DETECTING HACKERS (ANALYZING NETWORK TRAFFIC) BY POISSON MODEL MEASURE

Participants:

Surrey Kim, Random Knowledge Inc., and Randall Pyke, University College of the Fraser Valley, (Mentors)

Stanislava Peker, Concordia University

Benjamin Chan, Andrei Maxim, Radu Haiduc, Cornell University

Vilen Abramov, Kent State University

Bo Zeng, Purdue University

Hongwei Long, Weiguang Shi, Random Knowledge Inc., and Univ. of Alberta

Pengpeng Wang, Simon Fraser University

Yury Petrachenko, Robert Liao, Mengzhe Wang, Zhian Wang, Yulia Romaniuk, Univ. of Alberta

Shijun Song, Xuekui Zhang, Gabriel Mititica, Univ. of British Columbia

Song Li, Tzvetalin Vassilev, Univ. of Saskatchewan

Jiaping Zhu, Mahin Salmani, Nancy Azer, Univ. of Victoria

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:

a) defining all the different types of connection sessions,

b) 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,

c) determining the rates for $N(A \times (0, t])$ or equivalently its mean measure if the session generation indeed conform reasonably to the Poisson measure model, otherwise suggesting other suitable models, and

d) 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.

1 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 then a preset time-out value. Note: One session could include transmission of multiple packets.
- Inter-arrival time: Time between the arrivals of two consecutive sessions.

2 Introduction

2.1 What is the problem?

Since the Internet came into life in 1970s, it has been growing more than 100% every year. On the other hand, the solutions to detecting network intrusion is 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 behaviors such as port scanning, system exploring, virus and worm propagation from normal traffic.

Portscan Detection System (PDS), proposed by Random Knowledge Inc., 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. Because of the reliability on a proper model for normal network traffic, in this workshop, we concentrate on modelling the normal traffic, by the Poisson model in particular.

2.2 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 in accuracy of nanoseconds. It is well known that a university usually provides variety types of network services such as HTTP, FTP and Telnet. The arrival times 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.

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. By filtering out the outbound traffic from the original dataset, we still have a large dataset of size in Gigabytes.

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 is then conducted on the reduced dataset.

2.3 Goals

The purpose of this work is to design an efficient method to balance between missing malicious detections (false negative) and false alarms (false positive).

Our normal traffic model is built on the pattern of incoming connection sessions, which is defined by the time interval between the arrival of a packet up to the last packet arrival before the silent time is greater than a preset value. Then we are going to model the normal network traffic using session traffic. Sessions can model well the independent service usages of different clients. Intuitively packets belonging to a single service should arrive in consecutively. In other words, if the silent time is larger than a certain value, two packets separate by this silent time are less likely to belong to the same service. It follows that the independent events arrival model, namely Poisson process, can adequately fit the normal session traffic pattern. This is one major motivation of this research work.

The other motivation is based on the characteristic of the normal malicious hacking behaviors, namely the reconnaissance behavior. In reconnaissance, a hacker normally first tries to do a port scanning for vulnerable computer ports of the network service system. This is done by the means of sending probe packets to all the available ports, in a short time, to get the information whether this port is open, and what service is running. In traditional network traffic model using packets, this port scanning is just a tiny portion of the traffic, and difficult to detect. However, by sessions, dependent packets are grouped together into a single entity, and at the same time, the probe packets are themselves sessions by definitions. These probe sessions are related and not characterized by Poisson model. Thus, it is much easier to identify this type of malicious behaviors.

To justify the usage of Poisson process model note the following features: First, as mentioned before, the sessions representing different service requests are independent events; Then, it is widely accepted that the network traffic has a constant rate within a short time frame, say 5 minutes. Due to the time-dependence of arrivals, we propose this is an inhomogeneous Poisson process as illustrated in Figure 1. In this model proposed, the arrival rate λ is a time-dependent function. Thus, the overall session traffic, namely session arrivals, can be modelled piecewisely as homogeneous Poisson process.

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

A) 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.)

B) Test if the arrivals are inhomogeneous Poisson. If they are indeed inhomogeneous 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.

C) Otherwise, suggest other suitable model and estimating its parameters.

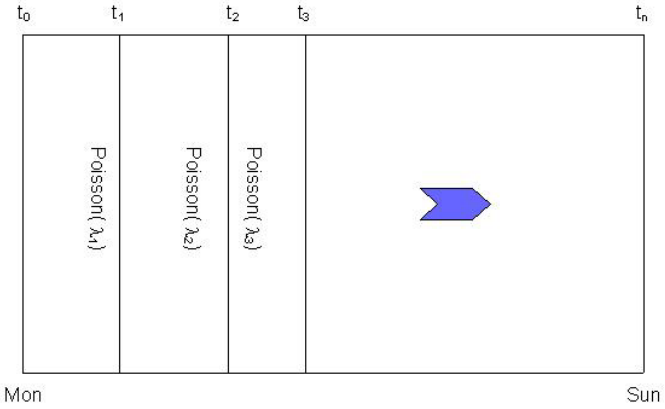


Figure 1: The inhomogeneous Poisson process model for network session traffic within a week.

3 Methodology

In this section we will describe the statistical framework for testing inhomogeneous Poisson model. Because of the close relation (See below for details.) between Poisson and exponential distributions, we are going to test if the data, session arrivals, is actually Poisson distributed by exponential distribution test and independence test of the inter-arrival time. In the following, we are going to first briefly recapitulate the definitions of Poisson and exponential distributions and their close relations. Then the Goodness-of-Fit test of exponential distributions and independence test are described as to how to test inhomogeneous Poisson distribution.

3.1 Poisson and exponential distributions

Poisson distribution is used to model the number of events happening per time interval, such as the number of customers arriving to a store per hour, or a 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

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

The mean and variance of Poisson distribution both equal λ.

Exponential distribution, as its name implies, describes a process whose probability is exponentially distributed. Using $f(x)$ and $F(x)$ to denote the probability density and cumulative distribution functions respectively, we have for an λ-exponential distribution,

$$\begin{aligned} f(x) &= \lambda \cdot e^{-\lambda x} \quad \text{for} \quad x \geq 0 \quad \text{or} \quad 0 \quad \text{otherwise}, \\ F(x) &= 1 - e^{-\lambda x}, \qquad \lambda > 0. \end{aligned} \tag{2}$$

The mean and variance of an exponentially distributed r.v. X are $1/\lambda$ and $1/\lambda^2$ respectively.

The Poisson and exponential distributions are closely related in the following fact. To say event arrival times conform to the Poisson distribution is equivalently to say that the inter-arrival times of these events are independent r.v. and exponentially distributed. This is why in the following, in order to test the network session traffic is Poisson distributed, we test both that the inter-arrival times of the sessions are exponentially distributed and their independence.

3.2 Goodness-of-fit test

To test that inter-arrival 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 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

$$S = \sum_{i=1}^{N} \frac{2i-1}{N} (ln[F(Y_i)] + ln[1 - F(Y_{N+1-i})], \tag{3}$$

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 $\lambda = 1/\bar{Y}$, $\bar{Y} = \sum_{i=1}^{N} Y_i/N$). As A-D test uses estimated mean, A^2 has to be multiplied by a constant 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 $A_*^2 \geq 1.341$ while using the 95% significance level.

The data files contained session inter-arrival 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).

3.3 Independence test

Next, we test whether the inter-arrival 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 $t_1, t_2, ..., t_n$,

$$r_k = \frac{\sum_{i=1}^{n-k} (Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}, \qquad \bar{Y} = \sum_{i=1}^{n} Y_i/n. \tag{4}$$

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, the probability that the autocorrelation will exceed $1.96/\sqrt{n}$ is .05.

We use binomial probability distribution with parameter .95 (the probability of success we seek) to check for independence of inter-arrival times. The density of the binomial r.v. X, representing the number of successes in N trials, is:

$$P(X = K) = N!/((N - K)! \cdot K!) \cdot (.95)^K \cdot (.05)^{N-K}. \quad (5)$$

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 inter-arrival times were uncorrelated. So if $P(X = K) \leq .05$, we reject the hypothesis that the inter-arrival times are independent. Similarly, we performed the binomial tests to check for the significance of autocorrelation sign (using parameter 1/2, and rejection region .025) within each subinterval and between the lags for the subintervals of the entire six-hour trace.

Alternatively, we can test the null hypothesis that the arrivals are independent in a given time series using the Box-Pierce or Ljung-Box tests [4]. The advantage of this test lies in its robustness, with not necessary the assumption of the normality of the arrivals distribution.

4 Results

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

4.1 Arrival pattern suggesting strong periodicity in inter-arrival time

Fig. 2 show strong periodicity in session inter-arrival time. This periodicity suggests that inter-arrival time can be represented by a function of time. (It would be of interest that the periodicity can be used to predict future inter-arrival time.)

4.2 Statistical tests reasonably confirming inhomogeneous Poisson

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

Table 1 summarize all the results of K-S and A-D statistics we obtained from the dataset. Both tests show that exponential distribution has a good fit to the inter-arrival

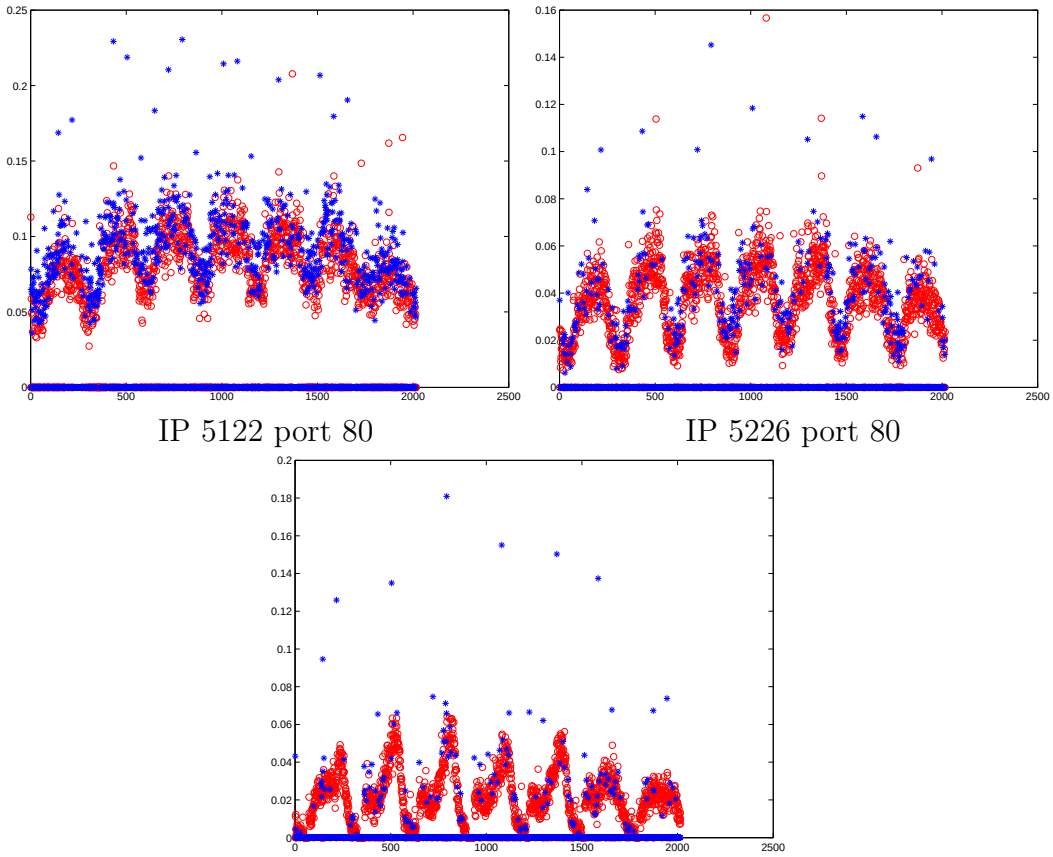


Figure 2: The data plot of the session inter-arrival time of different marks.

time. Thus, overall there is no strong evidence to reject our null hypothesis that the arrival of session requests is an inhomogeneous Poisson process.

To further validate independence amongst inter-arrival times, for all of the tested servers, autocorrelations, shown in Fig. 3 ,are not significant either in magnitude or in too frequently positive/negative. These figures can be viewed as an analog of confidence interval depicting how wide the autocorrelations spread out from 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 provides a percentage of points that pass the independence test.

5 Conclusion and Future Work

In this report, we have tested the inhomogeneous Poisson process by validating if the inter-arrival 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 independent exponen-

	IP 5122	IP 5226	IP 5264
K-S Test (a=1%)	84% Passed	93% Passed	97% Passed

(i) K-S tests summary where level of Significance is 99%

	IP 5122	IP 5226	IP 5264
K-S Test (a=5%)	83% Passed	63% Passed	82% Passed

(ii) A-D tests summary where level of Significance is 95%

Table 1: Test results

	IP 5122	IP 5226	IP 5264
Independence of Inter-arrival Times	96.5% Passed	99.2% Passed	98.2% Passed

Table 2: Percentage of data passing the independence test

tial distribution.

At this workshop we investigated that normal session traffic can be modelled by piecewise inhomogeneous Poisson. However, there are a few other questions to be answered as to further this anomaly detection study such as: how to extend inhomogeneous model on the other servers; to develop an algorithm by periodicity so as to forecast future arrival pattern; 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.

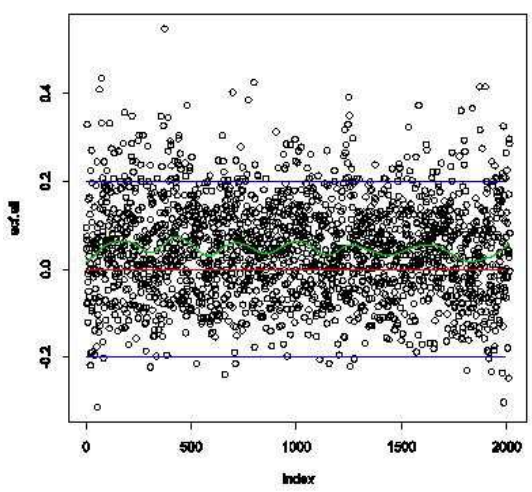
References

[1] Cisco Systems Inc., The Return on Investment for Network Security. Availabe at "http://www.cisco.com/warp/public/cc/so/neso/sqso/roi4_wp.pdf"

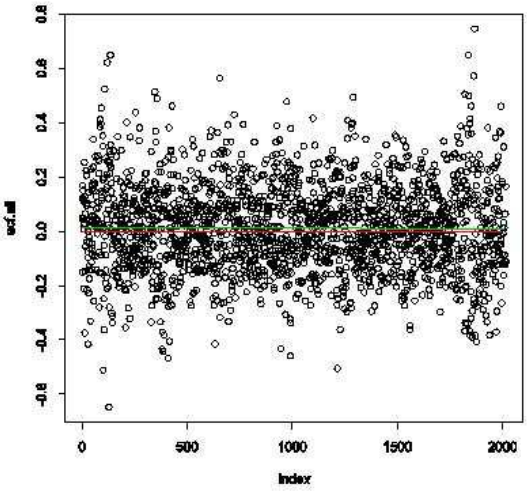
[2] T.W. Anderson and D.A. Darling, Asymptotic theory of certain goodness-of-fit criteria based on stochastic processrs, Ann. Math. Statist. 23(1952), pp. 193-212.

[3] T.W. Anderson and D.A. Darling, A test of goodness-of-fit, J. Amer. Statist. Assoc. 49(1954), pp. 765-769.

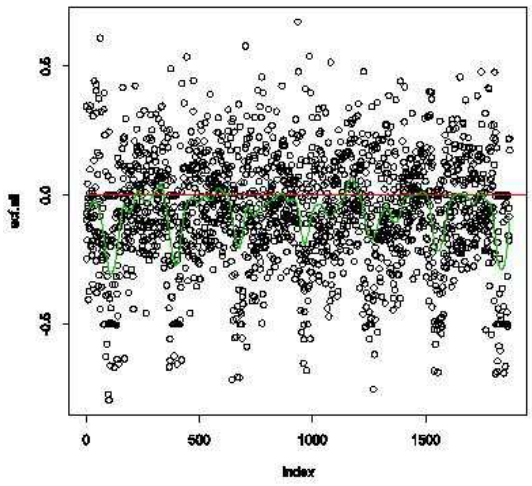
[4] Ljung, G. M. and Box, On a measure of lack of fit in time series models, Biometrika, Vol. 65, pp. 553-564, 1978.



IP 5122 port 80



IP 5226 port 80



IP 5264 port 80

Figure 3: The autocorrelation of the session inter-arrival time of different marks.