IDENTIFICATION OF ATTACK/MISUSE WITH EMAIL HEADER FORENSIC

AHMAD FAHRULRAZIE BIN MOHAMAD

A dissertation submitted in partial fulfillment of the
Requirements for the award of the degree of
Master of Science (Information Security)

Faculty of Computing
Universiti Teknologi Malaysia

DECEMBER 2014

This dissertation is dedicated to my family for their endless support and encouragement.

ACKNOWLEDGEMENT

Firstly, I would like to thank my parents for their loving and support that help during my study in UTM. Secondly, I would like to thank my supervisor Dr. Shukor Bin Abd Razak for her constant support and guidance during this project. He inspired me hugely to work in this project. His willingness to motivate and guide me contributed tremendously to our project. I have learned a lot from him and I am fortunate to have him as my mentor and supervisor.

Also, I would like to thank the authority of Universiti Teknologi Malaysia (UTM) for providing me with a good environment and facilities to complete this project.

ABSTRACT

Email becomes important communication nowadays; it was used in government sector, education sector, business sector and others. Because of its popularity it attracts offenders to commit crime in email communication. This study focuses on one type of email crime; spam email. Three popular webmail was choose namely as Hotmail, Gmail and Yahoo mail. Spam email was sent to each webmail to see the accuracy of each webmail in detecting the spam email. The results show Hotmail, Gmail and Yahoo mail are lack of accuracy in detecting those spam email. All spam emails was collected and information contained in the email header was analyze. Previous studies believe mismatch or forging information in the email header may indicate the behavior of spam emails. New email header forgery detection mechanism was developed to check mismatch or forging information in the email header. This study focus on the information contained in the Message-ID, Reply-To, From and Received field. Any mismatch or forging information in this field may indicate the behavior of spam emails. The mechanism will classify those emails that have mismatch or forging information in that particular features as spam emails instead classify as legitimate email since Hotmail, Gmail and Yahoo mail classified those spam email as legitimate email.

ABSTRAK

E-mel merupakan komunikasi penting pada masa kini, ianya digunakan dalam sektor kerajaan, sektor pendidikan, sektor perniagaan dan lain-lain. Oleh yang demikian ianya telah menarik pesalah untuk melakukan pelbagai jenayah dalam komunikasi e-mel. Kajian ini memberi tumpuan kepada salah satu jenayah dalam komunikasi e-mel iaitu e-mel spam. Tiga webmail popular telah dipilih dalam kajian ini iaitu Hotmail, Gmail dan Yahoo mail. E-mel spam telah dihantar ke setiap webmail untuk melihat ketepatan setiap webmail dalam mengesan e-mel spam. Kesemua e-mel spam tersebut dikumpulkan dan maklumat yang terkandung di dalam pengepala e-mel spam dianalisis. Kajian terdahulu percaya ketidaksepadanan dan pemalsuan maklumat di dalam pengepala e-mel akan menunjukkan tanda-tanda e-mel spam. Pengepala e-mel mekanisme pengesanan pemalsuan baru telah dibangunkan untuk memeriksa kesepadanan dan pemalsuan maklumat di dalam pengepala e-mel. Kajian ini akan memberi tumpuan kepada maklumat yang terkandung dalam bidang mesej-id, reply-to, from dan received. Sebarang ketidaksepadanan dan pemalsuan maklumat dalam bidang ini akan membawa kepada tingkah-laku e-mel spam. Pengepala e-mel mekanisme pengesanan pemalsuan akan memeriksa ketidaksepadanan dan pemalsuan maklumat di dalam pengepala e-mel dan akan mengklasifikasikan e-mel tersebut sebagai e-mel spam memandangkan Hotmail, Gmail dan Yahoo mail mengklasifikasikan e-mel spam tersebut sebagai e-mel yang sah.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Email message become important communication nowadays used from personal, business, education and others. Because of its popularity, it attracts the offenders to commit crime in email communication. Most of the people and organization believe email communication have many pitfalls since 52% of email users didn't believe it because of spam email (Boykin and Roychowdhury, 2004). Spam email is one of the crimes in email communication. It may cause network traffic problem and economical loss. High volumes of spam email may slow down the performance of mail server and delay the transportation of legitimate email since it already overload the mail server. Furthermore, malware programs like Trojan horses, viruses and phishing email may be placed inside the spam email.

Email message consist of email header and email body. Several problems have been highlight when identifying spam email based on the email body. Therefore, this study focuses on the information contained in the email header to identify the behavior of spam email. Several rules that lead to spamming activity are highlighted in this study.

Spammer continuous devise a way to trick the spam filtering system by changing or forging the information contained in the email header. Previous research related to the field believes that changing information in the email header may lead to the spamming behavior.

This study focus on three popular webmail namely as Hotmail, Gmail and Yahoo mail. Each webmail have its own spam filtering system to combat spam email. This study show that Hotmail, Gmail and Yahoo mail are lack of accuracy in detecting spam email due to incorrectly identified of spam email as legitimate email in each platform. Those incorrectly identified of spam email are collected from Hotmail, Gmail and Yahoo mail to analyze the information contained in the email header. Furthermore, Hotmail, Gmail and Yahoo mail have their own email header formats that differ between each other. Spammers may forge the information contained in the email header to deceive the spam filtering system in each webmail. This study will commonalize the webmail header and analyze the behavior of spam email itself.

Most of the researcher has no idea what type of spam filtering technique have been used by large email service provider like Hotmail, Gmail and Yahoo mail in detecting spam email. Experiments have been conducted in the previous work to evaluate the performance of several machine learning-based classifiers in detecting spam email. Potential features contained in the email header already highlighted by the researcher. This study will analyze the information contained in the spam email header that incorrectly identified in Hotmail, Gmail and Yahoo mail. Potential features contained in the email header are highlighted in this study. The result of the analysis may produce new email header forgery detection mechanism that may correctly identified those spam email. Problem background, problem statement, project aim, project objectives, project scopes, motivation and significance of the project and conclusion will be discussed on the next section.

1.2 Problem Background

Spam email become a serious threat to the organizations and email consumers. Apart from terrible financial lost it can also affect employees productivity. Most of the employees spend 10 minutes a day on average sorting through unsolicited messages (S.Hinde, 2002). The time spent by the employee will cause significant loss to the organization.

United States (US) Secret Service already conducted a study that estimates $5 Billion losses associated with email related crimes in the 1990's with victims' worldwide (Smith, 1999). A study conducted by Messaging Anti-Abuse Working Group (MAAWG) indicates that almost 89% to 92% of all sent emails are spam emails. It's already more than twenty years and we believe it become more damageable and destroyable from day to day.

Most of the spammer continually devises new way to deceive the spam filtering system. One of the technique uses by the spammers is to forge the information contained in the email header. We believe these techniques are carried out successfully by the spammers since several spam email are incorrectly identified in Hotmail, Gmail and Yahoo mail. Spam emails are classified as legitimate email in each platform. Most of the email consumers are lack of knowledge in identifying spam email. Some of them will reply the email message that leads the users to scam and phishing activities. Several spam emails may contain malware programs that harm the computer system of email consumers.

1.3 Problem Statement

There is no specific mechanism or technique that can produce low or no false positive and false negative result in detecting spam email since spammer continuously revise a way to trick the spam filtering system. Most of the spammer would like to change or forge the information contained in the email header to deceive the spam filtering system. These techniques are carried out successfully since large email service provider like Hotmail, Gmail and Yahoo mail also faced the same issues. This study will analyze the information contained in the spam email header and identify potential features in the header that may lead to the spamming behavior.

Furthermore, Hotmail, Gmail and Yahoo mail have it own email header format. The feature consists of mandatory features and optional features. Spammer may use this opportunity to deceive the spam filtering system in each webmail by changing or forging the information contained in the email header. The features contained in the email header will be discussed in detail on the next chapter. Several rules are implemented inside the spam email header that led to the spamming activities. These rules will check the behavior of information contained in the spam email header. For instance, all researchers in this field believe mismatch information of domain name in the Message-ID and From field may lead to spamming behavior. This rule will be placed inside the new email header forgery detection mechanism in detecting spam email since Hotmail, Gmail and Yahoo mail incorrectly indentify spam email that contained mismatch information of domain name between Message-ID and From field as legitimate email.

1.4 Project Aim

This study only highlight the issues in Hotmail, Gmail and Yahoo mail in detecting spam email only not responding on it. Spam email was sent to Hotmail, Gmail and Yahoo mail. Each webmail are lack of accuracy in detecting those spam emails. Most of the spam emails are incorrectly identified as legitimate email in Hotmail, Gmail and Yahoo mail. All incorrectly identified spam email in Hotmail, Gmail and Yahoo mail was collected and the information contained in the email header was analyzed. Potential features contained in the email header have been identified in this study. The features consist of Message-ID, Reply-To, From and Received field. In general, mismatch information or invalid domain name contained in the features may lead to spamming behavior since spammer may change or forge that information. Mismatch or forging information in the email header may be the rules that will be placed in the new email header forgery detection mechanism in detecting current spam email.

Furthermore, each webmail have different email header format. Spammers will use this opportunity to forge any features that can trick the spam filtering system. Therefore, this study will commonalize Hotmail, Gmail and Yahoo mail email header as well. The results of commonalize will be placed inside the new email header forgery detection mechanism.

New email header forgery detection mechanism will be developed at the end of the study. All the result obtains from the analysis will be placed inside the mechanism. The performance of the mechanism will be evaluated at the end of the study. The results will be comparing to Hotmail, Gmail and Yahoo mail to see the efficiency in detecting spam email compare to Hotmail, Gmail and Yahoo mail. We believe this mechanism will have high accuracy in detecting spam email since it contained the analysis information of spam email header that indicate the behavior of spam email.

1.5 Project Objectives

The objectives of this project are:

1. To commonalize Hotmail, Gmail and Yahoo mail email header.
2. To propose new email header forgery detection mechanism.
3. To evaluate the performance of new email header forgery detection mechanism.

1.6 Project Scopes

The scopes of his project are:

1. This project focuses on one Crime Related to Email Header (CREH) namely as spam emails.
2. This project focuses on the features in Hotmail, Gmail and Yahoo mail email header.
3. This project focuses on the incorrectly identified spam email in Hotmail, Gmail and Yahoo mail.
4. This project will evaluate the performance of new email header forgery detection mechanism.

1.7 Motivation and Significant of Project

Spam emails become a serious threat in email communication. Most of the email consumers will face this problem. Apart from terrible financial lost, spam emails can cause malicious actions in email communication. For instance, some spammers insert viruses, Trojan horse and phishing attacks in spam emails that can affect the credibility of email consumers. The survey show the problems cause by spam emails increase security concern and damageable from time to time.

Hotmail, Gmail and Yahoo mail are the popular webmail use by most organization and email consumers today. Yet, these platforms are lack of accuracy in detecting spam emails. Most of the spam emails are identified as legitimate emails. As mentioned above, spammer can carry out malicious actions in email communication. The attack may also affect the confidentiality of important data or information resides in the organization or users workstations. There are no specified techniques or mechanisms that produce low or no false positive and false negative results. Many studies are conducted to find efficient techniques that can counterpart all spam emails effectively. We believe it is impossible to develop the mechanism that produces no false positive results. Hopefully, the development of the new mechanism in this study may produce low false positive and false negative results in detecting spam emails. At least current spam email can be detect by this mechanism since the spammer may continually devise a ways to trick the spam filtering system by forging another features contained in the email header.

1.8 Conclusion

This chapter already discussed the overview of the studies that cover problem background, problem statement, project aim, project objectives, project scopes and motivation and significant of the project. The next chapter which is literature review will tell the readers about the main research of the studies. Several questions will be asked when writing literature review like what other researcher had done, what issues already discussed and what is the research gap that already identified. Real steps or methods taken to fulfill the studies will be discussed in chapter 3. Last but not least, initial results of the project will be discussed in chapter 4.

The information contained in each chapter may be helpful for research activities to come out with new email header forgery detection mechanisms that produce low or no false positive and false negative results in detecting spam email. The mechanism can act as supplement mechanism to analyze and identified any spam emails that incorrectly identified in Hotmail, Gmail and Yahoo mail system.

REFERENCES

About.com, 2010. *Definition of Email Body*. Retrieved 13 April 2012, from http://email.about.com/library/glossary/bldef_body.htm

Biggio B., Fumera G., Pillai I. and Roli F. 2011. *A survey and experimental evaluation of image spam filtering techniques*, Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi, Cagliari 09123, Italy.

Boykin O., and Roychowdhury V. 2004. *Personal Email Networks: An Effective Anti-Spam Tool.* Department of Electrical Engineering, University of California, Los Angeles, CA 90095.

MAAWG. Email Metrics Program: The Network Operators' Perspective. Report #10 {third and fourth quarter 2008, Messaging Anti-Abuse Working Group, S. Francisco CA, USA, March 2009.

Mokhtari M., Saraee M. and Haghshenas A. 2008. *A Novel Method in Scam Detection and Prevention using Data Mining Approaches*, Department of Electrical and Computer Engineering Isfahan University of Technology, Isfahan, Iran Mokhtari@ec.iut.ac.ir, Saraee@cc.iut.ac.ir, Department of Computer Engineering Iran University of Science & Technology, Tehran, Iran Haghshenas@comp.iust.ac.ir

Shih D., Chiang H. and Yen C. 2005. *Classification methods in the detection of new malicious emails*. Department of Information Management, National Yunlin University of Science and Technology, 123, Section 3, University Road, Touliu, Yunlin, Taiwan, ROC. Department of DSC & MIS Miami University, Oxford, OH 45056, USA.

TechieSouls Team, 2009. *Definition of email*. Retrieved 17 March 2012, from http://www.techiesouls.com/2009/08/10/definition-of-email/.

Al-Jarrah, O., Khater, I., & Al-Duwairi, B. (2012, January). Identifying Potentially Useful Email Header Features for Email Spam Filtering. In *ICDS 2012, The Sixth International Conference on Digital Society* (pp. 140-145).

Wang, C. C. (2004). Sender and receiver addresses as cues for anti-spam filtering. *Journal of Research and Practice in Information Technology*, *36*(1), 3-7.

Hu, Y., Guo, C., Ngai, E. W. T., Liu, M., & Chen, S. (2010). A scalable intelligent non-content-based spam-filtering framework. *Expert Systems with Applications*, *37*(12), 8557-8565.

Qaroush, A., Khater, I. M., & Washaha, M. (2012, September). Identifying spam e-mail based-on statistical header features and sender behavior. In*Proceedings of the CUBE International Information Technology Conference* (pp. 771-778). ACM.

Treviño, A., & Ekstrom, J. J. (2007). Spam Filtering Through Header Relay Detection. *Brigham Young University*.

Anh, N. T., Anh, T. Q., & Thang, N. X. (2010, October). Spam Filter Based on Dynamic Sender Policy Framework. In *Knowledge and Systems Engineering (KSE), 2010 Second International Conference on* (pp. 224-228). IEEE.

Tak, G. K., & Tapaswi, S. (2010). Query Based approach towards spam attacks using artificial neural network. *International Journal of Artificial Intelligence & Applications*, *1*(4).

Sanchez, F., Duan, Z., & Dong, Y. (2010, July). Understanding forgery properties of spam delivery paths. In *Proceedings of 7th Annual Collaboration*, *Electronic Messaging*, *Anti-Abuse and Spam Conference (CEAS)*.