FRIEDMAN STATISTICAL ANALYSIS AND COSINE THRESHOLD METHOD FOR TCP BASED MALWARE DETECTION

NUR AMEERA NATASHA BINTI MOHAMMAD

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS TESIS

JUDUL: Friedman Statistical Analysis And Cosine Threshold Method For DDoS Detection

SESI PENGAJIAN: 2016 / 2017

Saya __________NUR AMEERA NATASHA BINTI MOHAMMAD__________

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: NO. 20, Jalan Perdana 24	DR. S.M WARUSIA BIN S.M.M
Taman Bukit Perdana, 83000 Batu Pahat,	YASSIN
Johor.	
Tarikh: _________________	Tarikh: __________________

CATATAN: * Tesis dimaksudkan sebagai Laporan Akhir Projek Sarjana Muda (PSM)
** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

FRIEDMAN STATISTICAL ANALYSIS AND COSINE THRESHOLD METHOD FOR TCP BASED MALWARE DETECTION

NUR AMEERA NATASHA BINTI MOHAMMAD

This report is submitted in partial fulfilment of the requirement for the

Bachelor of Computer Science (Computer Security)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2017

DECLARATION

I hereby declare that this project report entitled

FRIEDMAN STATISTICAL ANALYSIS AND COSINE THRESHOLD METHOD FOR TCP BASED MALWARE DETECTION

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

STUDENT : ______________________________________ Date: ________
(NUR AMEERA NATASHA BINTI MOHAMMAD)

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

DEDICATION

To my beloved parents thank you
for always supporting me
and being there when I am feeling down

To my loyal friend thank you
for sharing your knowledge and helping me
in completing this project

To my supervisor thank you
for encouraging, motivating and believing
in me

ACKNOWLEDGEMENTS

Alhamdulillah. Thanks to Allah SWT, who with His willing give me the opportunity to complete this Final Year Project which titled Friedman Statistical Analysis And Cosine Threshold Method For TCP Based Malware Deteection. Firstly, I would like to express how thankful I am to Dr. S.M Warusia Bin S.M.M Yassin as my supervisor who had guided a lot of task during this semester in completing this Final Year Project. Deepest thanks and appreciation to my parents, family and my supportive friends and others for their assistance, encouragement, constructive suggestion and full support for the report completion, from the beginning till the end. Last but not least, my thanks to the members of Faculty of Information Communication and Technology UTeM, for commitment and cooperation during my Final Year project.

ABSTRACT

Intrusion Detection System (IDS) is a network security technology which inspects all inbound and outbound on computer network traffic and design for detecting suspicious patterns that attempts to perform security policy violation. IDS approach the goal to detect threats in various ways. Most of the IDS implement signature based which means that they operate almost as same as a virus scanner, by search for a known identity or any signature for each specific intrusion event. The Friedman test is used to test for differences between groups when the dependent variable being measured is ordinal and for continuous data that has violated the assumptions necessary to run the one-way Anova with repeated measures. As what has been described above, Friedman test has meaning of the non-parametric alternative to the one-way Anova with repeated measures. The problem statements for this project are that the lack of approach to examine the degree of behaviour of each packet more accurate and to ensure whether the unforeseen packets behaviours contain anomalous and non-anomalous activity is hard to differentiate. For the objective, it is to distinguish the degree of packet behaviour using Friedman statistical base analysis for detecting behaviour more correctly and also to differentiate the anomalous and non-anomalous packets behaviour more accurately using scoring method. For methodology, this project represents the method in Friedman statistical analysis for detecting packet behaviour including the specific steps that will undertake to produce accurate outputs. There are few steps to be focus in this project which include data preparation, data scoring where include anomaly score and normal score and last step is analyse data. Analyse data divided into standard deviation, mean method and normal. This project will contribute on proposing the new technique to identify intrusion by classifying activity as either anomalous or normal. Other than that, this project also contribute in distinguish the degree of packet behaviour using Friedman statistical base analysis for detecting behaviour.

ABSTRAK

Sistem Pendeteksi Intrusi (SPI) adalah teknologi keselamatan rangkaian yang memeriksa semua trafik rangkaian komputer masuk dan keluar dan direka untuk mengesan corak trafik yang mencurigakan dan cuba untuk melaksanakan dasar keselamatan pelanggaran. SPI meliputi matlamat untuk mengesan ancaman dalam pelbagai cara. Sebahagian besar SPI akan melaksanakan operasi signatur yang bermaksud bahawa mereka beroperasi hampir sama seperti pengimbas virus, dengan mencari identiti yang diketahui atau sebarang petunjuk untuk setiap acara pencerobohan tertentu. Ujian Friedman digunakan untuk menguji perbezaan antara kumpulan apabila sesetengah ciri yang diukur adalah ordinal dan data berterusan telah melanggar andaian yang perlu untuk menjalankan satu hala Anova dengan langkah-langkah berulang. Seperti yang telah dinyatakan di atas, ujian Friedman mempunyai pengertian alternatif bukan parametrik kepada sehala Anova dengan langkah-langkah berulang. Agak sukar untuk penyata masalah bagi projek ini untuk mengkaji tahap tingkah laku setiap paket yang lebih tepat dan untuk memastikan sama ada paket tersebut bertindak di luar jangkaan tingkah laku atau tidak dan sama ada mengandungi aktiviti ganjil dan bukan ganjil. Bagi objektif, ia adalah untuk membezakan tahap tingkah laku paket menggunakan analisis asas statistik Friedman untuk mengesan tingkah laku yang lebih betul dan juga untuk membezakan paket ganjil dan bukan ganjil tingkah laku dengan lebih tepat menggunakan kaedah pemarkahan. Metodologi, projek ini merupakan kaedah dalam analisis statistik Friedman untuk mengesan tingkah laku paket termasuk langkah-langkah tertentu untuk menghasilkan hasil kerja yang tepat. Terdapat beberapa langkah untuk menjadi tumpuan dalam projek ini termasuk penyediaan data, jaringan data di mana termasuk anomali skor atau skor normal dan langkah terakhir adalah menganalisis data. Menganalisis data dibahagikan kepada sisihan standard, kaedah min dan normal. Projek ini akan menyumbang dalam mencadangkan teknik baru untuk mengenalpasti pencerobohan dengan mengklasifikasikan aktiviti sama ada ganjil atau normal. Selain daripada itu, projek ini juga menyumbang dalam membezakan tahap tingkah laku paket menggunakan analisis asas statistik Friedman untuk mengesan tingkah laku.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Introduction

As data frameworks turn out to be progressively minded boggling, vulnerabilities and bugs of data framework are regularly abused by malignant clients to interrupt into data frameworks and trade off security, for example, accessibility, respectability and classification of data frameworks.

In this manner, the likelihood of interruptions into data frameworks dependably exists. For this situation, Intrusion Detection System is conveyed to distinguish suspicious exercises. The Intrusion Detection System (IDS) is an innovation of system security which initially worked for distinctive misuses against target application or PC. The goal of this system is to identify dangers and accordingly is set out network circle on the system framework, implying that it is not in the unaffected ongoing communication between the sender and recipient of data.

Intrusion Detection System (IDS) is kind of formula of observing the events happening in a personal computer or network circle or organize and breaking down indications of interruption. With a specific end goal to watch and look at the information for anomalous and non-anomalous practices, Anomaly Detection approach is utilized.

Inconsistency based IDS build up ordinary utilization design as gauge and derives anything that generally digresses from the typical examples as a conceivable interruption. The essential preferred standpoint of abnormality location is its ability to discover novel assault which accordingly it addresses the greatest impediment of abuse discovery. Be that as it may, because of the suppositions basic irregularity location components, their false alert rates are by and large high. Numerous inconsistency recognition strategies have been proposed in the writing to defeat this issue. These methods incorporate Statistical-based analysis.

In another word, this examination includes inspecting procedure of odd and non-peculiar conduct of parcel. Moreover, the way toward deciding these practices turns out to be all the more difficult these days. In my venture, distinctive methodologies of the factual based interruption discovery are investigated so as to perform examinations along these lines to create exact outcomes. Amid the investigating procedure, Microsoft Excel will use for creates charts and MySQL for putting away and physically deliver the outcomes for distinctive methodologies. The distinctive methodology is an algorithm of Friedman statistical analysis. Friedman test is a non-parametric measurable test produced Milton Friedman. Like the parametric rehashed measures ANOVA, it is has been programmed to find contrasts in treatments over various trial endeavors. However, statistical analysis can cause sampling error, correlation and causation error, construct validity and simplified solutions.

1.2 Problem Statement

Intrusion detection techniques usually divided into two classes which are signature based detection and anomaly based detection. In other perspective, Anomaly-based detection used statistical strategies to examine the behaviour of the target and afterward searches for deviations or inconsistencies. One of its essential qualities over signature-based detection is it do not give attention of having priori information of particular vulnerabilities or attacks and capacity to identify unknown attacks. The reference for this algorithm is very limited as it is can consider as a new technique to analyse DDoS packets. Thus it can be the disadvantage of statistical analysis which is likely to report an unacceptable number of false alarms because of the lack of approach to examine the degree of behaviour of each packet more accurate.

Besides that, statistical techniques can be major disadvantages when the researchers depend on statement that produced in a form of report from a certain distribution. This theory do not usually proved true for high dimensional especially on real data sets. The statistical theory can be randomly justified but there are still some of the statements that can be used to detect anomalies and it is not straightforward to choose the best result. On the other hand, it should include a deep study 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 it is difficult to differentiate the unforeseen packets behaviours and to identify the packets that contain anomalous and non-anomalous activity.

Table 1.1: Summary of Problem Statement

PS	Problem Statement
Ps1	The degree of packet behaviour is hard to be determined.
Ps2	Difficult to differentiate the unseen behaviours of a packet more accurately.

1.3 Problem Question

i. What approach can be use to examine the degree of behaviour of each packet more accurately?

ii. How the packet or data can be differentiated whether it comes from anomalous or non-anomalous behaviour?

1.4 Project Objective

Distinguish the degree of packet behaviour using Friedman statistical base analysis for detecting behaviour more correctly is one of the objective of this project. Furthermore, boundary between normal and anomalous behaviour is frequently not correct. Consequently the result of an anomalous perception can be normal or vice versa since the boundry was too close. Because of that, when we want to differentiate between two groups of dependent variable so Friedman approach can be used and for continuous data that has violated the assumptions necessary to run the one-way Anova with repeated measures. As what has been described above, Friedman test has meaning of the non-parametric ways or step by step to the one-way Anova with repeated measures.

The detection of anomaly based is depends on defining network behaviour. The network behaviour in linkage circle is predefined behaviour, and then it can be accepted or triggers the in the anomaly detection event. Anomalies are behaviour in data does not conform to a well define motion of usual behaviour. For that reason, to describe a pattern that represents usual behaviour and proclaim with any possible result of comment in the data that does not belong to this section can examine using Statistical-based technique. This techniques use statistical properties and statistical tests involves examining process of anomalous and non-anomalous behaviour of a packet. In statistical approach, terms such as mean and standard deviation or any other correlations are known as a moment. Thus possible hypothesis may be the output outside the set interval above or below the moment is said to be anomalous. The system is imperilled to change by making changes to the statistical rule data base and considering the aging data.

Table 1.2: Summary of Project Objectives

PO	Project Objective
Po1	To distinguish the degree of packet behaviour more accurately using Friedman statistical base analysis.
Po2	To differentiate the anomalous and non-anomalous packets behaviour more accurately using scoring based approach.

1.5 Project Scopes

1.5.1 Specific Tools

i. Microsoft Excel is utilized to ascertain the information and frame a graph to dissect the behaviour whether it is anomalous or non-anomalous.

ii. MySQL is used for stores the data.

1.5.2 Specific Data Used

i. Laboratory for Advanced System Research Dataset

- This will be focus on all header of packet but give more attentions on the time frame of DDoS attacks.

1.5.3 Specific Method

i. Friedman Statistical analysis

1.6 Project Contribution

In this theory, this venture might be preferences to individuals who in control for any digital assaults and will spare their time when they have to keep any assault as the assaults have been separate some time recently. Other than that, there have benefits in investigating the information of malware with more precisely by using Statistical Analysis by examining at the information or data to enhance the nature of results. Moreover, the statistical based anomaly detection technique does not require the foundation or background information about the objective target's system movement.

1.7 Thesis Organization

Chapter 1: Introduction
This part depicts more about this venture and prompts the following exercises to be created. In this section, we will talk about the problem statement, project question, project objective, project scope, project organization and the conclusion.

Chapter 2: Literature Review
Main section discusses the work involved from the resources that had been searched. The resources include journal, books, and magazine between 2012 until 2016 which have been done by other scientist. The section will talk about the basic audit of current issue and justification.

Chapter 3: Project Methodology
This part will examine each phases of the chosen methodology which are focus more on network architecture, datasets, milestones and Gantt chart. Milestones will clarify the activities arranges preceding the finish of the project and it will be apply from what we have learnt from project management. Also, it will explain in each phases of the exercises that will be done in the project and the system of the project methodology.

Chapter 4: Design

This section 4 is about system framework engineering which will discuss about system assumptions and as appropriate, conceivable situations, implementation approach and the metric estimation utilized as a part of project.

Chapter 5: Implementation

This part discusses about the movement required in the implementation phase and the normal yield. Other than that, this section includes environment setup, the parameters, factors, and the assumptions used in this project.

Chapter 6: Testing and Analysis

This section is about the movement in the implementation phase of the investigation in the project and will include graphical outcomes from the critical analysis using the collected data.

Chapter 7: Project Conclusion

In this part, it will abridge the project, report all execution and testing stage. Lastly, it will close the critical outcomes that will produce in this project. This part states the disadvantages or weakness and quality of this project.

1.8 Conclusion

In outline, this paperwork includes the introduction of project in details, the problem statements, question may be emerge from the project, the objectives, project scope, project contribution which is the benefits in the project and the thesis organization. On the following part, we will discuss in insight about Literature Review.

CHAPTER II

LITERATURE REVIEW

This chapter presents some of the existing intrusion detection techniques, which are focused on some metric measurement such as accuracy and detection rate. It is also have some explanations about categories of IDS and their functions. 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. In this chapter also will be discuss about some types of attacks which will be focus more on Distributed Denial of Service (DDoS) attack.

2.1 Introduction of Intrusion Detection System

According to J. P. Anderson, 1980 Intrusion Detection System is a struggle that cautious and unapproved efforts to get information, wrong use of information, or reduce a policy untrustworthy or impracticable. For example, Denial of Service (DoS) attack endeavours to absorb resources links, which during processing it had to be used in order to operate well. Another example are some of the famous world threat that exploit other crowds through the targeted network circle and taking privileged access to a host by taking benefits of known vulnerabilities (Bhuyan et al., 2014b).

Intrusion Detection System is a programming language that have detached to break down the concrete wall and network components in scope of confidentiality, integrity and availability stated by R. Heady, G. Luger, A. Maccabe and M. Servilla, 1990. This attack can be performing to enter the box whether from inside or outside and take over control of the security mechanism. They also stated that to shelter structure of the network systems, Intrusion Detection System (IDS) must provide a very well recognized mechanism, which collect and examine information from many part of link circle to recognize potential security gaps. According to R. Heady, G. Luger, A. Maccabe and M. Servilla also intrusion detection have many function which are observing and investigating element involve any possible activity perform by the user, measure integrity of log file, identify typical attacks patterns, study people policy violations of unformal activities. IDS assess the security by scanned possible vulnerabilities of a host or a network. It also can be used on intrusion activities assumption that may have different pattern from normal drawing (Bhuyan et al., 2014b).

According to Anderson, 1980, there are different classes of attack. It is divided into types which are external and internal. External intruders are unallowed users of the machineries spasm while internal intruders can access the system with major permission but do not have any rights for the main or basic or upper user mode. Moreover, internal intruders divided into two types which are masquerade internal intruders and clandestine internal intruders. A concealment internal intruder logs in as other users with valid access to targeted of sensitive data while a clandestine internal intruder is the most dangerous person as they has tendency to run off audit control for themselves (Bhuyan et al., 2014b).

According to H. G. Kayacik, A. N. Zincir – Heywood and M. I. Heywood, 2005 and also A. A. Ghorbani, W. Lu, and M. Tavallaee there are countless classes of Intrusion Detection Systems or attacks. Below are the summary of the attacks:

Table 2.1: Summary of the attacks

No.	Attack Name	Characteristics	Example
1	Virus	Have capability to do self-replicate which can make the program harm without owner knows and also have capability to enhance the infection graph of cloud file system if the system is access by another computer.	Trivial.88.D, Polyboot, Tuareg.
2	Worm	Have tendency to do self-replicating and from network service it can transmit on computer system without give alert to user and highly damage link by overshadowing network circle.	SQL Slammer, Mydoom, CodeRed Nimda.
3	Trojan Horse	Wicked unlike worm that can replicate itself but it has secret code that can be a reason of serious security problem because of backdoor that they created which functions to allow anything affected the system easily. We can conclude somebody take over on the computer system.	Example- Mail Bomb, phishing attack.
4	Denial – of – Service (DOS)	One of thousand ways to cut access to the system or network properties and can cause non-perfect functionality certain or targeted network circle, such as email.	Buffer overflow, ping of death (PoD), TCP