Web Attack Analytics

Author: Luis O. Sanchez
Advisor: Dr. Alfredo Cruz
Electrical & Computer Engineering and Computer Science

Abstract

Cyber attacks are on the rise. Information systems are constantly being attacked, generating great economic loss. Computer networks and web technologies are the preferred places for hackers to commit cyber crimes. Most computers and devices are connected to the Internet through a communication network. On many occasions network security administrators do not monitor adequately the cyber attacks that are occurring through the computer network, making it difficult to protect the organizations' networks. It is important for information security experts to learn how to use data mining tools to analyze cyber attacks in the computer network and to make better decisions regarding security. Based on this need this Master Project has been developed; so that network security administrators can use various programming algorithms to extract important data that transits the network, analyze it, and make decisions that contribute to enhance the protection of networks.

Introduction

One of the main goals of this project is to be able to assist security administrators through the use of tools such as RStudio, and programming code such as R Language. With the use of these tools they can learn to extract important data from datasets. Once security administrators have access to datasets that contain network and monitoring data, they can extract important information from these datasets and analyze it for subsequent decision making. It is crucial to develop effective and efficient countermeasures against attacks on the web and networks.

Background

The information age has changed the mode in which organizations and governments operate [3]. From the typewriter, pen and paper, to computers, devices, and computer networks. Each day organizations and governments integrate new information technologies into their operations in order to be competitive and succeed in the market [1]. But the transition to the information age has increased cybercrime aimed at computer networks, which is the main infrastructure of information systems. The proportion of cyber attacks is increasing daily and new attacks are emerging exponentially, making it difficult for security experts to maintain a safe and secure environment. Many of these experts do not have the tools and knowledge to be able to extract important information from network traffic to recognize the trends of cyber attacks on the network [2]. This makes it difficult for security experts to be able to analyze and make decisions that protect information systems [2]. Through the use of programming tools and coding these experts can provide better monitoring, extract important data, and protect their computer networks from new threats.

Problem

Will Information Security experts who do not have a background in computer science be able to learn how to use software tools and programming languages to extract important network data that will help them improve network security?

Since many information security experts have a background in information systems or information technology rather than computer science or engineering, they lack programming skills. That is why the Author demonstrates through programming examples how to extract and group important data from datasets on cyber attacks to computer networks.

Methodology

In order to do an appropriate data analysis extraction, the number of cyber attacks per each type of attack is extracted in the dataset and then grouped into a single data structure. Once the web attack data from all the datasets are grouped into a data structure, we proceed to apply a data mining algorithm. The first step was to obtain the result of the cyber attacks of each dataset together with its attack classification using the summary function. Below, Table 1 shows the outcomes of web attacks on each dataset.

Table 1 Datasets Summary

```
> NB15_1
      Fuzzers      Analysis     Backdoors           DoS      Exploits
            0          5051           526           534          1167          5409
      Generic Reconnaissance     Shellcode         Worms
         7522          1759           223            24
> NB15_2
      Fuzzers Reconnaissance     Shellcode      Analysis
            0          4668          3116           324           608
     Backdoor           DoS      Exploits       Generic         Worms
          370          4637         11103         27883            40
> NB15_3
      Fuzzers Reconnaissance     Shellcode      Analysis
            0          9137          5582           593           873
     Backdoor           DoS      Exploits       Generic         Worms
          759          5642         16574        118198            67
> NB15_4
      Fuzzers Reconnaissance     Shellcode      Analysis
            0          5390          3530           371           670
     Backdoor           DoS      Exploits       Generic         Worms
          666          4907         11439         61878            43
```

Before applying data analytics to the dataset, research was done on the dataset's data structure and which rows were to be used. Mydatawebattacksfour shows the data of the different types of attacks from the 4 datasets that was grouped in the data frame. For example, the fuzzers column shows the attack data of each data set NB15_1 5,051, NB15_2 4,668, NB15_ 3 9,137, NB15_4 5,390. Next, Table 1 shows the results of the web attacks of the four datasets integrated into a data structure called the data frame.

Table 2 Mydatawebttacksfour Data

```
> mydatawebattacksfour
  Datasets Fuzzers Analysis Backdoors  DoS Exploits Generic Reconnaissance Shellcode worms totalattacks
1  NB15_1    5051      526       534 1167    5409    7522          1759      223    24        22215
2  NB15_2    4668      608       370 4637   11103   27883          3116      324    40        52749
3  NB15_3    9137      873       759 5642   16574  118198          5582      593    67       157425
4  NB15_4    5390      670       666 4907   11439   61878          3530      371    43        88894
```

The main purpose of the project design is the identification and application of statistical algorithms for the extraction of important information that can be useful for a statistical analysis. The summary function was applied to the data frame mydatawebattacksfour to obtain the min, 1st quantile, Median, Mean, 3rd quantile and the max of the results of each column that is made up of the grouping of the 4 attack type datasets. For example the min of Generic column is 7,522, the max 118,198, the median 44,881 and mean 53,870. Another example the min of Exploits column is 5409, the max 16,574, the median 111,271 and mean of 111,31 attacks. Below, Figure 1 contains statistical information of the columns of the mydatawebattacksfour data frame.

```
> summary(mydatawebattacksfour[1:11])
   Datasets     Fuzzers          Analysis        Backdoors           DoS            Exploits
 NB15_1:1   Min.   :4668   Min.   :526.0   Min.   :370.0   Min.   :1167   Min.   : 5409
 NB15_2:1   1st Qu.:4955   1st Qu.:587.5   1st Qu.:493.0   1st Qu.:3770   1st Qu.: 9680
 NB15_3:1   Median :5220   Median :639.0   Median :600.0   Median :4772   Median :11271
 NB15_4:1   Mean   :6062   Mean   :669.2   Mean   :582.2   Mean   :4088   Mean   :11131
            3rd Qu.:6327   3rd Qu.:720.8   3rd Qu.:689.2   3rd Qu.:5091   3rd Qu.:12723
            Max.   :9137   Max.   :873.0   Max.   :759.0   Max.   :5642   Max.   :16574
    Generic      Reconnaissance    Shellcode         worms        totalattacks
 Min.   :  7522   Min.   :1759   Min.   :223.0   Min.   :24.0   Min.   : 22215
 1st Qu.: 22793   1st Qu.:2777   1st Qu.:298.8   1st Qu.:36.0   1st Qu.: 45116
 Median : 44881   Median :3323   Median :347.5   Median :41.5   Median : 70822
 Mean   : 53870   Mean   :3497   Mean   :377.8   Mean   :43.5   Mean   : 80321
 3rd Qu.: 75958   3rd Qu.:4043   3rd Qu.:426.5   3rd Qu.:49.0   3rd Qu.:106027
 Max.   :118198   Max.   :5582   Max.   :593.0   Max.   :67.0   Max.   :157425
```

Figure 1 Summary Mydatawebttacksfour Data Frame

Results and Discussion

This section will explain the results of the project and how it relates to the initial idea that was proposed. From a total of 2,540,047 rows of network packet traffic, a total of 321,283 (12.648%) web attacks were detected. The type of Web attack with the highest number of attacks was Generic with a total of 215,481 (67.06%), and Worms had the fewest attacks with a total of 174 (0.0541%). Below, Figure 2 shows a screenshot of Total Tuples, Total, Max & Min results.

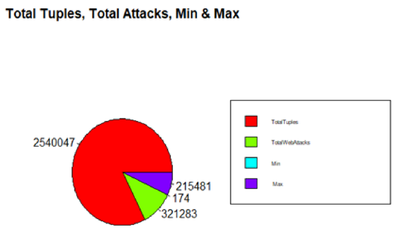

Figure 2 Total Tuples, Total Attacks, Max & Min

The cyber attack with the most occurrence and the least occurrence was selected, and statistical concepts of percentiles were applied. The percentiles of Generic attacks is 7,522(0%), 22,792.75(25%), 44880.50(50%), 75,958.00(75%) and 118,198.00(100%). The percentiles of Worms attacks is 24(0%), 36(25%), 41.50(50%), 49(75%) and 67(100%). Below, Figure 3 shows a screenshot of the most occurring web attack "Generic Quantile Percentages".

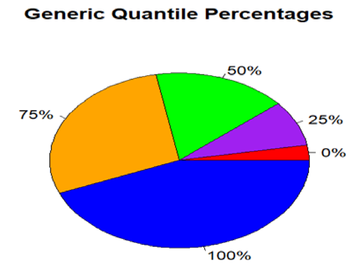

Figure 3 Generic Quantile Percentages

The data frame mydatawebattacks shows the sum of all the results of the attacks of each column from all datasets used in the project. Fuzzers with 24,246, Analysis 2,677, Backdoors 2,677, DoD 16,353, Exploits 44,525, Generic 215,481, Reconnaissance 13987, Shellcode 1,511, Worms 174 web attacks and totalAttacks 321,283. Being the column of generics with the highest number of attacks (215,481) and being the column of worms with at least (174) attacks. Below, Table 3 shows a screenshot of Mydatawebattacks.

Table 3 Mydatawebattacks1 Data Frame

	Fuzzers	Analysis	Backdoors	DoS	Exploits	Generic	Reconnaissance	Shellcode	worms	totalAttacks
1	24246	2677	2329	16353	44525	215481	13987	1511	174	321283

Conclusions

By following the steps, and selecting the most appropriate coding algorithms, the result was the creation of pieces of useful information for analysis and decision making. The main goal of carrying out this web attack analysis project is to be able to contribute to the field of information security administration. It facilitates knowledge and skills for future information security professionals with examples and techniques of gathering data. Using the RStudio and R Language, it was possible to extract information about network attacks in real time from four datasets. These network datasets were generated by the Cyber Range Lab of the Australian Center for Cyber Security (ACCS). The results obtained from the different types of attacks and the statistical data obtained can help experts to analyze cyber attacks. The objective of extracting data that can be used for analysis and decision making was achieved. The field of cybersecurity is complex, but with education and guidance security experts can learn the knowledge and skills to mine data in the information security field.

Future Work

Developing more research on methodologies, data analytics & data mining techniques to analyze web attacks is a great countermeasure for mitigation of cyber attacks. These skills required are in great demand. It is crucial for universities to do research in data mining and machine learning for optimal data extraction. They can also do capture the flag competitions and data mining on network traffic to identify new trends in cyber web attacks. Finally, researchers can develop new data analytics and mining programming libraries for different programming languages to obtain better data mining results with a more efficient use of the programming code.

Acknowledgements

I would like to thank to my advisor Dr. Alfredo Cruz for his guidance and advice. I also want to thank Carlos Vélez and Alexander López for their support.This material is based upon work supported by, or in part by the PUPR-NRC Scholarship. Grant Fellowship Award under contract/ award # NRC-HQ-7P-15-G-0006.

References

[1] Kaur, D. Kaur, P. (2015). Empirical Analysis of Web Attacks. International Conference on Information Security & Privacy (ICISP2015) Available:https://www.sciencedirect.com/science/article/pii/S1877050916000594.

[2] Berthier, R., Korman, D., Cukier, M., & Hiltunen, M, (2008). The Comparison of Network Attack Datasets: An Empirical Analysis. 2008 11th IEEE High Assurance Systems Engineering Symposium. Available:https://ieeexplore.ieee.org/abstract/document/4708862

[3]Yao, J., & Yao, Y. (2003). Web-based information retrieval support systems: building research tools for scientists in the new information age. 2012 Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003). 10.1109/WI.2003.1241270.