Identifying Network Traffic Features Suitable for Honeynet Data Analysis

Mohammed H. Sqalli[1], Syed Naeem Firdous[1], Khaled Salah[2], and Marwan Abu-Amara[1]
[1]Computer Engineering Department
King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia
{sqalli,snaeem,marwan}@kfupm.edu.sa
[2]Computer Engineering (Sharjah Campus)
Khalifa University of Science, Technology & Research (KUSTAR),
P.O.Box 573, Sharjah, UAE
khaled.salah@kustar.ac.ae

ABSTRACT

A honeynet is a solution designed by the Honeynet Project organization to gather information on security threats and it can be used to proactively improve network security. A honeynet captures a substantial amount of data and logs for analysis in order to identify malicious activities and this is a challenging task. The main aim of this work is to identify the best traffic features or parameters that can be used in an anomaly detection technique to identify anomalies in honeynet traffic. In this work, a detailed analysis of feature-based and volume-based parameters is carried out and the most appropriate features for honeynet traffic are selected. Unlike other techniques proposed in the literature, our work combines entropy distributions for feature-based parameters and volume distributions for volume-based parameters to evaluate the different features. The features were evaluated using real honeynet traces released by the Honeynet project organization and other sources.

Index Terms— Honeynet Traffic, feature evaluation, entropy, network security, network forensics.

1. INTRODUCTION

Computer network security is a major area of concern for different people from normal home users to businesses trying to protect their resources from unauthorized access. There is a constant threat from malicious users who are trying to disrupt normal operations or trying to steal sensitive or proprietary information.

The honeynet is a concept designed to gather information on security threats which can be used by the organizations to proactively improve their network security. A Honeynet can be used to assist system administrators in identifying malicious traffic in the enterprise network. By its very nature, a honeynet has no production value and should not be generating or receiving any traffic. Any traffic to or from the honeynet is suspicious in nature[1]. The key requirements to successfully implement a Honeynet, is data control, data capture, and data analysis [1].

Currently, a honeynet gathers a lot of network data and this sometimes makes it difficult to analyze this huge data set. Various types of data are collected based on which type of honeynet is used, e.g., Honeywall, Nepenthes, HoneyD, Dionaea, etc., and each tool has its own format. Various honeypot implementations result in "needlessstack" data overload (too much data and different types of data) and this is one of the main challenges for honeynet analysts [8]. A honeynet's real potential will not be realized until organizations can effectively deploy multiple honeynets and correlate the information they collect.

Honeynets do not include anomaly detection schemes to identify anomalies in the honeynet traffic. The main contribution of this work is to evaluate different candidate features that can be used with an anomaly detection technique to detect anomalies in honeynet traffic. The approach used in this work is to first identify candidate features for honeynet traffic, and then evaluate them to find the features that have the best detection capabilities. Unlike other techniques proposed in the literature, this work is based on the combination of entropy distributions for feature-based parameters and traffic volume distributions for volume-based parameters to evaluate the identified features.

The rest of the paper is organized as follows. Section 2 discuses the related work, in which we briefly summarize the existing research on using feature-based and volume-based parameters for anomaly detection. Section 3 discusses our proposed approach used in this work. Section 4 presents the evaluation of our approach and discussion of the obtained results. Section 5 presents the conclusions and discusses future work.

2. RELATED WORK

The main aim of anomaly detection techniques is to differentiate between normal and abnormal traffic. There has been recent focus on two main categories of detection techniques applied to network traffic: volume-based detection techniques [2] [3] [4] [5], and feature-based detection techniques [6] [7]. Volume-based detection

CCECE 2011 Niagara Falls, Canada
978-1-4244-9789-8/11/$26 ©2011 IEEE

schemes use volume changes to detect anomalies in the network traffic, such as flooding attacks and certain types of DoS attacks. In contrast, the feature-based detection schemes use the distributional changes of packet header details, such as IP addresses and port numbers, to detect anomalies.

Lakhina et al. [6] proposed an anomaly detection method using traffic feature distributions. They argue that distributions of packet features like IP addresses and ports along with the use of entropy are useful in detecting a wide range of anomalies in the network traffic. On the other hand, Nychis et al. [7] conducted an empirical evaluation of using entropy for anomaly detection. The authors mainly focused on the use of entropy for different traffic features and analyzed the detection capabilities of different traffic feature distributions.

In contrast, Kind et al. [8] proposed a new approach to the feature-based anomaly detection of Lakhina et al. [6]. In the proposed approach, the authors created histograms of the different traffic feature distributions and then modeled histogram patterns which were used to detect anomalies. Ping and Abe [4] proposed an IP packet size entropy-based DoS detection scheme in which changes in the IP packet size entropy (IPSE) is used to detect possible DoS attacks. The authors illustrated that the various applications have default packet sizes with respect to the request/response messages. In the presence of attacks, the generated packets are of identical sizes irrespective of the response from the victim.

Thonnard and Dacier [9] proposed a clustering-based approach to detect attack patterns in honeynet data. In their approach, they specifically use time signatures to cluster the honeynet data. They conducted experiments on large data sets collected from 44 worldwide distributed honeypots. The attack source is identified as an IP address that targets the honeypot on a given day with a certain port sequence.

3. ANALYZING HONEYNET TEST DATA

In order to detect anomalies in the honeynet traffic, we first need to analyze different honeynet traffic data sets to understand the difference between the normal and abnormal behavior. Honeynet traces were collected mainly from the honyenet.org site in which the scan of the month (SOM) challenges and Forensic Challenges are released [10]. We also used traces from hack.lu 2009 Information Security Visualization Contest [11] and from the honeynet deployment at KFUPM. The Honeynet traces that were used are listed in Table 1.

The traces provided by the Honeynet organization are instances of real compromises that were captured by different honeynet chapters. These traces were analyzed to identify the suitable characteristics / features that can be used for anomaly detection. In the first step, candidate features were selected (listed in Table 2) by analyzing the honeynet traffic and from literature. Then, these features were evaluated for their detection capabilities.

Table 1: Honeynet Traffic Test Datasets

Traffic Data Set Name & Source	Description	Traffic Details
Pcap Attack Trace, Honeynet.org	The network traffic related to an automated malware attack.	348 packets Duration: 16 sec
Scan 28 - Honeynet.org	Trace collected by the Mexico Honeynet Team - Italian blackhats break into a Solaris server.	Day1: 18843 Packets 24 Hours Day 3: 123123 Packets 24 Hours
Scan 14 - Honeynet.org	This trace is about a successful Windows NT attack.	6707 packets Duration: 20 Hours
Scan 19 - Honeynet.org	Trace of Redhat Linux 6.2 honeypot attack.	24440 packets Duration: 23 Hours
SSH Based Honeypot trace Information Security Visualization Contest - hack.lu 2009	Dataset collected from an SSH based honeypot. Includes anomalies such as network scans, rootkit file transfers, IRC traffic, etc	4323191 packets Total Duration: 12 days

Table 2: Traffic Features used for a Detailed Analysis

Traffic Features	Volume Features
• Source IP Address[6] • Destination IP Address[6] • Source Port[6] • Destination Port[6] • Packet Size Distribution[4] • Indegree & Outdegree[7]	• Average packet inter-arrival time • Total payload bytes received during the interval • Average payload size during the interval • Total packets received during the interval

Entropy distributions were used for feature-based parameters and volume changes or distributions for volume-based parameters. Entropy was calculated using:

$$H(X) = -\sum_{i=1}^{n} P(x_i)\ \log P(x_i)$$

Suppose we randomly observe X for a fixed time window w, then $P(x_i) = m_i/m$, where m_i is the frequency or number of times we observe X taking the value x_i, i.e.

$$m = \sum_{i=1}^{n} m_i$$

$$H(X) = -\sum_{i=1}^{n} (m_i/m) \log (m_i/m)$$

Where:

$H(X)$ = Entropy
m_i = number of packets with x_i as the traffic feature
m = total number of packets

The sliding window concept was used to gather entropy values in overlapping intervals as shown in Figure 1, so that any valuable information is not missed in cases where an anomaly overlaps across multiple intervals.

Figure 1: Sliding window used for calculating entropy

4. FEATURE EVALUATIONS

The candidate features were evaluated based on the traffic distributions seen during the anomalous events. The features were also evaluated based on their ability to differentiate between normal and abnormal traffic. All the honeynet traces mentioned in Table 1 were used for testing the features. For brevity, only the evaluations of SSH based honeypot trace are presented here. The SSH based honeypot trace includes 12 days of collected traffic. This trace mainly includes SSH traffic and many anomalies such as network scans, rootkit file transfers, IRC traffic, etc.

4.1. Single Feature Evaluation

Initially, individual features were evaluated and tested for whether they provide good variations when an anomaly occurs. Figure 2 shows the distribution of destination IP entropy for the SSH trace. The destination IP entropy represents the number of external connections initiated by the honeypot. The peaks indicate that the honeypot initiated a large number of connections to external IP addresses and are related to a network scan initiated by the honeypot.

The packet size entropy illustrated in Figure 3 shows many variations throughout the duration of the trace that are unrelated to the anomalies present in the trace. Therefore, the packet size entropy feature does not aid much in detecting anomalies.

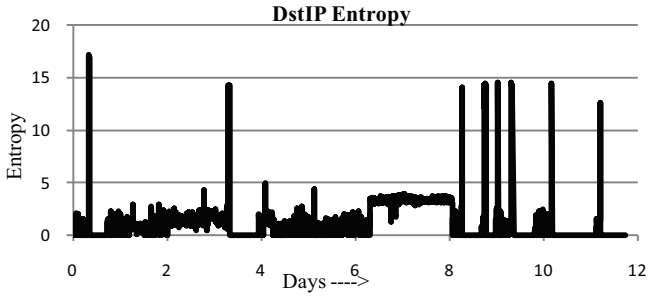

Figure 2: Destination IP entropy distributions of SSH based honeypot trace

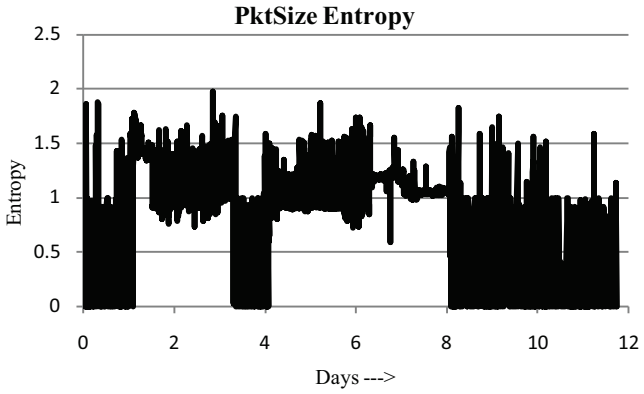

Figure 3: Packet size entropy distribution of SSH based honeypot trace

Volume-based features such as the total payload bytes also lead to variations during the anomalous events. Figure 4 shows that before a network scan event, a large data transfer took place. When we manually analyzed the trace, we found that this was related to a malicious file transfer which was later used to initiate a network scan activity.

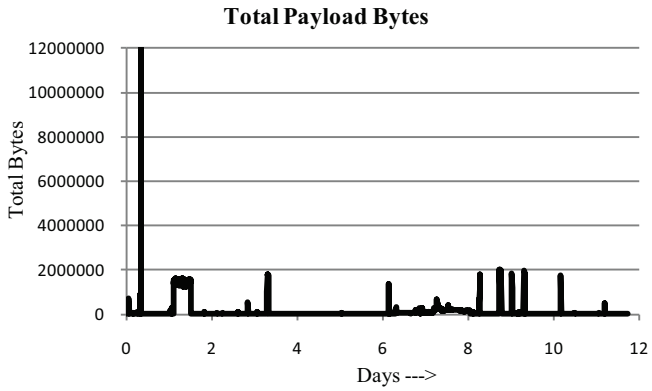

Figure 4: Total Payload Bytes distribution of SSH based honeypot trace

4.2. Evaluation of Two Features Combinations

Using individual features helps only in detecting certain anomalous events and it does not give a clear understanding of the anomaly that occurred. To get a better understanding of the behavior of the anomaly, we need to look into a combination of features. This is useful to detect certain anomalies that were not visible using a single feature. A number of combinations of the above listed features were tested to identify the useful features combinations and get a better understanding of the anomalies.

The combination of the destination IP entropy and the destination port entropy show visible groups, i.e., clusters, which can be used to differentiate between normal traffic and anomalies, i.e., outliers. In Figure 5, the group with a high destination IP entropy and a low destination port entropy indicates a network scan where a large number of IP addresses are being scanned for the SSH port. Figure 6 shows the combination of the destination IP entropy and the Packet size entropy. It shows a lot of variation or grouping along the packet size entropy axis indicating different packet sizes were seen in the trace.

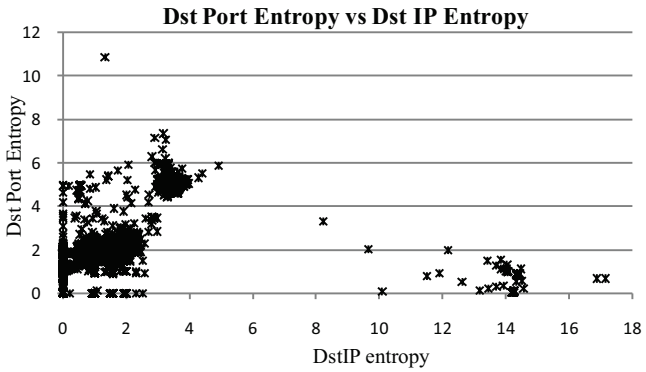

Figure 5: Destination Port entropy and Destination IP entropy combination of SSH honeypot trace

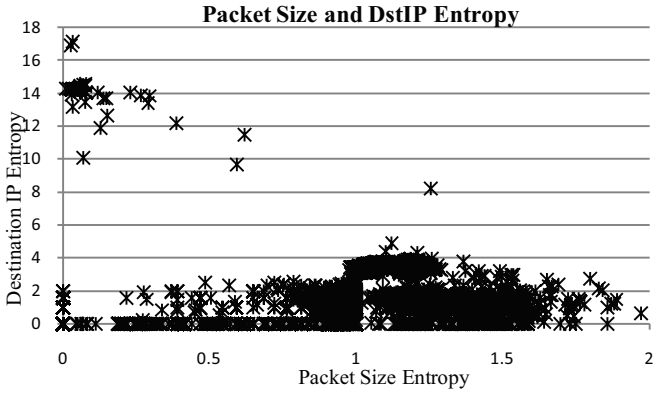

Figure 6: Destination Port entropy and Destination IP entropy combination of SSH honeypot trace

4.3. Evaluation of Three Features Combinations

When combining different features, we can see different patterns that can help us detect anomalous regions as well as normal regions. Using three features helps in getting a better visualization of the different clusters present in the honeynet data. We performed various tests using different combinations of the features to identify those features that provide the best distinction between normal behavior and outliers by showing distinct clusters.

The combination of source IP, destination IP, and destination port entropies is shown in Figure 7. This combination does not show many cluster regions and hence is not very useful.

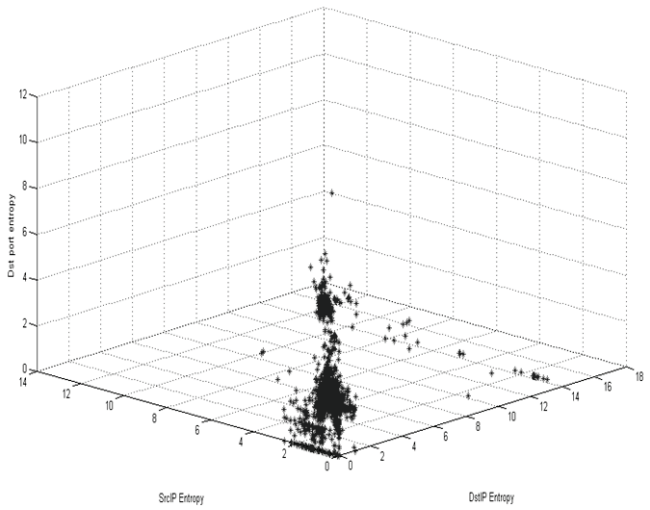

Figure 7: Combination of Destination port, Source IP and Destination IP Entropy values

The combination of source port, destination port, and destination IP entropies shows visible clusters, which can be attributed to different anomalous events. After manually analyzing the traffic, we found that in the Figure 8, cluster 1 with entropy values between 0 and 3 represents the normal traffic. The second cluster represents the scanning by the honeypot for different IRC channels. The third cluster includes a region where there were bruteforce attempts to log into the SSH service running on the honeypot. In this region, the source port entropy is high and the destination port entropy is low since many IPs are targeting the SSH port. The fourth cluster indicates the network scan performed by the honeypot; which scans the SSH port on the destination machines using different ports for each connection. The region closer to zero mostly represents the IRC traffic as there are machines communicating with each other using the IRC ports. Table 3 summarizes the findings of feature analysis by providing the detection capabilities of various features.

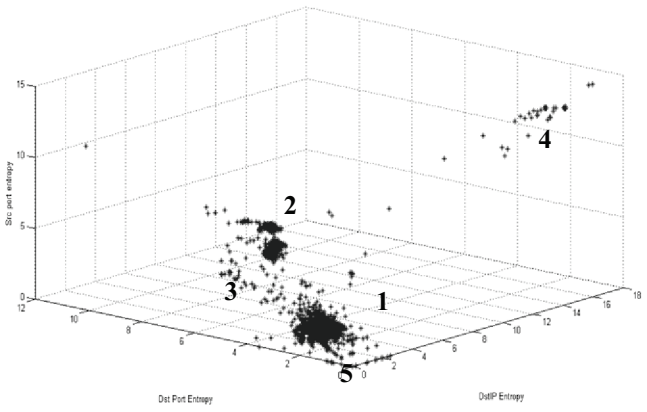

Figure 8: Combination of Destination IP, Destination Port, Source port entropy values

Table 3: Summary of Detection Capabilities of various features

Traffic Feature	Detection Capabilities
Packet Size Entropy	Shows good variations but does not help in understanding the anomaly.
Destination IP Entropy	Shows large variations during specific anomalies and gives good indication of anomaly.
Source IP Entropy	Shows less variations in the traffic compared to destination IP entropy.
Destination Port Entropy	Shows large variations for various anomalies.
Source Port Entropy	Shows large variations for various anomalies.
Average Packet Inter-Arrival Time	Shows good variations but not very useful in understanding the anomaly behavior
Total Payload Bytes	Shows good variations during most of the anomalies and when used with other features gives good understanding of the anomaly
Total Packets	Shows good variations during anomalies and very useful in understanding the anomalies.
Average Payload Size	Shows good variations during anomalies but does not aid in understanding the anomaly behavior.

Based on the features evaluation, we found that the following features are the most appropriate for the honeynet traffic:

- Destination IP Entropy (DIP)
- Destination Port Entropy (DP)
- Source Port Entropy (SP)
- Total Payload Bytes (TB)
- Total Packet Count (PC)

5. CONCLUSION

In this work, we have evaluated feature-based and volume-based parameters that can be used for anomaly detection in honeynet traffic. The candidate features were evaluated using real honeynet traces, and the features that show better detection capabilities were selected. It was discovered that the combination of the destination port entropy, the source port entropy, the destination IP entropy, the total payload bytes, and the total packets provide better detection capabilities for various anomalies. These features can be used to design a technique to detect anomalies in honeynet traffic. This will enhance the data analysis process and will reduce the overall time to analyze the honeynet traffic.

6. ACKNOWLEDGEMENT

The authors would like to acknowledge the support provided by the King AbdulAziz City for Science and Technology (KACST) through the Science and Technology Unit at King Fahd University of Petroleum and Minerals (KFUPM) for funding this work through project No. 08-INF101-4 as part of the National Science, Technology, and Innovation Plan. The authors would like to also thank all Saudi Honeynet Project team members for their feedback, especially Zubair Baig, Farag Azzedin, and Hakim Adiche.

7. REFERENCES

[1] Levine, J., et al. The use of Honeynets to detect exploited systems across large enterprise networks. in Information Assurance Workshop, 2003. IEEE Systems, Man and Cybernetics Society. 2003.

[2] Dainotti, A., A. Pescape, and G. Ventre. NIS04-1: Wavelet-based Detection of DoS Attacks. in Global Telecommunications Conference, 2006. GLOBECOM '06. IEEE. 2006.

[3] Haggerty, J., et al. DiDDeM: a system for early detection of TCP SYN flood attacks. in Global Telecommunications Conference, 2004. GLOBECOM '04. IEEE. 2004.

[4] Ping, D. and S. Abe. Detecting DoS attacks using packet size distribution. in Bio-Inspired Models of Network, Information and Computing Systems, 2007. Bionetics 2007. 2nd. 2007.

[5] Barford, P., et al., A signal analysis of network traffic anomalies, in Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment. 2002, ACM: Marseille, France. p. 71-82.

[6] Lakhina, A., M. Crovella, and C. Diot, Mining anomalies using traffic feature distributions, in Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications. 2005, ACM: Philadelphia, Pennsylvania, USA. p. 217-228.

[7] Nychis, G., et al., An empirical evaluation of entropy-based traffic anomaly detection, in Proceedings of the 8th ACM SIGCOMM conference on Internet measurement. 2008, ACM: Vouliagmeni, Greece. p. 151-156.

[8] Kind, A., M.P. Stoecklin, and X. Dimitropoulos, Histogram-based traffic anomaly detection. Network and Service Management, IEEE Transactions on, 2009. 6(2): p. 110-121.

[9] Thonnard, O. and M. Dacier, A framework for attack patterns' discovery in honeynet data, in Digital Investigation. 2008. p. S128-S139.

[10] Honeynet.org. http://www.honeynet.org/challenges.

[11] hack.lu, Information Security Visualization Contest, hack.lu 2009, http://2009.hack.lu/index.php/InfoVisContest. 2009.