A Chi-square testing-based intrusion detection Model

Nasser S. Abouzakhar and Abu Bakar

School of Computer Science, The University of Hertfordshire,
College Lane, Hatfield AL10 9AB, Hertfordshire, UK
{N.Abouzakhar, A.Bakar}@herts.ac.uk

Abstract

The rapid growth of Internet malicious activities has become a major concern to network forensics and security community. With the increasing use of IT technologies for managing information there is a need for stronger intrusion detection mechanisms. Critical - mission systems and applications require mechanisms able to detect any unauthorised activities. An Intrusion Detection System (IDS) acts as a necessary element for monitoring traffic packets on computer networks, performs analysis to suspicious traffic and makes vital decisions. IDSs allow cybercrime forensic specialists to gather useful evidence whenever needed. This paper presents the design and development process of a Network Intrusion Detection System (NIDS) solution, which aims at providing an effective anomaly based detection model using Chi-Square statistics. One of the design objectives in this paper is to minimise the limitations of current statistical network forensics and intrusion detection. Throughout the development process of this statistical detection model several aspects of the process of building an effective detection model are emphasized. These aspects include dataset pre - processing and feature selection, network traffic analysis, statistical testing and detection model development. The calculated / output statistical figures of this model are based on certain threshold values which could be used and / or adjusted by a forensic specialist for deciding whether or not a suspicious event took place.

The modelling and development process of this proposed anomaly detection has been achieved using various software and development tools. In this paper we focus on modelling dynamic anomaly detection using the Chi-square technique. It investigates a network traffic dataset collected by CAIDA in 2008 that contains signs for denial of service (DoS) attacks called backscatter. The normal dataset patterns are analysed to build a profile for the legitimate network traffic. Any deviations from these normal profiles will be considered anomalous. The dataset was pre - processed using Wireshark and T-Shark, the detection model was developed using MATLAB for different variants of denial of services attacks and promising results were achieved.

1.0 Introduction

The rapid growth of the Internet and WWW has made life faster and easier. On the other hand however, new types of crimes made their appearance and made life insecure as well. The growing dependence on the Internet has led to the appearance of various security problems and unpleasant incidents such as cyber attacks and intrusions. An intrusion into a network system is an unauthorised activity that compromises its security (such as integrity, confidentiality and availability) through a series of illegitimate events. To ensure integrity, confidentiality and availability of private information, a computer system or network resource, we need a system that monitors events, processes and actions within a system [1] [2]. Nowadays intrusion detection systems play a significant role in an organization's security infrastructure. The main focus of this section is to describe intrusion detection types, techniques and challenges of current intrusion detection systems. It also covers the problems faced during the dataset pre - processing and feature selection in terms of the techniques used during this phase.

The idea of intrusion detection was first introduced in 1980 by J. P. Anderson and the first intrusion detection model was proposed by D. E. Denning in 1987 [16]. The two major types of IDSs are Host-based IDS (HIDS) and Network-based IDS (NIDS). The HIDS monitor mostly the events on a host computer system, while the NIDS monitor the activity of a computer network system. Intrusion detection can be classified into two detection methods: misuse detection and anomaly detection. Misuse detection or signature based IDS can detect intrusion based on known attack patterns and familiar intrusive scenarios. Anomaly intrusion detection is based on an assumption that the behaviour of an intruder is different from that of normal users. It targets intrusions by identifying the deviation from normal activities behaviour and alerts from potential unseen violations and/or attacks. Anomaly detection systems are divided into two types: static and dynamic. Static anomaly detectors assume that part of the system being monitored will not change such as network protocols. Network traffic data or audit records represent appropriate scenarios for dynamic anomaly detection systems [17] [18].

One of the major security threats is denial of service (DoS) which often compromises the availability of a system or network. DoS attack including its distributed approach is an attempt to exhaust a network or computer resource, so to become unavailable to its legitimate users. Such resources could be network bandwidth, computing power or e-Commerce services [5]. DoS can be achieved by flooding a particular router or network or with an overwhelming traffic and/or by generating huge number of service requests to a server over short period of time. This makes resource and/or services unavailable to legitimate users. Many of the current security measures are no longer considered sufficient to provide reliable network security, especially against zero error malicious activities and intrusions. There are two basic classes of DoS attacks: logic attacks and resource attacks. Logic attacks tend to exploit current software flaws to degrade or crash a particular software system. However, in resource attacks the victim computer's CPU or memory, or network bandwidth are overwhelmed by a large amount of useless traffic and / or requests. There are many methods used to implement denial of service attack. The most commonly methods are TCP SYN flooding, ICMP flooding and RST attack [6] [7] [8] [15].

In this paper we investigate a network security denial of service dataset (called backscatter 2008) captured by the CAIDA to develop our proposed detection model. In backscatter attack, attackers spoof the source IP addresses of live systems selected randomly. The victim node responds to the spoofed source IP addresses. This response behaviour captured by the CAIDA is called backscatter. This dataset does not have traffic between the attacker and the victims. It has only reflections or responses from the victim to the spoofed IP addresses. Therefore, this dataset contains information that is useful for investigating a recent denial-of-service attack [9].

In anomaly detection, normally we have a long - term profile for each user or network system and then compare this profile with recent system events or incoming data. An anomaly event is signalled when there is a larger departure between the observed profile and normal profile. Network traffic data of normal profile / events are required for training the normal profile and other data events (normal and abnormal) for testing purposes [3]. Firstly, the dataset is decompressed using LZO utility into PCAP format. Then, Wireshark is used to visualise the dataset and T-Shark to convert the dataset format to CSV. Finally, the dataset in CSV format is used as an input to Matlab to develop the intrusion detection model. An intrusion detection system should be able to identify a substantial percentage of intrusions while maintaining the false alarms rate at an acceptable level. The major challenge for intrusion detection systems is the base rate fallacy. This is due to the difficulty in maintaining the standard high rate of detections with low rate of false alarms [4]. The base rate fallacy represents both, the false positives and false negatives.

2.0 System Model

In this section we introduce our proposed detailed architecture for our intrusion detection model. The following figure shows a typical generalized architecture of an intrusion detection model. Firstly, the model accepts a network traffic dataset as an input and extracts the TCP flags of each input packet [14]. A frequency distribution is generated and split into four categories as the number of RST, SYN-ACK, ICMP packets per second plus other TCP packets. The average number of packets (for each category) per second is calculated.

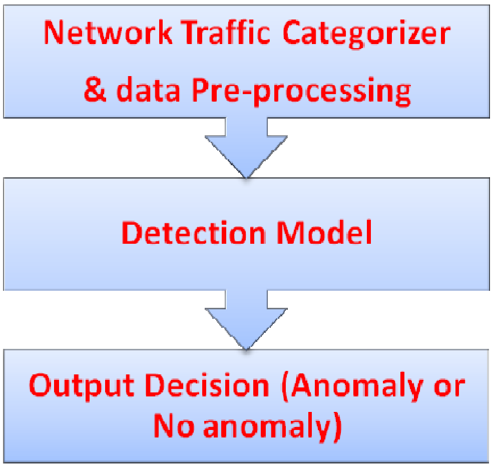

Network Traffic Categorizer & data Pre-processing
Detection Model
Output Decision (Anomaly or No anomaly)

Figure 1: Generalised architecture of Intrusion Detection

To test the model, samples of three minutes of network traffic are extracted from a 3 hours of backscatter-2008 dataset. The dataset has been sampled into 60 samples of three minutes. All samples represent the observed data and are categorised based on the TCP flags. After the three minutes sample categorization the model uses the Chi-square test to perform anomaly detection. In this step a chi-square value is computed from the observed and expected data, and then the chi-square computed value is compared with a chi-square tabulated value. An intrusion alarm is raised when the chi-square computed value is greater than the chi-square tabulated value [11]. Figure 2 shows all these test calculation procedures. The proposed model for the intrusion detection is depicted in this figure.

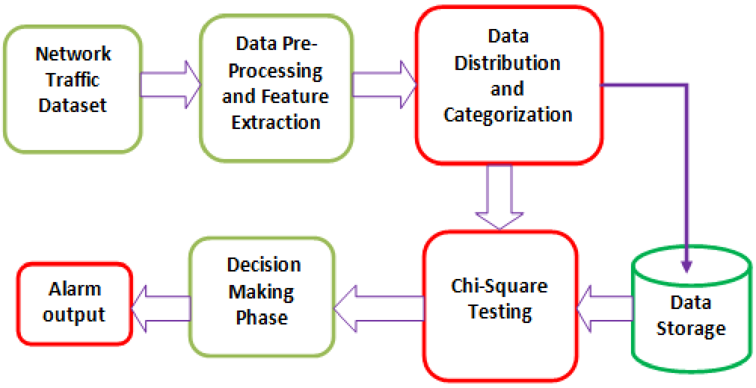


Figure 2: Chi-square Detection Model

As part of the detection model training process the data pre-processing stage splits the input dataset into smaller three minutes durations of PCAP format. In this phase the TCP packets are analysed. Only TCP flags information in the packets are processed and all other information from the dataset is removed. Using the T-Shark utility all worthless data is eradicated, but the remaining data is then converted into CSV format. The CSV format data is manipulated through MATLAB and only the useful features (TCP flags information) are extracted. Next, a frequency population distribution is generated for the whole data. This distribution includes two columns: the first column contains the categories RTS, SYN-ACK, ICMP and OTHER, and the second column contains the average number of packets per second. This generated data distribution is then stored in a data storage.

To test the detection model, and after data pre-processing and feature extraction phase, the data is passed to the next stage of data distribution. In this stage, all the TCP packets are categorized into four categories (RST, SYN-ACK, ICMP, Other). Then the average number of packets per second is calculated for each category. Another distribution is then generated, but this time it is the sample distribution. This sample distribution is also called observed data entries in chi-square testing. As part of the testing process the chi-square calculation is performed on the sample distribution against the stored population distribution. A chi-square value is calculated and then passed to the decision phase. In the decision phase the chi-square calculated value is compared with the chi-square tabulated value, which is also called critical value [10]. If the chi-square calculated value is greater than the critical value then an intrusion / anomaly alarm is raised.

3.0 Method

In the HIDS and NIDS systems a statistical based anomaly detection technique is used to depict the expected normal behaviour of an event [12]. The statistical based anomaly detection techniques overcome the problems associated with string based and rule based misuse detection. Univariate based HIDSs and NIDSs use one specific behaviour measure, however many intrusions use multiple factors and more events having impact on multiple behaviour measures. Thus, a multivariate anomaly detection technique is required to detect such intrusions. Some of these techniques are Hotelling T^2, multivariate cumulative sum (MCUSUM), and multivariate exponentially weighted moving average (MEWMA) [3]. These multivariate statistical methods can be applied for intrusion detection and for examining anomalous behaviour of particular events. These techniques, however, require computationally intensive procedures and processes in order to deal with a huge amount of high - dimensional data. In general, anomaly intrusion detection demands a minimum delay of processing of each event to ensure an early detection of any anomalies [3] [13]. Therefore, a robust multivariate anomaly detection technique with minimum computation cost such as Chi-Square testing would be an appropriate choice for intrusion detection.

Chi-Square Goodness-of-Fit Test Procedure: Chi-Square goodness-of-fit test is used to find out how much the observed values of a particular given sample are significantly different from the expected values of the distribution [19]. It is used to compare the observed sample distribution with the expected probability distribution. Chi-square tests theories about the whole distribution (not an individual parameter – unlike Z and t tests) rather than a single statistic from within that distribution. In chi-square we reject the null hypothesis when there is a difference between the observed and the expected frequencies. To develop the chi-square hypotheses test for the distribution of relevant variables, one must ensure that the following assumptions are fulfilled [11] [19].

- All the expected frequencies are either 1 or greater than 1.
- At most 20 percent of the frequencies are less than 5.
- A sample of data is drawn from a simple - random sampling method so that each possible sample of a given size is equally likely to be the one selected.

The null and alternative hypotheses for the test are:

H_o: the relevant variable has the specified distribution, and
H_1: the relevant variable does not have the specified distribution.

Observed and Expected Frequencies: The numbers of occurrences obtained from the dataset are called observed frequencies and are denoted by **O**. The expected frequencies we expect to obtain from the dataset distribution if the null hypothesis is true are denoted by **E** [10] [11]. The expected frequency for a category is calculated as follows:

$$\mathbf{E} = np \text{ (or } nf\text{)}$$

where n is the size of the sample, p is the probability or proportion of that category if the null hypothesis is true and f is the relative frequency. H_o is rejected, when the observations **O** are sufficiently different from the expected values **E**.

Degree of Freedom for goodness-of-Fit Test [10] [11]: In the goodness of fit test, the degrees of freedom (df) is calculated as follows:

$$df = k-1$$

where k represents the number of categories in the dataset.

The next step in the procedure is the selection of significance level **α**, so to test whether or not the assumptions for the expected frequencies are satisfied. The significance level **α** represents the max risk we are willing to take in rejecting H_o when it is in fact true. For this purpose of chi-square testing the significance level **α** value is decided based on the computer vulnerability [10]. For highly secured networks this value is selected to be small so the results are statistically significant. The Chi-Square test statistic for a goodness of fit test is calculated as follows:

$$\chi^2 = \sum_{i=1}^{k}\left[\frac{(O-E)^2}{E}\right]$$

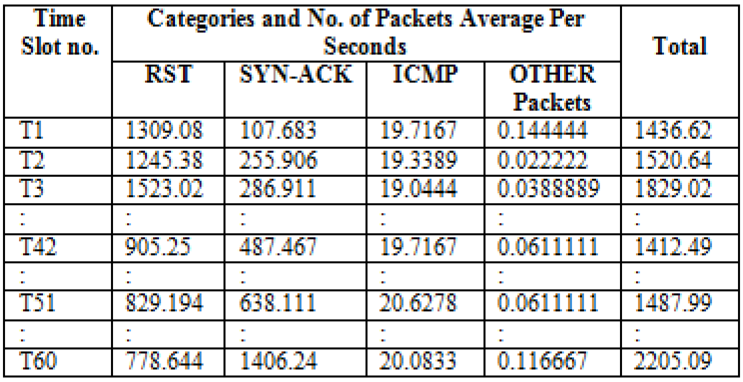
where **O** is the observed frequency for a category, **E** is the Expected frequency for a category and k is the number of observations in the sample (or number of categories in the dataset).

In this test statistic we check the χ^2 calculated value with χ^2 tabulated value, if the χ^2 calculated value is greater than the tabulated value at the significance level $\boldsymbol{\alpha}$ than we reject the null hypothesis H_o. This means that the observed value cannot be fitted in the distribution.

4.0 Experiments and Results

In this section we describe the main steps of experiments carried out. This includes the data sampling, time slots selection and relevant variables categorisation in terms of TCP and ICMP packets. Also, this section describes the achieved results in terms of the detection model outputs.

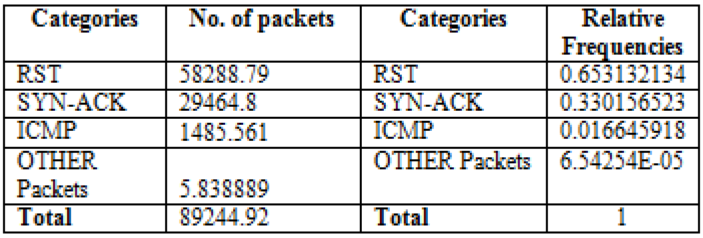
Chi-Square Goodness-of-Fit Test Calculation: In this experiment we have the three hour Backscatter-2008 dataset. Table 1 shows a sample of this data set in 60 time slots, each with 3 minute time slot. Table 1 also depicts the number of packets average per seconds, its categories based on TCP flags set and ICMP packets.



Time Slot no.	Categories and No. of Packets Average Per Seconds				Total
	RST	SYN-ACK	ICMP	OTHER Packets	
T1	1309.08	107.683	19.7167	0.144444	1436.62
T2	1245.38	255.906	19.3389	0.022222	1520.64
T3	1523.02	286.911	19.0444	0.0388889	1829.02
:	:	:	:	:	:
T42	905.25	487.467	19.7167	0.0611111	1412.49
:	:	:	:	:	:
T51	829.194	638.111	20.6278	0.0611111	1487.99
:	:	:	:	:	:
T60	778.644	1406.24	20.0833	0.116667	2205.09

Table 1: Time Distribution of Backscatter-2008 Data Set

For the Chi-square test calculation, a distribution for the whole dataset based on the TCP flag bit-set packets and the ICMP packets was produced, as shown in table 2a. The average packets per second for each category of packets were listed in table 2a as well. Table 2b shows the relative frequencies / ratios of the distribution needed for the remaining calculation. The relative frequencies can be easily calculated by dividing the average number of each category by the total numbers of categories.



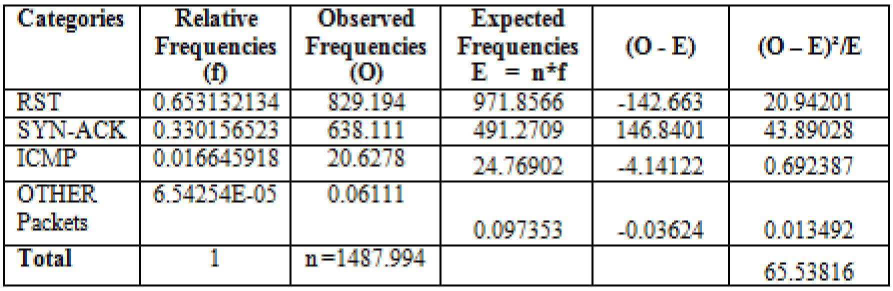
Categories	No. of packets	Categories	Relative Frequencies
RST	58288.79	RST	0.653132134
SYN-ACK	29464.8	SYN-ACK	0.330156523
ICMP	1485.561	ICMP	0.016645918
OTHER Packets	5.838889	OTHER Packets	6.54254E-05
Total	89244.92	**Total**	1

Table 2a: Packets distribution of the Backscatter-2008 dataset

Table 2b: Relative frequencies of the Backscatter-2008 dataset

In order to test the model's ability to detect any anomaly in any of the three minute slots, for example slot no. T51, we can derive the following hypothesis for this test as follows:

H_o: the T51 has the specified distribution i.e. there is no anomaly in T51, and
H_1: the T51 does not have specified distribution i.e. there is anomaly in T51.



Categories	Relative Frequencies (f)	Observed Frequencies (O)	Expected Frequencies E = n*f	(O - E)	(O – E)²/E
RST	0.653132134	829.194	971.8566	-142.663	20.94201
SYN-ACK	0.330156523	638.111	491.2709	146.8401	43.89028
ICMP	0.016645918	20.6278	24.76902	-4.14122	0.692387
OTHER Packets	6.54254E-05	0.06111	0.097353	-0.03624	0.013492
Total	1	n=1487.994			65.53816

Table 3: The χ^2 test calculation for T51 time slot

The χ^2 goodness-of-test statistic is

$$\chi^2 = \sum_{i=1}^{k}\left[\frac{(O-E)^2}{E}\right] = 65.54$$

Let us perform the hypothesis test at 5 % significance level so (α=0.05). There are 4 types of categories in the test so k = 4 and the degree of freedom df = 4 - 1 = 3. By checking the chi-square table and using α=0.05 and df = 3, we get the chi-square tabulated $\chi^2_{0.05}$ value as 7.82. So, the chi-square calculated value is greater than the chi-square tabulated value, therefore we reject the null hypothesis H_o and accept the alternative hypothesis H_1. This means that the time slot T51 is anomalous. In others words, we can safely say that there is denial-of-service attack during the T51 slot. It also means this that the observed entry is different from the expected entry [3] [19]. Nong Ye and Qiang Chen [13] stated that "the large difference between the observed and expected frequencies is an intrusion”. The difference between observed and expected frequencies for the T51 can be depicted in figure 3.

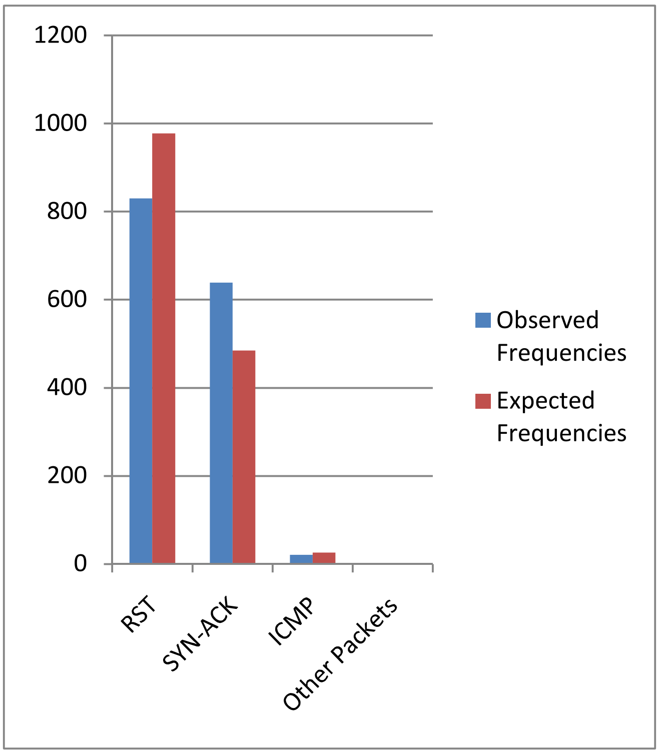


Figure 3: The differences between observed and expected frequencies in T51 slot.

Figure 3 shows that there is a clear difference between the observed and expected frequencies of the RST and SYN-ACK. This is an indication of a SYN flood attack and RST attack during the time slot 51. The purpose of testing T51 slot was to check for any anomalous events in this portion of backscatter-2008 dataset. The backscatter-2008 dataset for non - intrusive events during the time slot T42 has been tested as well. The calculated chi-square value is less than the critical value, so we cannot reject the null hypothesis H□ for this particular time slot. The acceptance of the null hypothesis means that there is no intrusive traffic in backscatter-2008 data during the time slot T42. This is due to the fact that there is a little difference between the observed and expected values, so we conclude that there is no intrusion at the T42 time slot. The calculated chi-square values for the backscatter-2008 dataset for all time slots from time slot T1 to T60 is shown in figure 4.

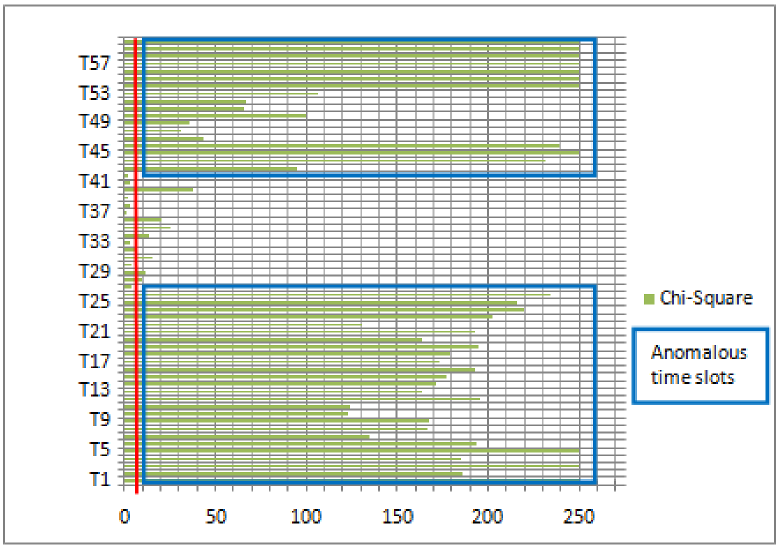


Figure 4: The detection model output

In figure 4 the X-axis represents the calculated chi-square values of the backscatter-2008 data set during the time slots T1 to T60. For the purpose of simplicity the large calculated chi-square values have been capped at 250. The chi-square tabulated threshold value is represented by the red vertical line which intersects with the X-axis at 7.82. All the chi-square values that are greater than the threshold value of 7.82 represent anomalous time slots.

5.0 Conclusion

In this paper we explored the concept of chi-square statistic in terms of detecting anomalous activities taking place in a computer network traffic and CAIDA backscatter-2008 dataset in particular. The paper also described the unique nature of the CAIDA dataset and explained the advantages of chi-square statistic in intrusion detection. Backscatter-2008 dataset analysis is quite challenging, as this data does not contain any direct traffic between the attacker(s) and victim(s). The unique property of the dataset is that it is only one way traffic as it has only reflection or responses from the victim node / network. The developed model, experiments and results analysis confirmed that chi-square is an interesting choice for statistical testing to detect various attacks in computer network systems.

The denial-of-service (DoS) attacks, such as TCP-SYN flood, ICMP flood and RST attack have been investigated. DoS properties have been studied and analysed through extracted features from network traffic and protocol header data such as TCP flags. Various software and tools were used including Wireshark, T-shark, LZO utility to pre-process the dataset. Matlab was used to develop the detection model and for coding development and implementation of chi-square statistic solution. The intrusion detection model performance depends on the input data distribution and its categorization. If the population distribution has been developed through proper statistical approach then the detection model should work well.

References

1 Innella P. McMillan O. (2001) An Introduction to Intrusion Detection System. http://www.securityfocus.com/infocus/1520 (visited August, 2009).

2 Javitz, HS, Valdes A. The SRI statistical anomaly Detector. Proceeding of the 1991 IEEE Symposium on Research in security and Privacy, May 1991.

3 Nong, Ye. Chen Q. An anomaly detection technique based on a chi square statistic for detecting intrusion into information System. Qual. Relib Engng. Int. 2001; 17: 105-102

4 Stalling, W. (2006) Cryptography and Network Security. Upper Saddle River, NJ 07458. Prentice Hall.

5 Seacord R. (2006). Secure Coding in C and C++. Upper Saddle River, NJ 07458. Pearson Education, Inc.

6 Ignatenko, O. Denial of service attack in the internet: agent-based intrusion detection and reaction. http://arxiv.org/PS_cache/arxiv/pdf/0904/0904.4174v1.pdf (visited August, 2009).

7 Manion, A. Pesnate L. Weaver G. (October 2001). Managing the Threat of Denial-of-Service Attacks. CERT Coordination Center. V10

8 Stevens W. R. (1993) , The Protocols. Boston, MA.TCP/IP illustrated (Vol.1). Addison-Wesley Longman Publishing Co., Inc 1993.

9 The CAIDA Backscatter-2008 Dataset - used in August, 2009, Colleen Shannon, David Moore, Emile Aben, and kc claffy. http://www.caida.org/data/passive/backscatter_2008_dataset.xml (visited August, 2010).

10 Mann, P. S. (2004) Introduction to Statistics. 5th Edition. Printed in the United States of America. Johan Wiley & Sons. Inc

11 Goonatilake, R. Herath, A. Herath S. Herath S. Herath J. (2007) Intrusion Detection Using Chi-square Goodness-of-Fit Test for Information Assurance, Network, Forensics and Software Security. Consortium for Computing Sciences in Colleges.

12 Javitz HS, Valdes A. The NIDES statistical component description of justification. *Technical Report A010*, SRI International, Menlo Park, CA, March 1994.

13 Jou, Y. Gong, F. Sargor, C. Wu, X. Wu, S. Chang , H. Wang, F. Design and implementation of a scalable intrusion detection system for the protection of network infrastructure. *Proceedings of the DARPA Information Survivability*

Conference and Exposition. IEEE Computer Society: Los Alamitos, CA, 2000; 69–83.

14 Postal J. RFC793 – Transmission Control Protocol. DARPA Internet Program for Protocol Specification. September 1981

15 Jeremy (2004) http://www.kerneltrap.org/node/3072 (visited August, 2009)

16 D. Yang, A. Usynin, J. W. Hines, "Anomaly-Based Intrusion Detection for SCADA Systems", *5th Intl. Topical Meeting on Nuclear Plant Instrumentation, Control and Human Machine Interface Technologies (NPIC&HMIT 05)* , Albuquerque, NM, Nov 12-16, 2006.

17 Chebrolu S. Abraham A. Thomas J (2004) Feature Deduction and ensemble design of Intrusion Detection System. Computer & Security (2005). 24, 295 – 307

18 Schneier B. Secerets & Lies, John Wiley & Sons, Inc., New York 2002.

19 Arthur Aron, Statistics for Psychology, 4e, 2006.