MALWARE DETECTION USING KRUSKAL-WALLIS STATISTICAL ANALYSIS AND TANIMOTO COEFFICIENT

NAFISATUN NAJA BINTI KAMARULZAMAN

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2017

BORANG PENGESAHAN STATUS TESIS

JUDUL: MALWARE DETECTION USING KRUSKAL-WALLIS STATISTICAL ANALYSIS AND TANIMOTO COEFFICIENT

SESI PENGAJIAN: 2016 / 2017

Saya, NAFISATUN NAJA BINTI KAMARULZAMAN

mengaku membenarkan tesis (PSM/Sarjana/Doktor Falsafah) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat Salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat Salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/)

_____ SULIT (Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)

_____ TERHAD (Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/ badan di mana penyelidikan dijalankan)

_____ TIDAK TERHAD

_____	_____
(TANDATANGAN PENULIS)	(TANDATANGAN PENYELIA)
Alamat tetap : PT41712, Jln Sentosa 2, Tmn Desa Sentosa, Teras Jernang. 43650 Bdr Baru Bangi, Selangor.	Dr. S.M Warusia Bin S.M.M Yassin
Tarikh: _____	Tarikh: _____

CATATAN: * Tesis dimaksudkan sebagai Laporan Akhir Projek Sarjana Muda (PSM)

** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

DECLARATION

I hereby declare that this project report entitled

MALWARE DETECTION USING KRUSKAL-WALLIS STATISTICAL ANALYSIS AND TANIMOTO COEFFICIENT

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT : _____ Date: _____
(NAFISATUN NAJA BINTI KAMARULZAMAN)

SUPERVISOR : _____ Date:_____
(DR. S.M WARUSIA BIN S.M.M YASSIN)

DEDICATION

To my beloved parents and those who have always by my side during my upside down.

ACKNOWLEDGEMENT

I wish to thank Dr. S.M. Warusia Bin S.M.M Yassin for all the beneficial comments and suggestions to review the paper and contributed to its technical content. I am also very grateful to authors of various articles on the internet who gives me become aware of the research currently ongoing in this field.

Last but not the least, I would like to thank all of my friends and family members who have always been there to give support and helped me to complete this research work.

ABSTRACT

Vulnerabilities have always been a worry for all software developers because of this, the protection of information systems against malicious activities and attacks in networks has been built, and one of them is intrusion detection systems (IDS). In this case, IDS is built to monitor a range of computer systems such as an information system, a network or a cloud computing for signs of intrusion. In order to observe and examine the data for anomalous and non-anomalous behaviours, Anomaly-based detection approach is used because of its ability to detect novel or "zero-day" attacks. Many anomaly detection techniques have been recommended in the literature to overcome this problem. One of them includes a Statistical-based detection that usually applying statistical analysis to examine and determine the behaviours of a subject such as packets or data. The current statistical based detection method have drawback in differentiate the anomalous behaviours more precisely. On the other hand, the lack of further analysis on anomalous behaviours results in high tendencies of wrongly examined malwares data. Thus, some ways are proposed to overcome the problem. First, distinguish the degree of packet behaviour more accurately using statistical base anomaly detection. Second, differentiate the anomalous and non-anomalous packets behaviour more accurately by exploring the Kruskal-Wallis and Tanimoto coefficient approach. Kruskal-Wallis test is used to find and produce accurate result on examining the packet behaviours. There are a few steps to be focused in this research which include data preparation, scoring method that focus on anomaly score, and analysing data that cover standard deviation, mean, Kruskal-Wallis and threshold based detection using Tanimoto Coefficient. This project contributes a better approach on detecting malicious attacks based on packet characteristics and also to propose a technique of

statistical analysis using Kruskal-Wallis to differentiate the anomalous and non-anomalous packet behaviour.

ABSTRAK

Kelemahan telah sentiasa menjadi kebimbangan untuk semua pemaju perisian kerana ini, perlindungan sistem maklumat terhadap aktiviti berniat jahat dan serangan dalam rangkaian telah dibina, dan salah satu daripadanya adalah sistem pengesanan pencerobohan (*Intrusion Detection System*). Dalam kes ini, sistem berkenaan dibina untuk memantau pelbagai sistem komputer seperti sistem maklumat, rangkaian atau pengkomputeran awan untuk mengesan tanda-tanda pencerobohan. Untuk memerhatikan dan memeriksa data untuk tingkah laku ganjil dan bukan ganjil, pendekatan pengesanan berasaskan *anomaly* digunakan kerana keupayaannya untuk mengesan novel atau serangan "*zero-day*". Banyak teknik pengesanan anomali telah disyorkan dalam kesusasteraan untuk mengatasi masalah ini. Salah satunya adalah termasuk pengesanan berdasarkan statistik yang biasanya menggunakan analisis statistik untuk memeriksa dan menentukan tingkah laku sesuatu subjek seperti paket atau data. Kaedah statistik pengesanan pada waktu ini mempunyai kelemahan dalam membezakan tingkah laku ganjil dengan lebih tepat. Sebaliknya, kekurangan analisis lanjut mengenai tingkah laku ganjil menyebabkan kecenderungan tinggi data malwares salah diperiksa. Oleh itu, beberapa cara yang dicadangkan untuk mengatasi masalah ini. Pertama, membezakan tahap tingkah laku paket lebih tepat menggunakan pengesanan asas anomali statistik. Kedua, membezakan tingkah laku paket ganjil dan bukan ganjil dengan lebih tepat dengan meneroka pendekatan pekali Kruskal-Wallis dan Tanimoto. Ujian Kruskal-Wallis digunakan untuk mencari dan menghasilkan keputusan tepat apabila meneliti tingkah laku paket. Terdapat beberapa langkah untuk memberi tumpuan dalam kajian ini termasuk penyediaan data, kaedah pemarkahan yang memberi tumpuan kepada skor anomali, dan menganalisis data yang meliputi sisihan piawai, min, Kruskal-Wallis dan pengesanan berdasarkan ambang menggunakan Tanimoto Pekali. Projek ini menyumbang pendekatan yang lebih baik dalam mengesan serangan berniat jahat berdasarkan ciri-ciri paket dan juga untuk mencadangkan satu teknik analisis

statistik menggunakan Kruskal-Wallis untuk membezakan kelakuan paket ganjil dan bukan ganjil.

TABLE OF CONTENTS

CHAPTER	SUBJECT	PAGE
	DECLARATION	**i**
	DEDICATION	**ii**
	ACKNOWLEDGEMENT	**iii**
	ABSTRACT	**iv**
	ABSTRAK	**v**
	TABLE OF CONTENTS	**vi**
	LIST OF TABLES	**ix**
	LIST OF FIGURES	**x**
	LIST OF ABBREVIATION	**xi**
CHAPTER I	**INTRODUCTION**	
	1.1 Introduction	1
	1.2 Motivation	2
	1.3 Problem Statement	3
	1.4 Project Question	4
	1.5 Project Objective	4
	1.6 Project Scope	5
	1.7 Project Contribution	5
	1.8 Thesis Organization	6
	1.9 Conclusion	8

CHAPTER II LITERATURE REVIEW

2.0 Intrusion Detection System 11
2.1 Scope of Detection 13
2.1.1 Host-Based IDS 13
2.1.2 Network-Based IDS 13
2.2 Analysis and Detection Mechanism 14
2.2.1 Signature Based Detection 14
2.2.2 Anomaly Based Detection 15
2.2.2.1 Statistical-Based Anomaly Detection System 16
2.3 Categories of Attack
2.3.1 Phishing 18
2.3.2 Spoofing 18
2.3.3 Denial-of-Service (DoS) 19
2.4 Similarity Analysis 19
2.5 Study and Analysis of Previous Work 20
2.6 Conclusion 25

CHAPTER III PROJECT METHODOLOGY

3.1 Introduction 36
3.2 Network Architecture 37
3.3 Data Source Collection 39
3.4 Generalised Framework 41
3.5 Software Specification 42
3.5.1 Microsoft Excel 42
3.5.2 MySQL 43
3.6 Project Milestone 43
3.7 Conclusion 44

CHAPTER IV FRAMEWORK DESIGN

4.1 Introduction 45

4.2 Multi-Layer Data Framework 46

4.2.1 Data Pre-processing 47

4.2.2 Scoring Phase 47

4.2.3 Kruskal-Wallis Test 48

4.2.4 Tanimoto Coefficient 49

4.3 Metric Measurement 50

4.3.1 Detection Rate 51

4.3.2 False Alarm Rate 51

4.4 Conclusion 52

CHAPTER VI IMPLEMENTATION

5.1 Introduction 53

5.2 Environment Setup 54

CHAPTER VII TESTING AND ANALYSIS

6.1 Introduction 68

6.2 Result and Analysis 68

6.3 Conclusion 70

CHAPTER VIII PROJECT CONCLUSION

7.1 Introduction 71

7.2 Project Summarize 71

7.3 Project Contribution 72

7.4 Project Limitation 72

7.5 Future Work 73

REFERENCES 74

LIST OF TABLES

Table 1.1: Summary of Problem Statement 4
Table 1.2: Summary of Problem Objectives 5
Table 2.1: Summary of Study and Analysis of Previous Works.....43
Table 3.1: Gantt Chart.....43

LIST OF FIGURES

Figure 2.1: Taxonomy of Intrusion Detection 10
Figure 3.1: Process of Methodologys 38
Figure 3.2: Network Architecture 39
Figure 3.3: Framework Flowchart 43
Figure 4.1: Multi-layer Data Framework 43
Figure 5.1: Sample of Training data in Excel 54
Figure 5.2: Sample of Testing data in Excel 55
Figure 5.3: Created schema 55
Figure 5.4: Created tables in MySQL 56
Figure 5.5: Import csv file into MySQL 56
Figure 5.6: Command to insert unique data into table training 57
Figure 5.7: Distinct data from training_distinct table in MySQL 58
Figure 5.8: Allocate '1' in matched data 58
Figure 5.9: Allocate '0' in unmatched data 58
Figure 5.10: Result after executing the query of allocating '0' and 1' 58
Figure 5.11: Command to calculate the value of *N* 58
Figure 5.12: Command to obtain the value of *R* 58
Figure 5.13: Calculation in normal profiling 58
Figure 5.14: Calculation to obtain normal score 58
Figure 5.15: Table of data in normal profile in MySQL 58
Figure 5.16: Table of data in normal profile table in MySQL 58
Figure 5.17: Command to allocate score into matched data 58
Figure 5.18: Table of data in testing data table in MySQL 58
Figure 5.19: Command to calculate total of normal score and anomalous score 58

Figure 5.20: Anomlous score and normal score in testing table in MySQL.....58
Figure 5.21: Distinct data from training_distinct table in MySQL.....58
Figure 5.22: Distinct data from training_distinct table in MySQL.....58
Figure 5.23: Distinct data from training_distinct table in MySQL.....58

LIST OF ABBREVIATION

IDS	Intrusion Detection System
NIDS	Network-based Intrusion Detection System
HIDS	Host-based Intrusion Detection System
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
DoS	Denial-of-Service
IP	Internet Protocol

CHAPTER I

INTRODUCTION

1.1 Introduction

The expanded dependence of ministry, combatant and business associations on Internet technologies to lead their regular field makes a stack of a brand-new test for the cyber protection. These days, the growth of internet and use of hardware and software has brought in a massive computerized change of data which encountered numerous issues like security and confidentiality of data. Vulnerabilities have always been a worry for all software developers because non-authorized or authorized users can misuse it. Thus, some tools are being intended and applied for a variability of exploitations in numerous kind of security violation. In this case, intrusion detection systems (IDS) is built to monitor a variety of computer systems such as an information system, cloud computing or a network for signs of intrusion.

Moreover, intrusions can be detected by using Intrusion detection systems (IDS) and defined as a try to make disruption on the security purposes for instance confidentiality, integrity, availability and non-repudiation. In order to observe and examine the data for anomalous and non-anomalous behaviours, Anomaly Detection approach is used.

Anomaly-based IDS does its job by monitoring the network traffic and analyse it with the accepted base station. The base station will recognise which bandwidth is used, what protocol and ports used. Then, it will alert the network administrator when traffic is detected. Moreover, it also can detect the anonymous attacks (Sobinsoniya, 2016). The patterns of anomaly detection are usually trained to create a baseline of normal network traffic. The detection can detect statistically significant deviations from normal. The main benefit of anomaly detection is its potential capability to detect novel or "zero-day" attacks. However, there are also the limitations of anomaly detection which are it may miss low-rate attacks and high rate of false alarms. Many techniques of anomaly detection have been recommended in the literature to overcome this issue. One of them includes a Statistical-based detection.

Statistical analysis methods provide measurements that can be compared across several methodologies. Moreover, this method can determine the behaviours of a subject such as packets or data. In my project, different approaches of the statistical-based intrusion detection are explored in order to perform detailed analysis thus produce accurate results. During the analysing process, Microsoft Excel will be used to generate graphs and MySQL for storing and manually produce the results of Kruskal-Wallis approach.

1.2 Motivation

Intrusions are described as any set of actions that can harm the integrity, availability or confidentiality of network resources. In the past few years, it showed

that the number of intrusions in networking has increased rapidly. Moreover, there are many new hacking tools and intrusive methods built for illegal activities. Therefore, anomaly detection is a detection technique used to capture and detects any abnormality from the normal behaviour as an anomalous behaviour. With high detection performance that is based on finding abnormal data by statistical measurement, we proposed an analysis on an intrusion detection system using the Kruskal-Wallis statistical test to detect outlier data and find out the results. The accuracy of results produced for continuous enhancement of intrusion detection capabilities and its numerous approaches is the motivation for this research.

1.3 Problem Statement

Intrusion detection techniques commonly contain two main IDS techniques which are signature-based detection and anomaly-based detection. Anomaly-based detection, on the other hand, uses statistical techniques to detect abnormality of behaviour to look at how impressive the statistical method in detecting intrusions. One of its primary strengths compared to the signature-based detection is if new attacks fall out of the normal traffic patterns, it can still be detected despite the fact for which signature does not exist. Unfortunately, anomaly detection engines are difficult to define the rules. It requires much time because each protocol that is being analysed must be defined, implemented and tested for accuracy. Thus it can be the disadvantage of statistical analysis which is likely to report an unacceptable number of alerts such as false alarms due to its failure to rapidly adapt to certain difference in a user's behaviour.

Besides that, statistical techniques can be major disadvantages because it is susceptible to be trained by attackers. Other than that, the parameters and metrics have difficult settings. Furthermore, when the researchers depend on the supposition that the information is produced from a specific distribution. This assumption usually unrealistic especially for high dimensional real data sets. It is difficult to

pick the best measurement as it is not a straightforward task. Moreover, it should include more knowledge of all well-known attacks that are considered to understand how they influence the normal network behaviour. This issue can lead to huge numbers of false positive since existing of statistical-based anomaly detections proposed are less efficient to examine the malware data accurately.

Table 1.1: Summary of Problem Statement

PS	Problem Statement
Ps1	Current statistical based detection method have drawback in differentiate the anomalous behaviours more precisely.
Ps2	Lack of further analysis on anomalous behaviours result in high tendencies of wrongly examined malwares data.

1.4 Problem Question

i. How the packet or data can be differentiated whether it comes from anomalous or non-anomalous behaviour?
ii. What techniques can be used to analyse the behaviour of packet or data?

1.5 Project Objective

The anomaly-based detection is based on defining the network behaviour. This method differentiate whether the node is normal or abnormal in view of its behaviour. Therefore, Statistical-based technique is used to determine a region denoting normal behaviour and state any view in the data that does not relate to this area. This technique uses statistical properties and statistical tests involve examining the process of the anomalous and non-anomalous behaviour of a packet. In statistical approach, mean, standard deviation or any other correlations are known as a moment. Thus, the event that drops outside the set interval above or below the

moment is assumed to be anomalous. The system is exposed to change by considering the aging data and making changes to the statistical rule database. By analysing web traffic and processing the information with statistical algorithmic program, Statistical-based Intrusion Detection (SBID) systems proficient of observing for anomalies in the fixed normal network traffic patterns. Plus, an anomaly score is given to all packets that indicate the level of abnormality for the particular event. If the anomaly score is higher than a certain threshold, the IDS will produce an alarm. The way to any SBID framework is its capability to differentiate and recognize normal from a typical anomalous network activity.

The limit amongst the normal and anomalous behaviour is regularly not exact. In addition, characterizing a normal region that includes each possible normal behaviour is exceptionally troublesome. Along these lines, an anomalous observation that lies near to the limit can really be normal and the other way around. Besides that, at the point when abnormalities are the consequence of malicious activitites, the malicious adversaries regularly adapt themselves to make the anomalous observations seem normal, in this manner making the task of examining normal behaviour more troublesome. Kruskal-Wallis approach is implemented in determining packet behaviours. In addition, a threshold called Tanimoto Coefficient is used to examine the anomalous and non-anomalous behaviour accurately.

Table 1.2: Summary of Problem Objectives

PO	Project Objective
Po1	To distinguish the degree of packet behaviour using statistical base anomaly detection and scoring method more accurately.
Po2	To differentiate the anomalous and non-anomalous packets behaviour more accurately by exploring the Kruskal-Wallis approach.

1.6 Project Scopes

1.6.1 Specific Tools

i. Microsoft Excel is used to calculate the data and create a graph to analyse the behaviour whether it is anomalous or non-anomalous.
ii. MySQL is used as database to store all the data.

1.6.2 Specific Data Used

i. UCLA CSD Dataset

1.6.3 Specific Method

i. Kruskal-Wallis Test
ii. Tanimoto Coefficient

1.7 Project Contribution

In this thesis, this project may be rewards to people who are in charge of stopping cyber-attack. This approach provide a detection on malicious attacks based on packet characteristics. Besides that, there are benefits in analysing data of malware more accurately using Statistical Analysis by examining the behaviour of data or packets whether it is anomalous or non-anomalous with a specific end goal to enhance the nature of results. Moreover, the statistical based anomaly detection technique does not require the foundation information about the target system's normal activity.

1.8 Thesis Organization

Chapter 1: Introduction

This chapter describe more about this project and lead to the next activities to be developed. In this chapter, we will discuss the problem statement, project questions, project objectives, project scope, project organization and the conclusion.

Chapter 2: Literature Review

This chapter describes about the related work form the resources that had been gained. The resources include journals and books created in range between 2012 until 2017 which have been carried out by other researcher. The chapter will discuss the critical review of existing problems and justification.

Chapter 3: Project Methodology

This chapter discusses each stages of the selected methodology which are milestones and Gantt chart. Milestones will clarify the activities designs preceding the finish of the undertaking and it will be connected from what we have learnt from project management. Besides, it will define every stages of the activities that will be carried out in the project and the framework of the project methodology. Other than that, network architecture and data source collection will also be explained briefly.

Chapter 4: Design

Chapter 4 introduces the multi-layered framework in selected methods conducted. The proposed methods for analysing the anomaly of packet behaviour can be divided into groups. This framework provides full requirements needed for features selection and extraction in pre-processing process, scoring method, developing and evaluation to the final results of the proposed approaches.

Chapter 5: Implementation

This chapter discuss the activity tangled in the implementation stage and the expected output. Besides that, this chapter will include environment setup, the parameters, variables, and the assumptions used in this project.

Chapter 6: Testing and Analysis

This chapter is about the activity in the implementation phase in the analysis of the project and will include graphical results from the critical analysis using the collected data.

Chapter 7: Project Conclusion

In this chapter, it will summarize the project, report all implementation and testing phase. Finally, it will conclude the important results that will be achieved in this project. This chapter also states the limitation and the strong point of this project.

1.9 Conclusion

In summary, this chapter includes the introduction of project in details, the problem statements, question might be arise from the project, the objectives, project scope, project contribution which is the benefits in the project and the thesis organization. On the next chapter, details about Literature Review will be discussed.

CHAPTER II

LITERATURE REVIEW

This chapter presents some of the existing intrusion detection techniques, which are focused on some metric measurement such as false alarm rate, false negative rate, accuracy and detection rate. It also covers the detailed of background information on statistical-based intrusion detection system and describes the statistical analysis techniques used in intrusion detection field.