Volume 4, Issue 9, September 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Study on Statistical Analysis on Security Attack Logs

Rajesh Kumar
M.Tech Scholar
PTU Regional Center (SBBSIET) India

Er. Tajinder Kaur
Assistant Professor
SBBSIET Padhiana (Jalandhar), India

Abstract— The attack data have been collected with the help of various security devices and which is now becoming very huge and diverse, thereby to analyse the collected data manually is just impossible in current trend. The attacks logs can be collected through many ways such as Honeypots, Malware collectors, Distributed Deployment of Intrusion detection system. The volume of collected logs is increasing, thereby there is a need to analyse the data optimistically through some good class of intelligent statistical algorithm to get the inference from the data. With the continuous increase in collected raw data of network attacks, there is a need to perform systematic analysis to extract the useful and relevant information from huge amount of traffic data sets to assist the security analyst. Here in this paper, we describe the analysis process and methods based on statistical analysis techniques which provide us the internals about data set collected on Honeynet or on some other collection testbed. The suitable techniques to apply the intelligent algorithms on the collected data set are being presented. Finally we conclude that based on the applied analysis techniques, the inference results can be further used by any security analytics to get the intelligent and useful information from the raw data.

Keywords— Network Security, Data Mining, Statistical Analysis, IDS, Network Traffic.

I. INTRODUCTION

In today's life, everybody is directly or indirectly affected by the internet, computer networks as the applications on internet are becoming so popular that internet is creating the space for everybody's life. Almost on daily basis, there is security incident reported by the security companies, for example numerous, from individual user's information loss, to worms and computer viruses, to large scale criminal behaviour precipitated by organized crime and nation states.

With the increase usage of internet, so does the potential thread to the global information infrastructure to increase. To protect the cyber world from various kind of spreading attacks such as botnets, DDoS attacks, there are lots of computer security techniques intensively studied in the last decade, namely cryptography, firewalls, anomaly and intrusion detection. The network infrastructures can be protected by placing these security techniques in any suitable place in an organization. Network Intrusion Detection System (NIDS) is also one of the security devices which is places in a network to protect the attacks spreading in an organization.

During the past few years, due to exponential growth in the field of computation network and increase in the volume of population connected to the internet, the data resources and the collection of data is becoming very vast which is currently cannot be handled by the traditional cyber security tools and techniques, thereby there is a need to implement the basic scientific research in the field of the cyber security. As per the report of by the National Research Council [1-3], scientific research and implementation it can produce a better understanding of why cyberspace is as vulnerable as it is and that such research can lead to new technologies and policies and their effective implementation, making cyberspace safer and more secure."

As stated above the data sets in science is growing every year and same in the field of the cyber security, which is directly reflecting our understating of the data with the subject knowledge of science, there is a critical need of applying the good class of intelligent statistical algorithm to get the inference from the so vast collected data set [4].

A. The significance of the problem

In the field of cyber security and its applications, there is a need to collect the attack traces and log them for scientific analysis which provide the factual and analytical information to defend against the cyber incidents As a fundamental building block of repeatable science, we see the lack of freely available raw data as an issue that is both critical for success and a problem that can be addressed through better mathematical modelling and techniques. Here in this research, we propose deal with the problem of attack data collection through deployment of baits system in the form of honeypots or honeynet. Further the logs collected on the honeypots will be processed through good analytical or data mining algorithm to get the statistical information. In this research, we deals the problems of : 1) the need for real-world data on which to base network models, 2) the need for developing analysis methods to get the inferential and statistical information, and 3) the need to handle large amount of security data and process the huge logs of collected data set.

Collection of Malicious Network Traffic:

The experimental study of attacks on the Internet is a very active research domain, and it has gained a lot of attention in recent years. Many valuable initiatives exist for capturing or monitoring malicious activities. Broadly speaking there are two approaches used to monitor unsolicited traffic.

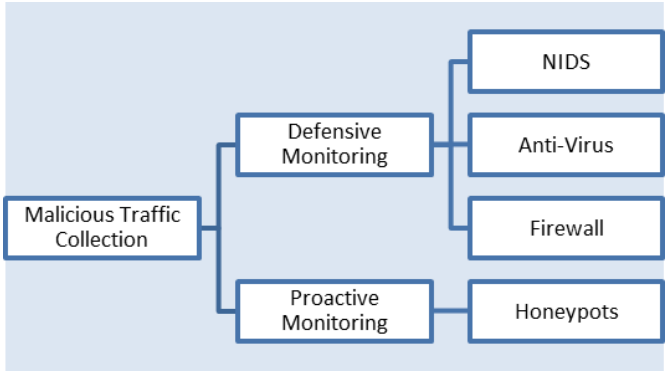

Figure 1: Detection techniques

Before we pursue to apply the statistical techniques in network security or on the collected attack data through different methods, we need to understand the basic fundamental of network security as well as tools and techniques used in network security to collect the network attack logs. Basically as per global scenario, the network security is commonly understood to use the following network security devices to protect the network:

- Firewalls
- VPNs
- AV products
- Intrusion Detection System (IDS)

Firewalls:

In terms of network connectivity, a firewall is basically a software or hardware which control the inbound and outbound traffic based on the analysis of packets and take the decisions based on the applied rule set weather the traffic packets should be allowed or blocked. A firewall acts as a trusted wall between the internal secured network and external network which is not assumed to be secured network

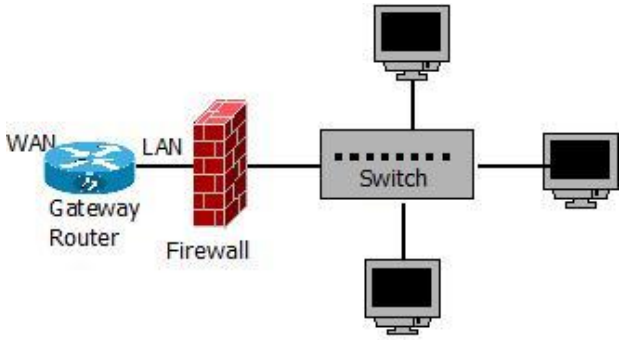

Figure 2: Placement of the Firewall

VPNs

The extension of the private network across the public network such as the internet is performed through technology known as virtual private Network (VPN). The data can shared across the public network just like it is a trusted private network. The connectivity of the remote site or users is performed through this technology that uses the public network usually known as internet. It uses the virtual connections which are routed through the public internet for connectivity of the remote sites or users.

Anti-virus

Anti-virus can be defined as a software program or set of programs which are designed to prevent the computer from the virus or other malicious software which are dangerous to the functionality of the computer. These malicious programs which harm the normal functionality of the computer can be in the form of Trojan, worms, bot programs etc. To prevent the computer, installed anti-virus software should be up-to-date because the computer without the anti-virus installed will be infected within a minutes of connecting to the public internet.

IDS

An intrusion detection system (IDS) monitors the network traffic for suspicious activities in the network. An IDS can be defined as a device or software application program which inspect the complete inbound and outbound traffic for any suspicious patterns in the traffic which may indicate a network or system attacks from a intruder.

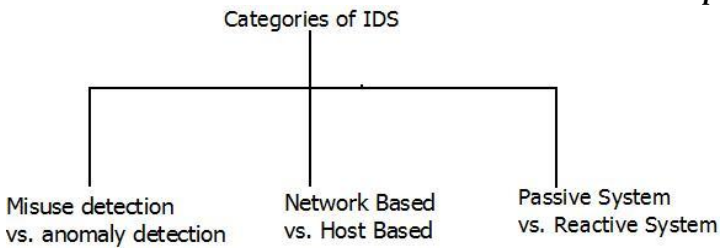

Figure 3: Categories of IDS

The placements of these four network security devices as discussed above are critical to any organization to protect the network from outside attacks. However the collection of logs through these devices is becoming vital for any organization to analyse the attack occurred on the network and to defend against the attacks.

B. Statement of the problem

With the induction of advance data collection tools and techniques and increase in volume of collected attack data, there is need to apply alternative analysis techniques from the field to statistics which will help in advance and systematic analysis mechanisms to perform the systematic analysis of collected attack data. As the size of the collected data is increasing, there is a need to find, discover, and utilize alternative analysis techniques in the field of network security to analyse security logs. As the security data collection tools are growing and continuous to improve, the quantity of the collected data is increasing exponentially. This collecting and utilizing the techniques to analyse the data is important as collecting the attack data.

C. Significance of the problem

Based on the study of the various data collection tools and devices which should be placed in any organization to protect the network, it is necessary to put the proper analysis mechanism to analyse the network security logs. The main contribution of this research is as followings:

- Study of various security data collections, tools and techniques.
- To identify the discipline which can be applied in security?
- Identity the tools and techniques for systematic analysis of security logs
- Develop a report which can be useful for any security researchers, network administrators.

The format of the remaining paper is: section 2, defines and explains the technology that has been employed and discusses the background and motivation in brief and other detection approaches. Section 3 deliberates the tools and techniques as well as requirement of implementation of systematic statistical analysis mechanism. Section 4 discusses conclusion and future work of the research problem.

II. BACKGROUND AND MOTIVATION

A Attack Data Collections tools and techniques

Client Firewall

Most recent operating systems come with built in and "enabled by default" firewall package. Starting with Windows XP service pack 2 and since, firewall has been enabled by default on all Microsoft operating systems. This provides basic protection for an average home user. Based on a latest study in European Union countries in 2010 [5], less than 50 percent of the users had their firewalled enabled. Users decide to disable windows firewall because of compatibility issues with other programs. This results in significant threat to the security of the host system.

Antivirus

Antivirus software is the basic security tool installed in end user computer. They mostly rely on signature based detection where executable files are matched against a signature database of known viruses. New versions have run-time scanning feature that scans the file in real time and avoids execution, if a threat is detected. Signature based detection however results in the antivirus engine failing to detect variants of known viruses, therefore a constant update of antivirus signature database is essential to provide basic protection.

BACKGROUND STUDY OF VARIOUS HONEYPOT SOLUTIONS FOR ATTACK DATA COLLECTION

The domain of network security can be categorized broadly into two categories known as defensive security and proactive network security. In terms of defensive network security, most of the traditional security devices such as firewall, Intrusion detection system and encryptions are working on this principle to protect the network from some misuse based on the pre-determined or pre-configured rules or patterns also known as signatures. In the black hat community which is formed by the attackers, this community is becoming very intelligent in terms of evading the signature based approach. For example, the traditional signature based approach in the form security devices or application security software's have a less capabilities to defend and to protect against the unknown malwares [6-10] which are spreading in the network. In the last few years, it has become more and more clear that these traditional, network-based defence techniques have severe limitations.

Thereby there is need for proactive based approach in the form of honeypots or honeynet which are closely monitored computing resources to be probed or attacked by the attackers. The attacker's activities are being logged into the honeypot which are further analysed by the security experts to get the clue of the attacks or incident performed by the attackers.

There are several possible ways to classify honeypots. Some of the more popular are by the level of interaction available to the attacker, the type of data collected, and the type of system configuration [11-14].

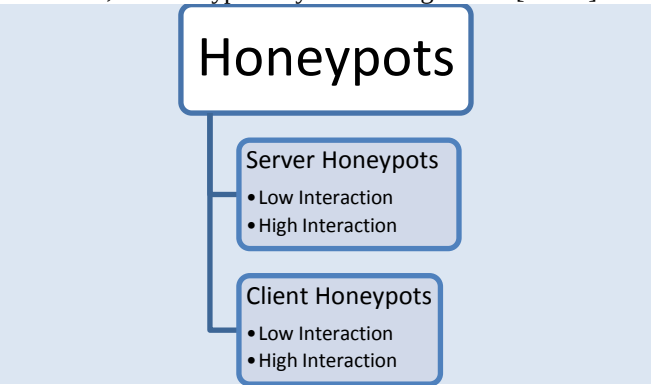

Figure 4: Classification of Honeypots

Honeypots can be classified into two main categories. Firstly, they can be based upon their level of interaction with an attacker. This can be further categorized as:

- Low-interaction: Emulate a variety of host services. These mimic real services but are implemented as a sandbox environment and run as an application. E.g. honeyd [Provos, N and Holz, T (July 26, 2007). Virtual Honeypots: From Botnet Tracking to Intrusion Detection. US: Addison-Wesley Professional.] And nepenthes.
- High Interaction: Attacker is given the freedom to interact with a real operating system and their every attempt is logged and accounted for.

The second Honeypot category is identified by the way they are deployed in a network. This includes:

- **Production Honeypots:**
  They are placed within an organization's production network for the purpose of detection. They extend the capabilities of intrusion detection systems. Such Honeypots are developed and configured to integrate with the organization's infrastructure. They are usually implemented as low-interaction Honeypots sitting within the server farm, but implementations may vary depending on available funding and requirements of the organization.
- **Research Honeypots**:
  These are deployed by network security researchers – the *white hat hackers*. They allow complete freedom for the attacker and, in the process; it is possible to learn their tactics. Using Research Honeypots zero day exploits, Worms, Trojans and viruses propagating in the network can be isolated and studied. Researchers can then document their findings and share them with system programmers, network and system administrators, various system and anti-virus vendors. They provide the raw material for the rule engines of IDS, IPS and firewall systems.

Tools and Techniques used in data collection and its analysis:

Table: Summary of Discipline and Techniques

Data Sources	**Subject Domain**	**Techniques**
Honeytrap [15]	Low Interaction Honeypot	Preprocessing and Data Cleaning
Flow data R-Project [16]	Data & Text Mining	Detection
Weka [17]	Machine Learning algorithms	Classifications
R-Project Orange [18]	Statistical, Data Mining	Pattern Identification and Trend analysis
R-Project	Statistics and Machine Learning	Prediction
Argos [19]	High Interaction Server Honeypot	Server Side Honeypot
HiHAT [20]	High Interaction Server Honeypot	Server Side Honeypot

III. METHODOLOGY

Here we discuss the processes which we have adopted during the collection of attack data and their analysis. Firstly we collect the security logs from various sources of data collection tools, then pre-process the data to extract the useful information from it as well as to apply the any good class of machine learning and statistical algorithm on it. For the sake of the research implementations and applying the mathematical statistical algorithms on the processed data, we have implemented the honeypot technology. With the help of honeypots, which is basically a bait resource or computer developed for the purpose of attacks data collection only, we will be able to collect the attacks logs in the form of network traffic and downloaded malware samples. Further these collected network security logs can be processed to make it in the form of processed data such as CSV, Excel, or arff format so that any scientific researcher will be able to apply the good class of analytical algorithm on the processed data set. At the end of the complete research, the outcome in the form of inferential will be helpful for the security/network administrators, major incident response etc.

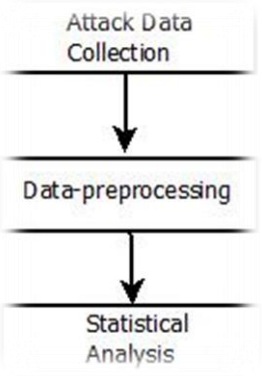

Figure 5: Process Building Blocks

Data Analysis Tool:

R-Project

R is a free software environment for statistical computing and graphics. It runs on a wide variety of UNIX platforms, Windows and MacOS.**R** is a free software programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls and surveys of data miners are showing R's popularity has increased substantially in recent years.

R is a sophisticated statistical software package, easily installed, instructional, state-of-the-art, and it is free and open source. It provides all of the common, most of the less common, and all of the new approaches to data mining.

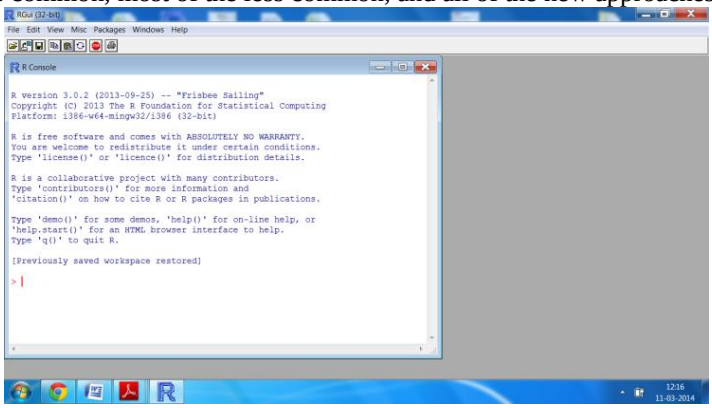

Figure 6: R-console

```
> tcp_data1
  X80 X123 X4613 X2529 X6000  X
1 IP1   40     0    15     0  0
2 IP2    0    12    14     0  0
3 IP3    0     0    35     0  0
4 IP4    0     0    34     0 17
>
```

Figure 7: Sample of R-data

Above figure depicts the sample data set loaded into R-project. It is basically the data set created manually from honeypot traffic which includes the 5 TCP ports as 80, 123, 4613, 6000. This is extracted for study purpose to test the statistical algorithms on network data.

Poison Distribution

Depending upon the usage and requirements as well as on the goal what we want to achieve, there are tons of statistical and data mining techniques embedded with R tool that we can use as per our requirements. Here we just wanted to do principal component analysis of the selected data set, correlations and generate some histograms.

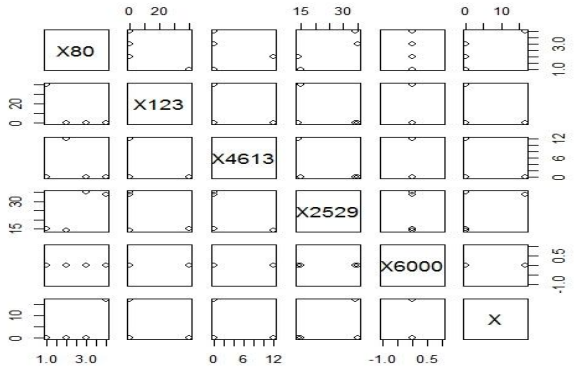

Figure 8: Plot of Data set

IV. CONCLUSION AND FUTURE WORK

In this paper, we presented some of what we believe are the most important problems in cybersecurity for open science environments and highlighted those areas where mathematics and statistics could provide new approaches and solutions. The use of mathematics and statistics in this field is relatively new and much remains to be done. We also believe that the type of analysis needed to address problems in cyber security will likely come from the use of non-traditional methods or techniques. Here we only presented the overview and need of statistical analysis in cyber security to analyse the collected logs. Further we propose to implement the whole system as complete research of this prototype implementations. Hereby we propose to design and development of automatic attack data capturing mechanism through open source tools and development of data processing engine for the purpose to data formatting of collected data set and then applying an optimized classical statistical algorithm with the help of open source tools.

ACKNOWLEDGEMENTS

We would like to sincerely thank Er.Tajinder Kaur (Assistant Professor) for her contribution and help in writing this paper.

REFERENCES

[1] Seymour E. Goodman and Herbert S. Lin, editors. Toward a Safer and More Secure Cyber-space. National Academies Press, Washington, DC, 2007. Committee on Improving Cyber-security Research in the United States, Computer Science and Telecommunications Board.

[2] Charlie Catlett (Ed.). A scientific research and development approach to cyber security. Report submitted to the U.S. Department of Energy, 2008.

[3] Daniel M. Dunlavy, Bruce Hendrickson, and Tamara G. Kolda. Mathematical challenges in cybersecurity. Technical Report SAND2009-0805, Sandia National Laboratories, February 2009.

[4] Matt Bishop. Computer Security Art and Science. Addison Wesley, 2003.

[5] Internet usage in 2010 – Households and Individuals

[6] Skoudis, E., and Zeltser, L., *"Malware: Fighting Malicious Code"*, Prentice Hall,2003, Page 3, ISBN = 978-0131014053.

[7] [Provos, N., McNamee, D., Mavrommatis, D. W., K and Modadugu, N., *the Ghost In The Browser Analysis of Web-based Malware*. 2007. [Online]. Available at:http://www.usenix.org/events/hotbots07/tech/full_papers/provos/provos.pdf [Accessed 11 Feb 2009]

[8] Secure Browsing | Malware Protection | Trustwavehttps://www.trustwave.com/securebrowsing/

[9] Google Safe Browsing www.google.com/tools/firefox/safebrowsing/

[10] Detection and Analysis of Drive-by-Download Attacks and Malicious JavaScript Code www.cs.ucsb.edu/~vigna/.../2010_cova_kruegel_vigna_Wepawet.pdf

[11] Reto Baumann and Christian Plattner. Honeypots, 2002.

[12] Amit Lakhani. Deception techniques using honeypots. Master's thesis, University of London, UK

[13] Feng Zhang, Shijie Zhou, Zhiguang Qin, and Jinde Liu. Honeypot: A supplemented active defense system for network security.

[14] www.cs.arizona.edu/~collberg/Teaching/466-566/2012/.../report.pdf

[15] honeytrap – A Dynamic Meta-Honeypot Daemon, honeytrap.carnivore.it/

[16] The R Project for Statistical Computing, www.r-project.org/

[17] Weka 3 - Data Mining with Open Source Machine Learning , www.cs.waikato.ac.nz/ml/weka/

[18] Orange – Data Mining Fruitful & Fun, orange.biolab.si

[19] Argos - An emulator for capturing zero-day attacks, www.few.vu.nl/argos

