) &= P(0 \text{ events in } (s, s+t] | T_2 = s) \\ &= P(0 \text{ events in } (s, s+t]) = e^{-\lambda t}. \end{aligned}$$

Repeating the same argument we see that the $\{T_n\}$ are independent identically distributed exponential random variables with mean $1/\lambda$.

2.4.2 Goodness-of-fit Test

To test that interarrival times between sessions are λ-exponentially distributed, we use Anderson-Darling (A-D) test, which checks if a given sample is drawn from a population with a specified distribution [2, 3]. There are other methods to test for goodness-of-fit, such as Kolmogorov-Smirnov (K-S) or chi-square tests. However, A-D test is more appropriate in our case, as it doesn't require the true population parameters, but uses those estimated from available data [2, 3]. The test statistic is $A^2 = -N - S$, where N is the sample size, and

$$S = \sum_{i=1}^{N} \frac{2i-1}{N} \left[\ln(F(Y_i)) + \ln(F(Y_{N+1-i})) \right]. \tag{2.2}$$

In the above equation, Y_i are sample values (sorted in ascending order), and F is the cumulative distribution function of the specified distribution (in our case, exponential with $\lambda = 1/\bar{Y}, \quad \bar{Y} = \sum_{i=1}^{N} Y_i/N$). As the A-D test uses an estimated mean, A^2 has to be multiplied by a correction factor, so that the actual used statistic is $A^2_* = A^2 \cdot (1 + .6/N)$. The null hypothesis that the sample is drawn from a given distribution is rejected if $A^2_* \geq 1.341$ at the 95% significance level [7].

The data files contained session interarrival times for a six-hour trace of internet traffic. The sessions were determined based on the time-out values. Each six-hour interval was subdivided into 5-minute subintervals to test whether λ was constant during this subinterval (A-D test).

2.4.3 Independence Test

Next, we test whether the interarrival times were independent within each time interval, as well as between the first lag of the 5-minute subintervals. For this, we used the autocorrelation function: given measurements $Y_1, Y_2, ..., Y_n$ at times $t_1, t_2, ..., t_n$,

$$r_k = \frac{\displaystyle\sum_{i=1}^{N-k} (Y_i - \overline{Y})(Y_{i+k} - \overline{Y})}{\displaystyle\sum_{i=1}^{N} (Y_i - \overline{Y})^2}, \qquad \overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i. \tag{2.3}$$

Autocorrelation is significant in two ways: it can be too strong in magnitude or too frequently positive/negative. Note that for a time series of N samples from an uncorrelated white noise process, $P(r_k > 1.96/\sqrt{N}) = .05$.

We use binomial probability distribution with parameter $p = .95$ (the probability of success) to check for independence of interarrival times. The density of the binomial r.v. X, representing the number of successes in N trials, is:

$$P(X = k) = \binom{N}{k} p^k (1-p)^{N-k}, \qquad k = 0, 1, \ldots, N. \tag{2.4}$$

Within each 5-minute subinterval, we calculated autocorrelation values and checked (using binomial distribution) whether sufficiently many of them were below $1.96/\sqrt{N}$ to claim that the interarrival times were uncorrelated. So if $P(X = k) \leq .05$, we reject the hypothesis that the interarrival times are independent. Similarly, we performed the binomial tests to check for the significance of the autocorrelation sign (using $p = .5$ and a rejection region of .025) within each subinterval and between the lags for the subintervals of the entire six hour trace.

Alternatively we could have tested the null hypothesis that the arrivals are independent in a given time series using the Box-Pierce or Ljung-Box tests [5]. The advantage of the A-D test lies in its robustness and the fact that it does not assume the arrival times are normally distributed.

2.5 Results

Our test results are based on the three servers that have the highest traffic volume, namely, web servers with IP addresses 5122, 5226 and 5264. This is a starting point about which we can frame a set of algorithms for model diagnostics. Later, these algorithms can be systematically extended to the other servers.

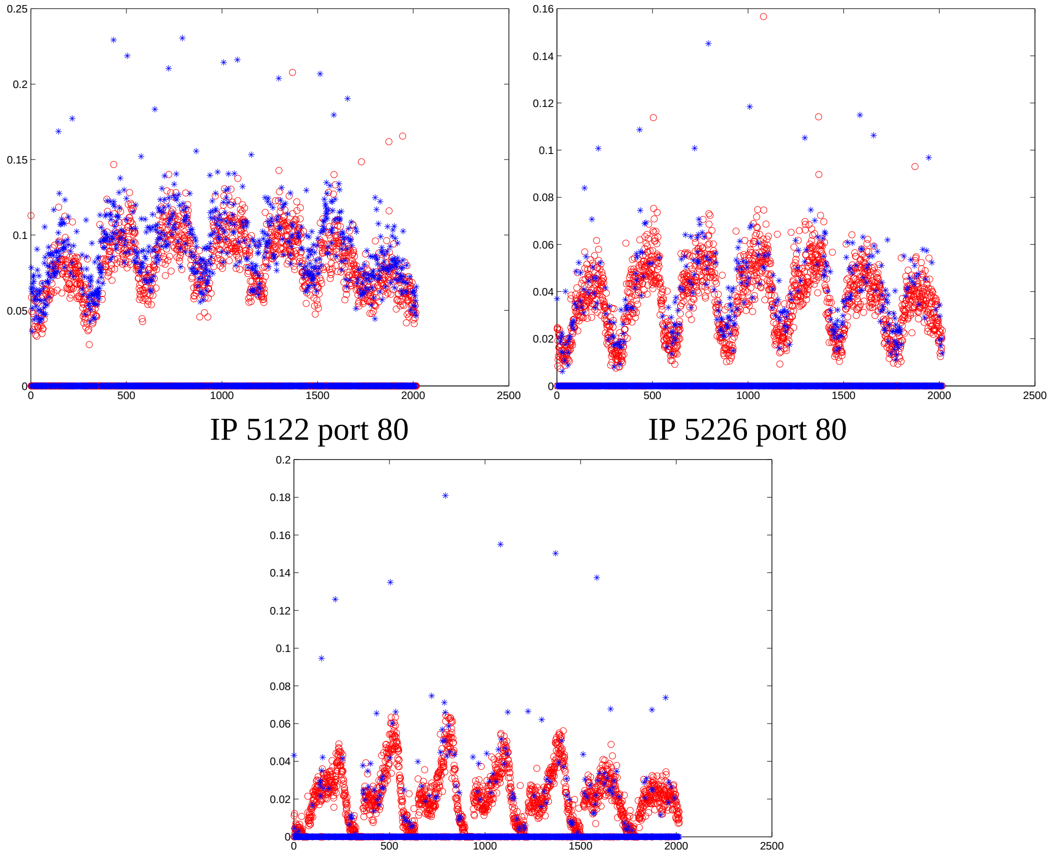
2.5.1 Arrival Pattern Suggesting Strong Periodicity in Interarrival Time

Figure 2.1 shows strong periodicity in session interarrival times. This periodicity suggests that a piecewise constant interarrival rate is a reasonable first assumption. (It would be of interest to use the periodicity to predict future interarrival times.)

2.5.2 Statistical Tests Reasonably Confirming Nonhomogeneous Poisson

Based on periodicity, our test procedures assume that interarrival times of a given server will depend on time of day. Recall that our null hypothesis is that the arrivals follow a nonhomogeneous Poisson process. It is a convincing indicator of a nonhomogeneous Poisson process if the interarrival times follow an exponential distribution with a moderately time varying rate. For convenience we used intervals of equal length and started with 5 minutes as basic unit.

Table 2.1 summarizes the results of K-S and A-D tests that the interarrival times are distributed exponentially on the time scale of five minutes. As a result there is no strong evidence to reject the claim that the arrival of session requests follows a nonhomogeneous Poisson process with a piecewise constant arrival rate.



IP 5122 port 80

IP 5226 port 80

IP 5264 port 80

Figure 2.1: The data plot of the session interarrival time of different marks.

To further validate independence amongst interarrival times for all of the tested servers, autocorrelations, shown in Figure 2.2, are not significant either in magnitude or in too frequently positive/negative. These figures can be viewed as an analog of a confidence interval depicting how wide the autocorrelations spread out from the zero line.

In addition to visual validation, the tests on individual points by Box-Pierce statistics does not show strong evidence against the independence hypothesis. Table 2.1 provides a percentage of points that pass the independence test.

2.6 Conclusion and Future Work

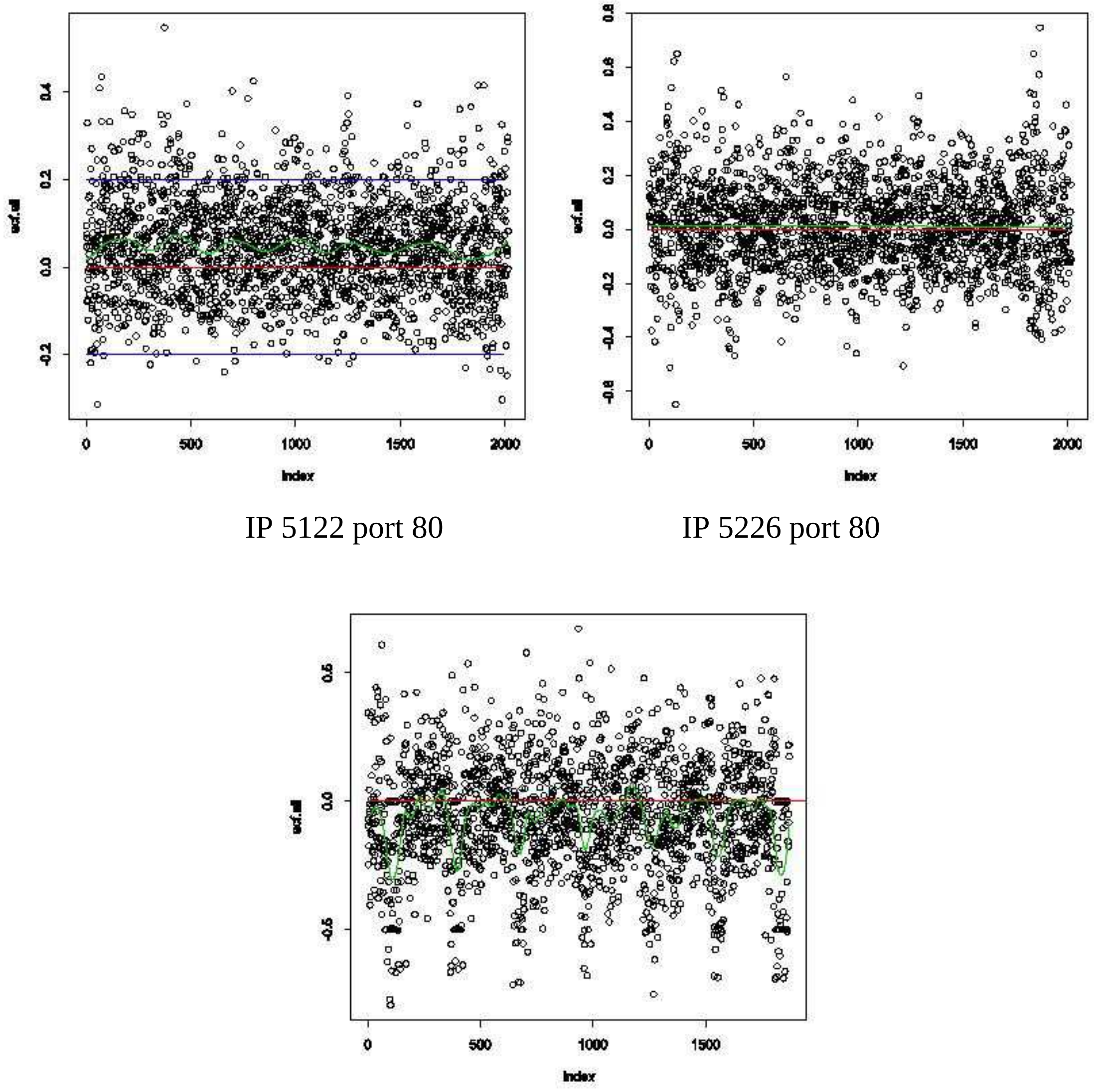
In this report, we have tested the nonhomogeneous Poisson process by validating if the interarrival times were independently exponentially distributed with time varying rates. In most respects the results of three tested servers do not differ noticeably from each other. Both visual and numerical tests have reasonably confirmed an independent exponential distribution.

At this workshop we investigated the modelling of normal session traffic with a nonhomogeneous Poisson process in the specific situation when the arrival rate is piecewise constant. However, there are a few other questions to be answered as to further this anomaly detection study such as: how to

Table 2.1: Percentage of sessions requests that pass.

	IP 5122	**IP 5226**	**IP 5264**
K-S $(\alpha = 1\%)$	84%	93%	97%
A-D $(\alpha = 5\%)$	83%	63%	82%
Independence	96.5%	99.2%	98.2%

extend the nonhomogeneous model on the other servers; to develop an algorithm by periodicity so as to forecast future arrival patterns; and to apply a method to group marks by dependence. As a result of joint efforts of PIMS and Random Knowledge Inc., some of workshop team members are participating in an ongoing project to develop anomaly detection algorithms.



IP 5122 port 80

IP 5226 port 80

IP 5264 port 80

Figure 2.2: The autocorrelation of the session interarrival time of different marks.

Bibliography

[1] *The Return on Investment for Network Security,* (2002). Cisco Systems Inc. `http://www.cisco.com/warp/public/cc/so/neso/sqso/roi4_wp.pdf`

[2] Anderson, T. W. and Darling, D. A. (1952). *Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes*. The Annals of Mathematical Statistics, 23(2), pp. 193-212.

[3] Anderson, T. W. and Darling, D. A. (1954). *A test of goodness of fit*. Journal of the American Statistical Association, 49(268), pp. 765-769.

[4] Leemis, L. M. (1991). *Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process*. Management Science, 37(7), pp. 886-900.

[5] Ljung, G. M. and Box, G. E. P. (1978). *On a measure of lack of fit in time series models*. Biometrika, 65(2), pp. 297-303.

[6] Paxson, V. and Floyd, S. (1995). *Wide-Area Traffic: The Failure of Poisson Modeling*. IEEE/ACM Transactions on Networking, 3(3), pp. 226-244.

[7] Stephens, M. A. (1974). *EDF Statistics for Goodness of Fit and Some Comparisons*, Journal of the American Statistical Association, 69(347), pp. 730-737.