A PHISHING E-MAIL DETECTION APPROACH USING MACHINE LEARNING TECHNIQUES

by

KENNETH FON MBAH

B.Sc, University of Dschang, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Computer Science

In the Graduate Academic Unit of Faculty of Computer Science, UNB

Supervisor:	Ali Ghorbani, PhD, Computer Science
Examining Board:	Huajie Zhang, PhD, Computer Science, Chair
	Rongxing Lu, PhD, Computer Science

This thesis is accepted by the

Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

January, 2017

Dedication

To God Almighty: Without you, no human work is possible. Thank Lord for giving me courage, health, strength and great men on my way for the achievement of this work. May the fruit of my labour never goes in vain but should always serves as a testimony of your infinite goodness on earth.
To my Parents Mn Fon Tebit Tenjoh Confort and late Mr Fon Sakios: Here I'm dad, your prayers to God to take me through my master program have been accomplished. Though the road was so narrow, i gave the best of me. May your joy be expressed and celebrated in heaven. Thanks for your efforts and sacrifices toward the satisfaction of my needs. You remain in my heart. To my mother Fon Tebit Tenjoh Confort alias Mami Elder, PCC Bafoussam Cameroun, I love you, you will never feel the absence of dad for we, your children will continue to show you love and always be close to you.
To my Brothers, Roland Njeck and Tenjoh Kingsley: I thank you for your unlimited assistance and encouragement.
I dedicate this work to you all.

Abstract

According to APWG reports of 2014 and 2015, the number of unique Phishing e-mail reports received from consumers has increased tremendously from 68270 e-mails in October 2014 to 106421 e-mails in September 2015. This significant increase is a proof of the existence of Phishing attacks and the high rate of damages they have caused to Internet users in the past. Because no attention is made in the literature to specifically detect Phishing e-mails related to advertisement and pornographic, attackers are becoming extremely intelligent to use these means of attraction to track users and adjusting their attacks base on users behaviours and hot topics extracted from community news and journals. We focus on detecting deceptive e-mail which is a form of Phishing attacks by proposing a novel framework to accurately identify not only e-mail Phishing attacks but also advertisements or pornographic e-mails consider as attracting ways to launch Phishing. Our approach known as Phishing Alerting System (PHAS) has the ability to detect and alert all type of deceptive e-mails so as to help users in decision making. We are using a well known e-mail dataset and base on our extracted features we are able to obtain about 93.11% accuracy while using machine learning techniques such as J48 Decision Tree and KNN. Furthermore, we equally evaluate our system built based on these above features and obtained approximately the same accuracy while using the same dataset as input to our system.

Acknowledgements

I highly express my deepest appreciations to all of those who showed love, support and assistance for the completion of this report.

To God almighty: Who during all this time, a myriad of miles away from my family has continued to show me favour and grace, and put educated men on my way for me to get to this stage.

To My Supervisor Prof. ALI GHORBANI: Your quality of imminent researcher, your professional rigor, your leadership, your expertise and your great scientific culture impose respect and admiration. I want to express my deep gratitude regarding my reception in the Faculty of Computer Science UNB, Fredericton Canada, the support and assistance you expressed for the realization of this work. Thanks again for introducing me in the world of research mainly in the domain of security during my days as member of the Information and Security Centre of Excellence (ISCX) UNB where you serve as director. May your ardor, rigor and meticulousness in work be a model for me in the job market.

To Dr. Arash Habibi Lashkari and all researchers in Information

and Security Centre of Excellence (ISCX) lab, UNB, Fredericton, NB, Canada 2016: for their assistance and collaboration.

To My Lovely Fiancee Tiwa Diffo Edwige: Who always show me love each minute of her life.

To My Lovely Sisters and Brother Fri Fon Carine, Fon Atche Julienne, Fon Olive, Fon Ezem Britta, Boris and Brenda Ezem: Who have always been close to me no matter the miles separating us and who always share my pains and joy.

To my friends Kengne Charles, Guy Mouafo Tagoukam and others: Who have always been supporting and encouraging me.

To all The FON's Family and Relatives

Table of Contents

Dedication **ii**

Abstract **iii**

Acknowledgments **v**

Table of Contents **x**

List of Tables **x**

List of Figures **xi**

Abbreviations **xii**

1 Introduction **1**

1.1 What is Phishing? 2

1.1.1 Why Phishing activity increases? 3

1.1.2 Concluding Remarks 4

1.1.3 Report Organization 5

2 Phishing attacks and Detection Approaches **6**

2.1 Phishing attacks 6

2.1.1 Deceptive phishing attack 7

2.1.2 Malware-based phishing attack 8

2.1.3 Keyloggers and Screenlogger 9

2.1.4 Session hijacking 9

2.1.5 Web trojan 10

2.1.6 Hosts file poisoning 11

2.1.7 System reconfiguration attack 12

2.1.8 Data theft 12

2.1.9 DNS-Based phishing 12

2.1.10 Content-Injection Phishing 13

2.1.11 Man-In-The-Middle phishing 13

2.1.12 Search Engine Phishing 13

2.2 Social and Financial Implications of Phishing 14

2.3 Detection approaches 18

2.3.1 URL-Based approach 19

2.3.2 Content-Based approach 21

2.3.3 Combination of URL-based and Content-based approach 22

2.3.4 Concluding Remarks 22

3 Related Works 24

3.1 Recent Works on Phishing 24

3.2 Features 31

3.2.1 Common Features in previous researches 31
3.2.2 Concluding Remarks 33

4 The Proposed Detection Framework 44

4.1 Proposed solution: Phishing Alerting System 44
4.1.1 Alerting System Flow Chart 45
4.1.2 Alerting System Classification Process 46
4.1.3 Feature Selection 50
4.1.4 Features Definition 51
4.1.4.1 Phishing emails Detection 51
4.1.4.2 Advertisement emails Detection 51
4.1.4.3 Pornography emails Detection 52
4.2 Datasets 52
4.2.1 Concluding Remarks 53

5 Experiments And Evaluations 54

5.1 Experiment 54
5.1.1 System Features Test 55
5.2 Evaluations 56
5.2.1 Alerting System Testing 56
5.2.2 Concluding Remarks 56

6 Conclusions and Future Work 57

6.1 Conclusion 57

6.2 Future Works 58

Bibliography **64**

Vita

List of Tables

3.2 Selected Features From Related Works 32

3.1 Lists Of Features 34

4.1 Lists Of 18 Selected Features 50

5.1 Experimental Result Using 18 Selected Features 55

List of Figures

1.1 Phishing e-mails Life Cycle 3

2.1 Deceptive Mobile Attack 7

2.2 Deceptive Phishing Attacks Setup Stages 8

2.3 denial of service (DoS) 9

2.4 Manipulating the token session executing the session hijacking attack 10

2.5 Tribe Flood Network Attack 12

2.6 Redirecting Link 13

2.7 Man-In-The-Middle 13

2.8 Fake email from attacker 16

2.9 Fake Paypal website 16

2.10 Fake Facebook 20

4.1 Alerting System Flow Chart 45

4.2 Example of Email Header 53

5.1 Experiment With 18 Features 55

List of Symbols, Nomenclature or Abbreviations

APWG: Anti-Phishing Working Group.
FTP: File Transfer Protocol.
IRC: Internet Relay Chat.
PHAS: Phishing Alerting System.
DoS: Denial-of-service attack.
TLD: Top-level domain.
N_F: is the total number of features
F_i: is the feature name

Chapter 1

Introduction

The exponential connectivity to Internet of some devices such as computers, phones and other electronic devices facilitate their users to connect far or near to a myriad of organizations and systems. Hence, Internet can be defined as a crossroad where its users could meet and share data. This is the main reason phishers use this means of sharing data as a contact point to widely carry out Phishing activities by planting malware onto PCs so as to mislead users to counterfeit websites. According to APWG reports of 2014 and 2015, the number of unique phishing e-mail reports received from users has increased tremendously from 68270 e-mails in October 2014 to 106421 e-mails in September 2015 that makes phishing detection one of the hot research topics. Though efforts have been made in the literature to detect phishing e-mails, no attention was made to additionally detect some variants of phishing e-mails such as Advertisement e-mails and Pornography e-mails that are two attraction means by which e-mail users are mislead to Phishing. In this dissertation, we focus on deceptive phishing attacks and its variants such as attacks through advertisement and pornography e-mails and proposed a novel Phishing e-mail Alerting Framework to accurately detect and alert suspicious Phishing e-mails to users. The system is a rule based system that uses eighteen features extracted from links, websites content or e-mails content. We take into consideration advertisement e-mails and pornographic e-mails for these are some major ways of attraction used by Phishers. We evaluate our approach by testing our features with some machine learning techniques such as J48 Decision tree and KNN, then we tested the performance of our System using the same dataset of 9308 e-mails in which 6951 e-mails were legitimate and 2357 were Phishing spam. We claim that our system will perform well with live e-mails because links embedded in live e-mails are accessible and will facilitate content-based analysis.

1.1 What is Phishing?

According to APWG trend report of 2014 [4], Phishing is a criminal mechanism employing both social engineering and technical subterfuge to steal consumers' personal identity data and financial account credential. Phishers attempt to acquire these credentials -user names, password and credit card details by masquerading as trustwarthy entities while exchanging data with Internet users via email messages or during users interaction with a system.

In some cases, they appear as a third party in between users and a legitimate system without any awareness of their presence. As an example, a user connected to a bank system through his machine meanwhile the phisher appears in form of a malicious code install on the user's machine with the objective to intercept all keystrokes that will later be analyzed by the hacker in order to extract potential credentials from it. Attackers deceptive phishing activities operations follow a well know life cycle. Ammar Almomani, B. B. Gupta, Samer Atawneh, A. Meulenberg and Eman Almomani in [3] show us the life cycle of phishing emails, in Figure 1.1 as presented below.

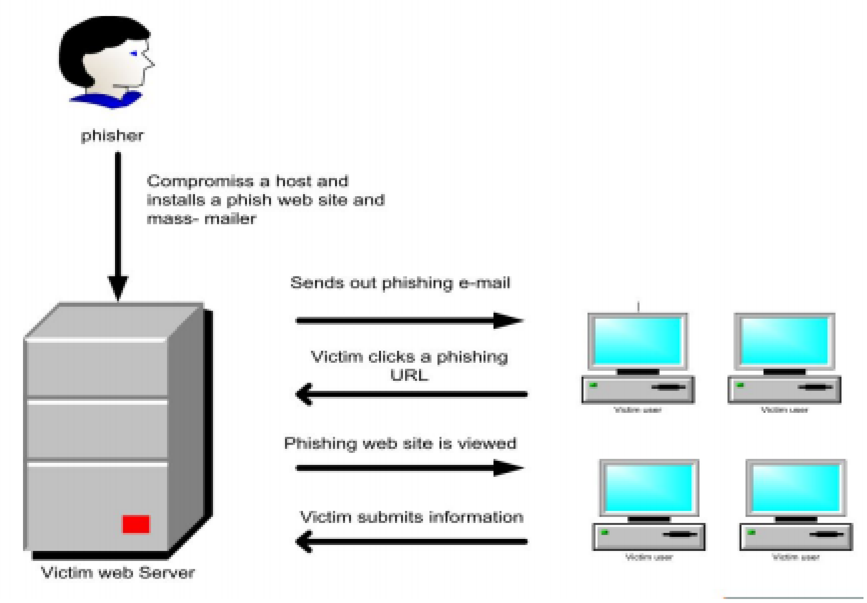


Figure 1.1: Phishing e-mails Life Cycle

1.1.1 Why Phishing activity increases?

Many reasons have contributed for the increase of Phishing activities. [3] points out that the necessity of technical resources to execute phishing attacks can be easily achieved through public and private sources. Equally, the automation of some Phishing technical resources have facilitated non-technical criminals to conduct phishing activities without any effort. One of the famous attack known as social engineering attack, increases rapidly because some Internet users are totally unaware of phishing and consequently cannot take any precaution when conducting online activities. Connected to systems such as bank, e-commerce systems, some Internet users lack knowledge concerning the policies of the system they are connected to, and ways for contacting system owners for issue related to privacy. This gives a door open to people conducting phishing to carry out their activities. Phishers are becoming more organized in their ways of thinking and operating. Their organization resulted to the creation of new ready-to-use phishing kits embedding items such as pre-generated HTML pages and emails for popular banks and e-commerce sites, scripts for processing user input, email and proxy server lists and even hosting services for phishing sites. With these kits creations, anyone connected to Internet can easily carry out phishing activities. According to APWG trends report of 2014, one additional reason this activity increases is due to the cheapness and freeness of domain name registration in some TLDs. The report stated that, in 95321 domains used for phishing, about 27253 (28,6%) were maliciously newly created domains due to the cheapness and freeness of domain name registration. This is one major reason the number of people acquiring domain names for fake activities increases exponentially every day.

1.1.2 Concluding Remarks

This chapter outlines some reasons phishing attacks increases and continue to make more victimes. It equally highlights the need to protect users e-mail account based on the high number of phishing e-mails reported to APWG.

1.1.3 Report Organization

This report provides a framework for emails phishing detection based on features extracted both from URLs and emails contents and is a snapshot of phishing attacks mode of operation with the intention to help email users to avoid being victims of Phishing. It's organized into six important chapters. The first chapter gives a brief definition of phishing and highlight some points making phishing a hot topic nowadays. Chapter two presents some well known phishing attacks, their modes of operation and detection approaches used in the literature. Furthermore, a list of features we extracted from literature review and our selected list of features are presented in chapter three of this report. Chapter four highlights our proposed framework for phishing emails detection, then chapter five talks about the experiments and evaluations of our framework and lastly Some perspectives that could be added to the proposed framework with the purpose to increase the detection rate of our system are highlighted in the conclusion chapter six as future works.

Chapter 2

Phishing attacks and Detection Approaches

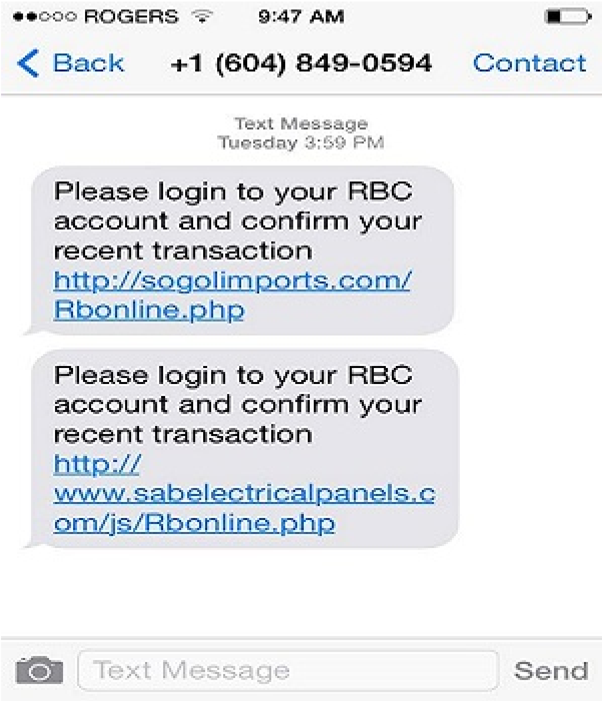
2.1 Phishing attacks

According to the 2016 Mobile World Congress [27], a myriad of types of phishing attacks exist for instance: `Deceptive phishing`, `Malware-based phishing`, `Keyloggers and screenloggers`, `Session hijacking`, `Web trojans`, `Hosts file poisoning`, `System reconfiguration attacks`, `Data theft`, `DNS-based phishing`, `Content-injection phishing`, `Man-in-the-middle phishing`, and `Search engine phishing` just to name some few.

2.1.1 Deceptive phishing attack

Deceptive phishing attack: This attack refers to social engineering attacks where users receive messages or e-mails which redirect them to bogus websites with the aim to steal personal information such as bank account number, social insurance and security numbers, account user name and passwords just to name some few. Hacker send messages alerting a problem that has to be solved rapidly while proposing to follow a link for solutions. As an example, the message could be *"your account has been compromised, click here to solve the problem"* or *"please confirm your account here"*.

The following Figure 2.1 is a real world example of deceptive message we received on a cell phone.



Please login to your RBC account and confirm your recent transaction http://sogolimports.com/Rbonline.php

Please login to your RBC account and confirm your recent transaction http://www.sabelectricalpanels.com/js/Rbonline.php

Figure 2.1: Deceptive Mobile Attack

The expectation of phishers is to mislead users to bogus websites where their personal information will be stolen.

In the case above, the hacker has created unique phishing form *Rbonline.php* in two different domains *sogolimports.com* website *"http://sogolimports.com/"* and *sabelectricalpanels.con/*. This is a typical deceptive attack because based on the message content the hacker is targeting potential RBC Bank clients meanwhile the confirmation links proposed by the messages has a domain which does not belongs to RBC bank. Only a small number of vigilant users are able to quickly identify them. It is important to understand how phishers set up themselves to achieve this type of attack. The following Figure 2.2

from [15] shows how deceptive phishing attacks are set up.

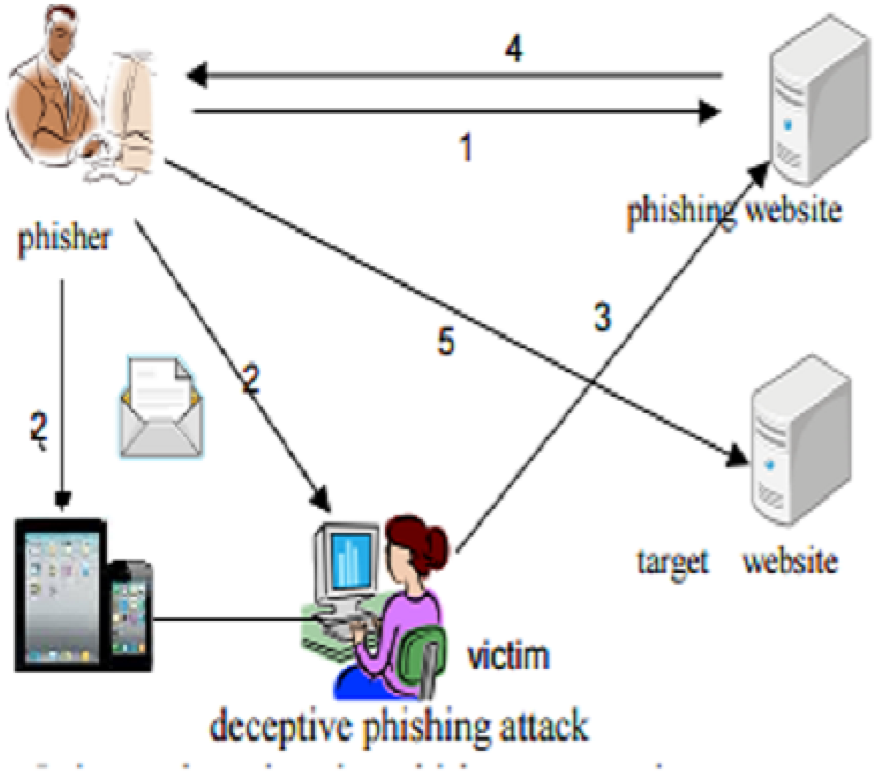


Figure 2.2: Deceptive Phishing Attacks Setup Stages

Deceptive attack steps:

1. The Phisher has to set a phishing website where all information entered to this site is posted to him. After user credentials are posted to the hacker, the hacker might effectively redirect the user to his/her original bank account while planing to connect to this account after the authorized user is disconnected.

2. After the phishing website is setup, the hacker has to broadcast phish-

ing messages to potential victims phones or PCs. This luring messages are meant to attract users to follow some links to bogus sites.

3. Non vigilant users follow links inserted in messages or e-mails they received, leading them to bogus site.

4. After the user connects and enters his Bank account number and other credentials, the phishing website will post them to the hacker.

2.1.2 Malware-based phishing attack

Malware code is installed on users' PC when he tries to open a malicious file attached to an email or download a file from a malicious website. This code could have the aim to compromised and join the corresponding PC on which it is installed to a botnet and the phisher as the botmaster will be able to conduct a general DoS as shown below in Figure 2.3 from [33].

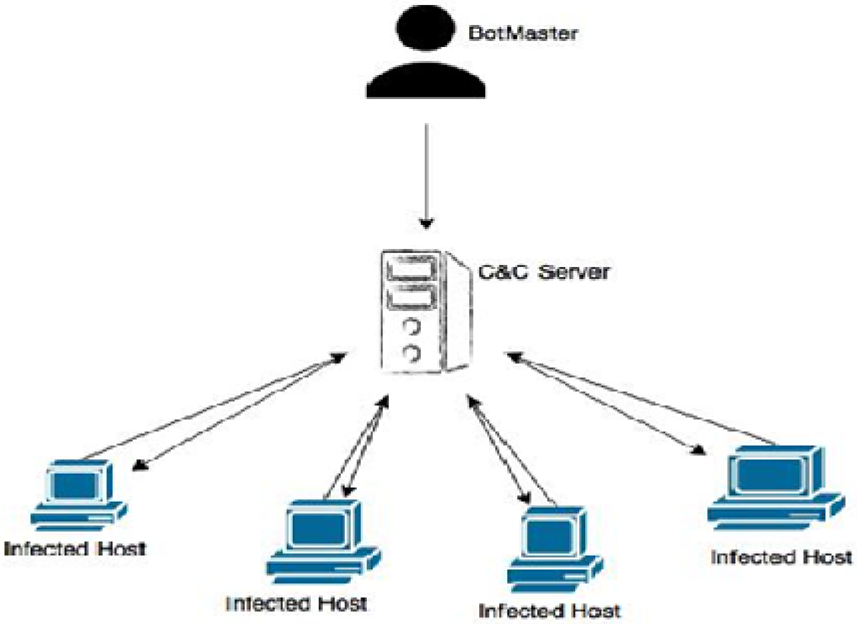


Figure 2.3: denial of service (DoS)

2.1.3 Keyloggers and Screenlogger

Keyloggers and Screenlogger describe respectively their program's function. They are respectively in charge of saving all characters entered via keyboard and all screen-shot of pages opened by users. Based on their respective functionalities, they can be used for legitimate services or for hacking. however when installed illegally on a PC's with the objective to track users, steal their personal information and post them to an unknown party, then they can be seen as malware.

2.1.4 Session hijacking

This attack monitors user activities on the Internet and once the user enters a targeted account credential to be submitted to a given server, the attack will steal the credential for future unauthorized activities in the place of the legitimate user by manipulating session execution token. The main objective of this attack is to intercept the token corresponding to a user session execution and later use this token to connect to the server as the legitimate user. Figure 2.4 below from [26] clearly presents this scenario.

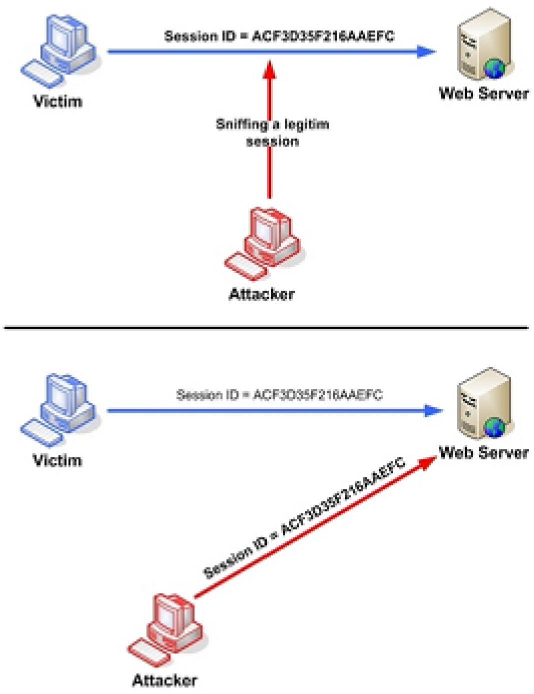

Session ID = ACF3D35F216AAEFC
Victim
Web Server
Sniffing a legitim session
Attacker
Session ID = ACF3D35F216AAEFC
Victim
Web Server
Session ID = ACF3D35F216AAEFC
Attacker

Figure 2.4: Manipulating the token session executing the session hijacking attack

This figure above shows session sniffing which is one way by which session hijacking attack operates. Overall there exist five ways or processes of operation of session hijacking attack:

1. Monitoring: This mode of operation of an attacker consists to monitor the flow of packets the victim shares with the target and predicts the sequence number.

2. Sniffing: In this attack mode of operation, the attacker places himself between the victim and the target server.

3. Session ID prediction: In Session ID prediction, the attacker takes over the session.

4. Session DE synchronization: In this form of attack, the attacker makes sure to broke the connection between the victim's machine and the target machine.

2.1.5 Web trojan

This is an invisible attack that occurs when a user is accessing a system such as online bank systems. When the user tries to enter his/her account credential, an invisible page pop up in order to collect user credentials and forward them to phishers.

Trojan does not propagates by self-replication but relies heavily on the exploitation of an end-user. It sometimes appears as an unsuspected mobile application, program, or hyperlink with the intention to enhance end-user attraction to click links that generates many problems such as the installation of backdoors on the computer system, the corruption of files in a subtle way, the uploading and downloading of files files, files encryption, allowing remote access to the victim's computer, making screenshots in order to Phish for bank or other account details which can be used for criminal activities. There exist a myriad of Trojan payload among which we have the Remote Access, Data Destruction, Downloader, Server Trojan(FTP, IRC, Email, Proxy, HTTP/HTTPS, etc.), Security software disabler and Denial-of-service attack(DoS).

2.1.6 Hosts file poisoning

In this type of attack, hacker poisons the host file because some Pc run system that always lookup "host names" in their hosts file before any Domain name system is carry out. Hence, hackers are sure to mislead users to bogus sites.

2.1.7 System reconfiguration attack

This is a type of attack that modifies some settings in a user's Pc in order to conduct malicious activity.

2.1.8 Data theft

In this type of attack, hosts are being infected and used as a point of contact to legitimate servers with the objective to compromise them and steal confidential communication, governments secret information. This way of compromising host can be seen in Tribe Flood Network attack where the Tribe flood network host instructs the daemon host to attack the victim host as shown in Figure 2.5 from [25].

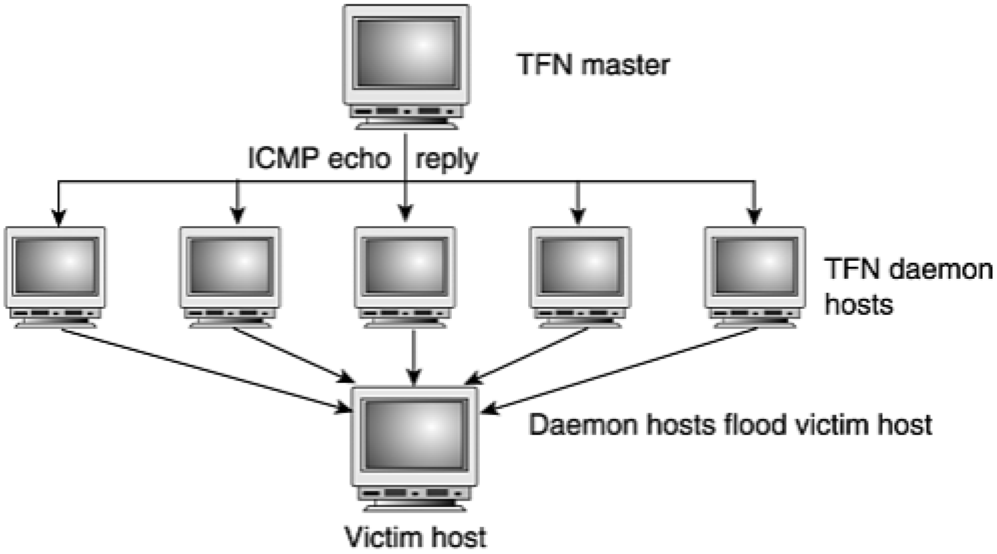


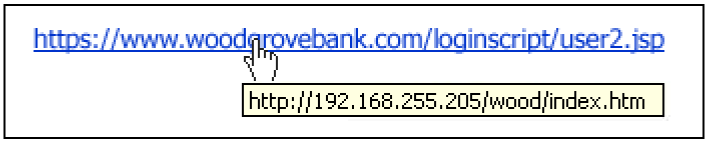
Figure 2.5: Tribe Flood Network Attack

2.1.9 DNS-Based phishing

Known as ”pharming”, which is a term given to the act of modifying hosts file or Domain Name System (DNS)-base phishing in order to return fake content for any legitimate content request via URLs.

2.1.10 Content-Injection Phishing

In this attack, codes are added or modified from legitimate sites content in order to either redirect users to phishing websites as presented below taken from Microsoft report [22], or by collecting user credential through forms and posting them to hackers. The below Figure 2.6 is a real example of this type of attack.



https://www.woodgrovebank.com/loginscript/user2.jsp

http://192.168.255.205/wood/index.htm

Figure 2.6: Redirecting Link

2.1.11 Man-In-The-Middle phishing

This is One of the most dangerous attack in the sense that all user activities are seen and can be manipulated by a third party.An active user interacting with a legitimate system without a SSL connection is easily mislead with the presence of man-in-the-middle who could records all information entered by the user and wait until he/she is deactivated to carry out unauthorized action using the information recorded as shown in [11]. The following Figure 2.7 is a good example illustrating this type of attack.

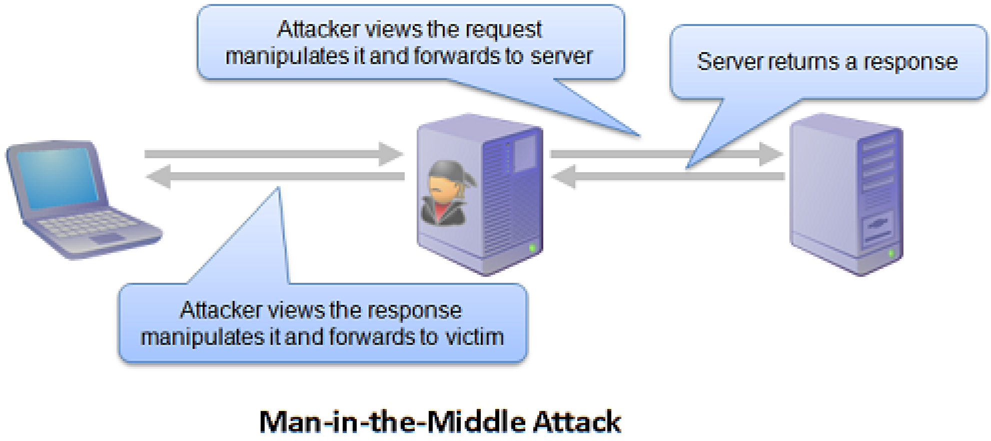

Attacker views the request manipulates it and forwards to server
Server returns a response
Attacker views the response manipulates it and forwards to victim
Man-in-the-Middle Attack

Figure 2.7: Man-In-The-Middle

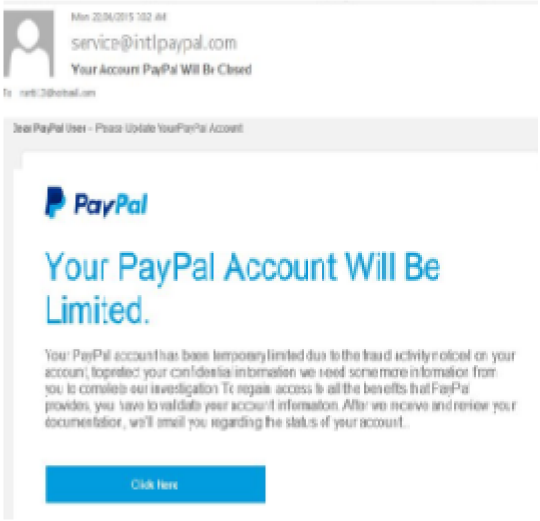
2.1.12 Search Engine Phishing

In this attack, hackers create attractive websites and have them recognized as legitimate with search engines. Once this is established, users are obviously fooled into giving up their secret information.

2.2 Social and Financial Implications of Phishing

The favourable economic, technological condition as well as socials media -facebook, twitter etc have massively contributed to the increase of phishing attacks in recent years. These attacks have impacted our society in many ways and have cause financial damages all over the world. In this section we will talk about potential victims of phishing, how they are being phished, by introducing some scenarios illustrating various phishing attacks modes of operations and further talk about the financial implication of phishing. Phishing techniques as mention in section 2.1, are used by criminal organizations around the world to acquire personal data via emails and webpages in order to fool financial institutions, disrupt computer operations, ruin reputations, destroy important data or lead Internet users into huge threat with shocking consequences such as loss of billions of dollas, running up of enormous debts that could lead to repossession of property. Phishing has a negative impact on the economy through financial losses experienced by businesses and consumers, along with the adverse effect of decreasing consumer confidence in online commerce and bank transactions. All Internet users with no exception could be a victim of this criminal act if no secure measures are taken into consideration. Hence, phishing attack affects everyone capable of carrying out any activity on the Internet and aims to steal money from accounts as legitimate owner. After collection of stolen bank information from users, Phishers can connect as legitimate users and launch a ransomware attack which is the act of blocking users important data that could only be released if they have paid a certain amount of money. People are getting more expose to Phishing attack cause by the high rate of data they exchange on social media nowadays. There exist a myriad of senarios by which phishing is setup.

1. A User receive an email in his email inbox which appears to come from a legitimate organization or bank such as PayPal stating that the user PayPal account may be suspended unless he log in and update his credit card details. This email contains a redirecting link that will redirect the user to a fake PayPal website containing a form with input fields for credentials collection. If the user log in and enters his credit card details into this form fields, then his credit card credentials for PayPal will be posted to the attackers. These type of scenario always ends with a "page not found" message set by the attacker after a successful attack. This page not found informs the PayPal user of a problem with the server responsible of the non availability of his PayPal account request. Figure 2.8 and 2.9 below extracted from [28] respectively illustrate this scenario that shows us a typical PayPal fake website that could be quickly identify by the differences in the domain name with that of the legitimate PayPal website. However, we can see some similarities in images logo with that of a real PayPal website.



Mon 22/06/2015 1:02 AM

service@intlpaypal.com

Your Account PayPal Will Be Closed

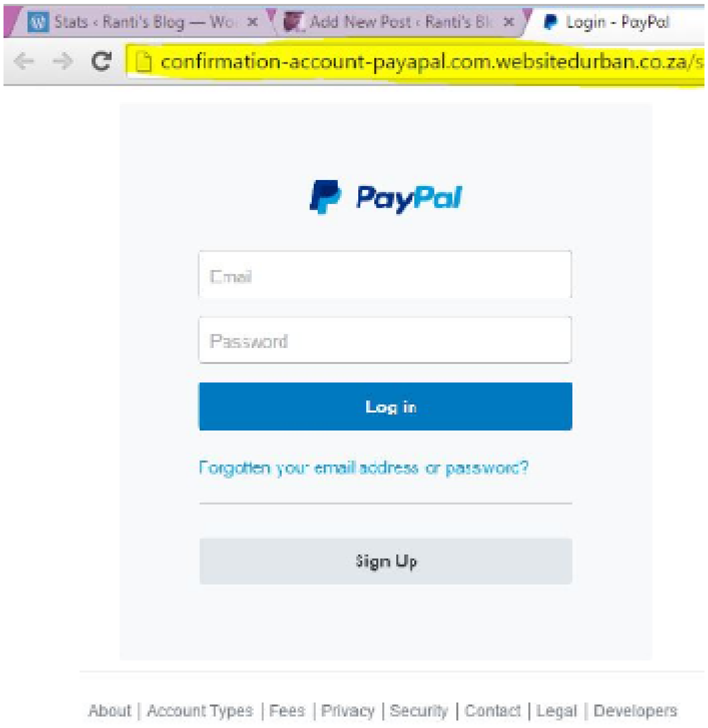
To: ranti_2@hotmail.com

Dear PayPal User - Please Update YourPayPal Account

Your PayPal Account Will Be Limited.

Your PayPal account has been temporary limited due to the fraud activity noticed on your account, to protect your confidential information we need some more information from you to complete our investigation. To regain access to all the benefits that PayPal provides, you have to validate your account information. After we receive and review your documentation, we'll email you regarding the status of your account.

Figure 2.8: Fake email from attacker



Stats ‹ Ranti's Blog — Wo
Add New Post ‹ Ranti's Bl
Login - PayPal

confirmation-account-payapal.com.websitedurban.co.za/s

PayPal
Email
Password
Log in
Forgotten your email address or password?
Sign Up
About | Account Types | Fees | Privacy | Security | Contact | Legal | Developers

Figure 2.9: Fake Paypal website

This scenario is classified under unawareness of threat and unawareness of policy that are two major factors criminals have been able to take advantage of. People should be aware of phishing attacks and various policies of organizations they deal with in order to avoid being victims of phishing attacks. Most organizations or banks don't communicate with their clients through e-mail and clearly mention this in their policies in order to avoid phishing. But because of the unawareness of these policies by most people, phishing continue to increase tremendously.

2. The Second scenario by which a user falls into Phishing is by clicking on an obfuscated URL. According to [12], there exist four types of commonly used URL obfuscation techniques that are classify in this report as phishing scenarios.

    (a) Type One refers to the host name obfuscation in which the host name is replaced with an IP address written in many formats (hexadecimal format, dword format, etc) in order to hide all signs that could help identify the fake URL. The obfuscated URL could content the targeted organization name for luring purposes. Example of Type one URL:
    http://210.80.154.30/ test3/signin.ebay.com/ebayisapidllsignin.html
    or http://$0xd3.0xe9.0x27.0x91 : 8080$/www.paypal.com/uk/login.html

    (b) The second type refers to obfuscating the host with another valid looking domain name but the URL path has been crafted and con-

tains the name of the organization being phished for redirection purposes. a typical example of this is as follow:
http://21photo.cn/https://cgi3.ca.ebay.com/eBayISAPI.dllSignIn.php or http://2−mad.com/hsbc.co.uk/index.html

(c) Type three attack focus on the long length of host name to achieve their attacks. The long length of host name is due to the fact that the hacker always try to include some legitimate domain names tokens in order to lure users. Example,
http://www.volksbank.de.custsupportref1007.dllconf.info/r1/vm/

(d) Type four refers to Domain name misspelled or unknown. Hackers make use of some similarity between words characters and numbers to misspell domain name or include characters to know domain names. Example, http://www.confirmation−account−payapal.com/ ; http://www.paypaI.com where ”l” is replace by ”I” in paypal.

3. A user could fall into phishing when he visits a malicious website making use of simple HTML redirection technique. The simple HTML redirection technique consists to make use of the content of the web page for obscuring the destination of a hyperlink by using a legitimate URL within an anchor element but have its ”href” attribute point to a malicious website. and example is shown as follow:
<a href =”www.hacker−payapal.com”> www.paypal.com < /a> ; In this case, www.paypal.com will redirect the user to a phishing website.

An aware user of this type of phishing scenario could easily avoid it by paying attention on the information display in the web browser status bar. In many cases, phishers construct phishing e-mails containing images and when displayed the appear to be legitimate images from legitimates organizations. These images are always logos and often belong to well known organizations such as Banks for attraction and luring purposes. If a user clicks on any of these images, then he will be redirected to malicious websites. Example:
<a href =”www.hacker−payapal.com”><img src= ”mimicpicture.jpg” alt=”Paypal”>< /a> ;

2.3 Detection approaches

There exist three phishing detection approaches:

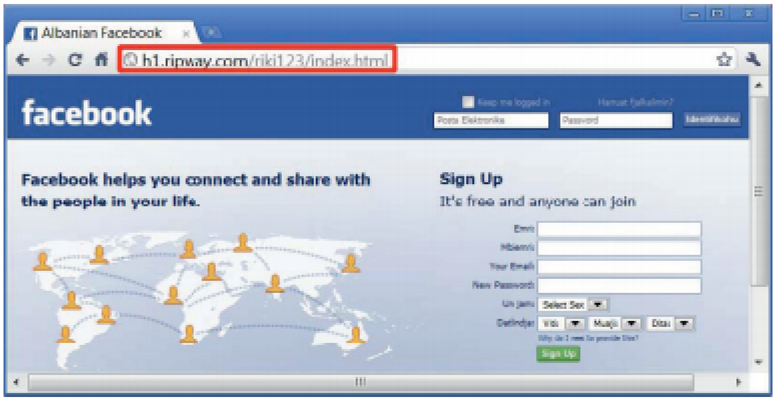
1. URL-Based approach
2. Content-Based approach
3. Combination of URL-based and Content-based approach

2.3.1 URL-Based approach

This approach uses only features extracted from a given URL to detect phishing. Ma, Justin et al in [20] presented features extracted from URL such as the length of the host name, length of the entire URL, the number of dots in

URL and lexical features or lexical properties as main features that could help in identifying phishing website with good accuracy. In addition to this features, Lakshmi, V Santhana and Vijaya, MS in [18] and Sanglerdsinlapachai, Nuttapong and Rungsawang, Arnon in [29] point out the the obfuscation of domain names with an IP address, the presence of '@' symbol in the URL and the number of dash in domain as some features characterizing phishing activities. The URL protocol combine with other feature also helps in phishing website identification process as mention in [2] and [14]. Some researchers such as Ma, Justin and Saul, Lawrence K and Savage, Stefan and Voelker, Geoffrey M in [21] and Le, Anh and Markopoulou, Athina and Faloutsos, Michalis in [19] mentioned the Whois lookup as a good feature that could increase the detection accuracy because it provides domain name registration information which is helpful to identify the owner of a website. They also mentioned a blacklist of words, that could be matched with words collected from URLs to determine phishing URLs.

The URL-Based approach is used for detecting deceptive phishing websites as well deceptive e-mails. Imagine an e-mail proposing a link to a user to connect to his/her Facebook account in order to confirm a registration. Once the user follow the link the following fake facebook page appears where credentials are being stealing. The following Figure 2.10 from [16] is an illustration of this scenario.



Facebook helps you connect and share with the people in your life.

Sign Up

It's free and anyone can join

Figure 2.10: Fake Facebook

One famous way hackers use URL-Based obfuscation approach is to make use of shortened URLs that can hide efficiently some indices for an obfuscated URL. Shortened URLs are short in length, in complexity and are a combination of service provider site and unique number or word. These types of URLs are to lead the users to the corresponding original website and are generated by some service providers that make their services affordable to all. For example if the service provider is http://goo.gl, then the shortened URL for http://en.wikipedia.org/w/index.php?search=shortened+url+&title=Speci -al%3ASearch&go=Go is http://goo.gl/VmwBNh. In case the service provider is http://bitly.com/ then the shortened URL becomes http://bit.ly/1kfh1P0. However, the knowledge of shortened URL is used by hacker to lead people to counterfeit websites. These types of URL are sent by hackers in email content and are practically difficult to be identify without any sufficient detection framework that takes into consideration shortened URL feature.

URL can equally be obfuscated by make use of the symbol '@' in the URL. Text appearing before this symbol is considered to be a comment. Therefore a web page URL containing this symbol should be directly classify as suspicious for this is a luring way used by hackers. An example of this could be http://www.Paypal.com/@http://Payapal.com. In this URL, http://www.Paypal.com will be considered as a comment and the URL that will be opened by the browser is http://Payapal.com which is a fake website.

2.3.2 Content-Based approach

This approach focuses on features extracted from a website HTML code or the content of an e-mail. Some URL features as named above are still useful in the content-base approach when dealing with Links extracted from the HTML code or e-mail content. Web pages containing more external links than internal ones and password field input are classified as suspicious. Basnet, Ram B and Sung, Andrew H and Liu, Quingzhong in [6] explained that a website content with more external links than internal links is an attempt to achieve some similarities and styles from external sources with the objective to steal user credential. Equally, Moore, Tyler and Clayton, Richard in [23] state that phishing sites using original graphics from original websites can now be detected. However, nowadays, most phishing sites use local copies of graphics to achieve similarities. Hence, to display a single phishing website, a huge number of distinct images is needed and can be seen as a feature for phishing detection.

Another feature for content-Based approach is the website tag <form> that can help to comfirm a web page is phishing. this tag is a means by which user's information could be leaked to phishers. Hence, in case an email contains a URL that leads to a website page containing the tag <form>, then this page as well as the email are considered to be suspicious.

2.3.3 Combination of URL-based and Content-based approach

This approach uses features selected from URLs and their corresponding content for bogus websites detection. This approach is more efficient for us because some URLs are well crafted and could not be quickly detected but their corresponding web pages contents could help to extract features that will help to classify them as phishing. Hence, though some research have succeeded to get good accuracy by using either each of these approaches, we strongly believe by combining the two approaches will lead to an efficient set of features that will help to get a high detection accuracy.

2.3.4 Concluding Remarks

Generally, this chapter deeply investigates how phishing attacks operate and some existing detection approaches such as URL based approach, content based approach and the combination of both approaches. It highlights some types of phishing attacks and presents for each of them the operational mode. Each of these attacks has the objective to steal users personal information and definitely has implications on society and financial institutions. The social and financial implications of phishing attacks as a whole are presented in this chapter as well as some scenario illustrating how people are being phished. These scenarios draw our attention to collect a set of features extracted from URLs embedded in email, web pages and web pages contents to accurately detect phishing attacks.

Chapter 3

Related Works

Many researchers have contributed to Phishing detection. from our reading, we selected a myriad of features for Phishing attacks detection as summarized in table 3.1 below.

3.1 Recent Works on Phishing

We selected a good number of research papers related to Phishing detection from 2007 to 2015 and summarised their works in order to select a collection of features commonly used that could yield high Phishing detection accuracy. In 2007, [12] proposed a method for Phishing web page detection base on the structure of web page URLs. They achieved their framework by studying set of URLs extracted from various known Phishing attacks and used the following features to achieve their framework { F_2, F_{98}, F_{34}, F_{45}, F_{99}, F_{100},

F_{101}, F_{102}, F_{103} }

In 2008, [5] used sixteen features for Phishing Detection, with some techniques such as Support Vector Machine (SVM), Biased Support Vector Machine (BSVM), Neural Network (NN) and Self Organizing Map (SOMs) to evaluate their framework. The set of features they used were: { F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11}, F_{12}, F_{13}, F_{14}, F_{15}, F_{16} }.

In 2009, [20] put forth a Phishing detection framework based on URL. They believe websites are characterized based on two groups of features, URLs lexical features and host-based features, then are classify base on their relationships. They evaluated their approach using some machine learning technique such as Naive Bayes (NB), Support Vector Machine (SVM) and Logistic Regression (LR) with features:{ F_{31}, F_{30}, F_{32}, F_{33}, F_{34}, F_{35}, F_{19}-F_{20}, F_{36} }

in 2009, [21] contributed in Phishing detection by combining host-based and lexical features from website's URL. Lexical features focus on URL appearance meanwhile host-based features are mainly focusing on identifying phishers and their ways of operation. Their lexical features account for 62% of the total number of features. They claimed their framework could yield good accuracy for real time detection because of the live source of labelled dataset of URLs and because the deployed system for feature collection was done in real time. They evaluated their framework using: Perceptron, Logistic Regression with Stochastic Gradient Descent, Passive-Aggressive (PA) algorithms and Confidence Weighted (CW) algorithms. Below are their elected

features:{ F_{34}, F_{36}, F_{81}, F_{36}, F_{82}, F_{83}, F_{30},F_{84}, F_{85}, F_{31}, F_{32} }.
In 2010, [29] used some Anti-Phishing and Network analysis tool (CANTINA) heuristic features with a new additional attributes that are combined with the new ones (-Domain Top page Similarity) to detect Phishing pages. The experiment was done in three metrics: first, with the intend to test CANTINA's reduced features, secondly, to test the new feature and thirdly to test the performance of some machine learning features using the combination of CANTINA's reduce features and the new feature. Their experiment used some machine learning techniques such as Naive Bayes (NB), Neural Network (NN), Support Vector Machine (SVM), Random Forest (RF), J48 Decision tree and Adaboost with the features set:{ F_{37}, F_{24}, F_2, F_{19}-F_{20}, F_{38}, F_{40} }.
In 2011, [19] used lexical features from URLs that are resistant to obfuscation techniques in order to identify Phishing. They claimed that their method can yield good accuracy by comparing the detection rate of their method with that of other techniques using combination of lexical and external features. They tested their idea using only lexical features, then mixed lexical and other external features. Bath-based SVM, online perceptron, confidence-weighted (CW) and adaptive regularization of weight (AROW) algorithms were used for testing mixed features. Though their technique yielded an accuracy of 1% less than when features are mixed, they still claimed their framework was best because features selection might be of great problem when mixed features are used. { F_{30}, F_{86}, F_{67}, F_{35}, F_{87}, F_2, F_{64}, F_{88}, F_{46}, F_{89}, F_{90}, F_{91}, F_{92}, F_{93} } are their features.

In 2011, [6] defined a set of rules in order to identify Phishing web pages. Two groups of rules were distinguished: -the simple rules, based on web page URL and -the higher complex and time consuming rules based on the analysis of meta-data, query search engines and blacklists - the search engine based rules, - the red flagged keywords based rules, - the obfuscation based rules, - the blacklists based rules, -the reputation based rules, - the content based rules. In order to build the rules, they identify and examine a tactic employed by phishers over a well known dataset of Phishing websites. Rules are initially assigned the same weight and while using a carefully chosen threshold, if the number of rules found in a web page is superior to the threshold, then the page is considered as Phishing. Decision tree and logistic regression were used and performed good results. The following are the features used: $\{ F_{29}, F_{70}, F_{71}, F_{72}, F_{76}, F_{7}, F_{28}, F_{38}, F_{95}, F_{25}, F_{96}, F_{31}, F_{42}, F_{8}, F_{97}, F_{2}, F_{24}, F_{91}, F_{19}\text{-}F_{20}, F_{34}, F_{35} \}$.

In 2011, [17] paid attention on obfuscation techniques operate on URLs domain name in order to detect Phishing web pages. Below are features used for this purpose: $\{ F_{4}, F_{5}, F_{76} \}$.

In 2011, [2] contributed in webpages Phishing detection by checking some characteristics of the web page source code that are not in respect with the W3C standards. A webpage is considered secured when the computed security percentage is 80% or higher, doubtful when it is between]50%-80%[and Phishing when it is less than 50%. Below are features selected to achieve their framework:$\{ F_{41}, F_{9}, F_{2}, F_{24}, F_{10}, F_{1}, F_{42}, F_{43}, F_{44} \}$.

In 2012, [18] extracted identities from some features such as Meta title, meta description, content attributes and ”href” attributes of tag <a> for detecting Phishing attacks. These features are tokenized with the objective to retained the first five keywords with high weight as identity set. Secondly, they extracted seventeen features base on the identities extracted in the previous step. Finally, the techniques used for evaluation were Multi Layer Perceptron (MLP), Decision tree induction and Naive Bayes. Their set of features was as follow: $\{ F_{17}, F_{18}, F_2, F_{19}\text{-}F_{20}, F_{21}\text{-}F_{22}, F_{23}, F_{24}, F_{25}, F_4, F_{26}, F_{27}, F_{28}, F_{29}, F_{30}, F_{31} \}$.

In 2012, [14] used a feature vector of size 23, in which four were structural features from URLs, nine were lexical features and ten were features targeting mostly brand and websites. SVM was used for experiment with feature set: $\{ F_2, F_{24}, F_{25}, F_{19}\text{-}F_{20}, F_{46}, F_{47}, F_{11}, F_{48}\text{-}F_{53}, F_{54}\text{-}F_{63} \}$.

In 2012, [7] evaluated two features selection techniques: The correlation-based and wrapper-based feature selection techniques. The correlation-based technique has the ability to generate a subset of features with the goal to improve the classification accuracy and reduce the feature dimension while exploiting the predictability of one variable with another. The wrapper based technique uses a machine learning algorithm while taking into consideration the fact that the method that has to use the feature subset should yield a good accuracy. Their experimental result demonstrates that a feature selection technique can improve classification results when trained and tested on a disjoint subset of dataset. They also show in their experiment that the wrapper-based classification technique significantly improved the accuracy compare to the correlation-based technique even though it was extremely slow. The feature selection techniques were evaluated using the following machine learning techniques: Naive Bayes (NB), Logistic Regression (LR) and Random Forest (RF). Features used were as follow: $\{ F_8, F_{24}, F_{42}, F_{29}, F_{31}, F_7, F_{38}, F_{19}\text{-}F_{20}, F_{46}, F_{94}, F_2, F_{24}, F_{11}\text{-}F_{16} \}$.

In 2013, [31] proposed a framework to detect hidden URLs based on lexical features. An URL is hidden if its corresponding page is hosted within a legitimate site without the site's administrator being aware. Hence, the HTTPs authentication becomes inefficient in detecting Phishing URLs. This approach uses only lexical features as shown below and yielded a high accuracy: $\{ F_{36}, F_{107} \}$.

In 2013, [9] used lexical and domain features extracted from Phishing URLs to detect Phishing websites. They Equally evaluate the effectiveness of machine learning based Phishing detection when they targeted websites are known. An optimal set of features was selected for this purpose: $\{ F_{64}, F_{65}, F_{66}, F_{67}, F_{68}, F_{69}, F_{70}\text{-}F_{75}, F_{76}, F_3, F_{77}, F_{78}, F_{79}, F_{80} \}$.

In 2014, [30] put forth a technique called Tabsol to fight against Tabnabbing. Tabnabbing is a recent variant of Phishing attack in which a malicious page opened in a tab disguises itself to a popular website's login page such as Gmail, Facebook login pages with the objective to steal credentials. The framework indentifies Tabnabbing attacks base on hash value comparison of the web page at different instances. This means that if any inconsistency of hash values is found between two states of a web page, then there is Phishing activities going on. An example of web page states are: When a page is focus(occupies the screen) and when the page regains its focus after the focus has been lost. Below are the features used in Tabsol: { F_{104}, F_{105}, F_7, F_{38} }.

In 2014, [32] proposed a technique for detecting Phishing web pages based on the discrepancy between the claimed identity and the domain name owner of the website. From a given web page, this technique extracts domain name, then tries to build a strategy to determine the domain name based on brand names from the web page content and compare to see if the extracted domain name matches the domain name generated based on brand names. In case of any mismatch found, they concluded a Phishing activities taking place. Below are features used for this purpose: { F_{39}, F_{29}, F_{104}, F_{106} }.

In 2015, [10] put forth a framework for Phishing web page detection based on URL analysis. The analysis consists to extract lexical features combine with bag-of-words approach for Phishing detection. They claimed that their framework achieved good accuracy while maintaining a low time. n-grams features, counting features, length features, pattern features and ratio features are some types of features they extracted from URLs. The features were: { F_{108}, F_{34}, F_{92}, F_{35}, F_{68}, F_{94}, F_{64}, F_{65}, F_{66}, F_{100}, F_{80}, F_{109}, F_{110}, F_{111}, F_{112}, F_{113} }.

Overall, an e-mail could redirect to a website via its links. More Phishing e-mails nowadays come in an attractive ways mostly as advertisements e-mails or pornography e-mails. None of these above researchers took into consideration advertisement e-mails and pornography e-mails and they equally do not pay attention on defining a set of rules that could lead to good classification. Based on the various operational modes of phishing attacks, we can conclude that an efficient detection of phishing email attacks could be successful by identifying a good set of features that may serve as rules for our framework. Hence, We proposed an Alerting system for detecting and alerting Phishing e-mails based on a well selected set of features. Rules are defined based on good collection of features which have shown high detection rate.

3.2 Features

The below table contains 112 features associate with their descriptions.

3.2.1 Common Features in previous researches

Table 3.2 below is a list of 20 commonly used features by researchers for websites Phishing detection. We select these features base on the number of times they are use in the set of research works presented above. These features were selected based on their rank value. We compute the Rank value of each feature as follow:

$$R_{F_i} = n/N$$

where $i \in \{1......N_F\ \}$, N_F is the total number of features, "n" the number of times a feature F_i has been used in all the set of papers read and "N" the total number of Papers. R_j represents the reference "j" which is the paper read, F_i the feature name. $j \in \{1......N_R\}$ where N_R is the total number of references or papers.

Commonly Used Features In Related Works	**Rank**
-IP Based URL (F_2)	0.529
-Number od Dots In Page Address (F_{19})-F_{20}, -Length of Host Name (F_{34})	0.352
-Symbol @ (F_{24}), -BlackList (F_{31})	0.294
-Number of Domain (F_4), -tag <form> (F_7), -Search Engine (F_{29}), -Whois Lookup (F_{30}), -Length of Entire URL (F_{35}), -Input Form (F_{38})	0.235
-Domain Age (F_3), -Number of Links (F_8), -Phish Words (F_{11} = update, confirm), -Server Form Handler (F_{25}), -Lexical Features or Textual Properties (F_{36}), -Iframe (F_{42}), -Number of dash in URL hostname (F_{46}), -Domain Token Count (F_{64}), -Domain Age Rank (F_{76})	0.176

Table 3.2: Selected Features From Related Works

None of these above researchers have worked on e-mail Phishing detection based on rules. Equally, these researchers did not take into consideration features for detecting advertisement and pornographic e-mails for they are attractive ways to launch deceptive attacks. We put forth an Alerting System framework base on rules that has the ability to detect and alert Phishing e-mails with its variants such as ad e-mails and porn e-mails.

3.2.2 Concluding Remarks

As mentioned in chapter 2 section 2.3.4, when phishing attacks mode of operation is known, features become accurate means for detection. However, before feature selection process, we found it necessary to carry out literature review on phishing attacks. This chapter outlines 112 features used in the literature for detection phishing attacks with a brief explanation for each of them. We give more priority to features with high occurrence from our literature review and create a set of features for our alerting system framework.

Table 3.1: Lists Of Features

Features	**Feature name**	**Description**
F_1	HTML Formatted e-mails	HTML formatted e-mails are mostly used for Phishing attacks because of their capability to use hyper-links and to insert some PHP functions in the HTML code which is a method mostly used for malicious code insertion.
F_2	IP-Based URL	IP-Based URL Is a URL or a Link in which the domain is an IP address. The IP could be written in many formats so as to hide all clues of identification.
F_3	Domain Age	A Domain Age shorter than 30 days characterizes domains created by attackers with the aim to quickly operate.
F_4	Number of Domains	A huge Number of Domains in a URL or link is one way to redirect users from one web page to another or from an e-mail message to a bogus website.
F_5	Number of Sub-Domains	Using a myriad of sub-domains is a way to facilitate URL obfuscation.
F_6	Token "JavaScript"	JavaScript codes are used to collect users credentials. Hence, the presence of the key word "JavaScript" in website html code could be suspicious.
F_7	Tag <Form>	Tag <Form> contains a method "Post" that defines the link where user input will be submitted. Hence, this feature validates the authenticity of this link.
F_8	Huge Number of Links	Huge Number of links is a characteristic of Phishing e-mails or Phishing websites for this is a means of similarities achievement.
F_9	Image Source	Image source commonly noted as "src" could contain a link value leading to a foreign domain. It could be suspicious in the sense that one can infer that the actual website is trying to copy an original image.
F_{10}	Matching Domains	Matching Domains is a characteristic of legitimate websites where The domain from an e-mail header should match with that of any link found in its body.
F_{11}-F_{16}	Keywords	Keywords are clues for detecting deceptive e-mails. Some commonly used ones are "verify", "update" and so forth.

F_{17}	Foreign Anchors	Foreign Anchors link to foreign websites. Hence, the usage of a huge number of these anchors could be an attempt to obtain a fake website similar to the original one.
F_{18}	Nil Anchor	Nil Anchor points nowhere. If the attribute ”href” of tag ¡a¿ belongs to the following set (” ”, ”void”, ”JavaScript”), then it is a Phishing sign.
F_{19}-F_{20}	Excess Dots in Page URL	Excess Dots in Page URL is when the URL contains more than five dots and is seen as suspicious.
F_{21}-F_{22}	Excess Slashes in URL	Excess Slashes in URL is when the URL contains more than five slashes, hence is consider as suspicious.
F_{23}	Foreign Anchors in Identity Set	Identity set is a set of words define to identify legitimate Foreign Anchors. Foreign anchors from a website are then tokenized one after another and their tokens are matched with tokens in this set. The non-membership of an anchor tokens in identity set is suspicious.
F_{24}	Symbol ’@’	Symbol ’@’ is a mean to introduce a comment to browsers. All text appearing before symbol ’@’ is a comment. Therefore the presence of this symbol in a link or URL is misleading and consider to be Phishing.
F_{25}	Server Form Handler	Server Form Handler takes different values depending on the value of the attribute ”action” which is an attribute of tag <Form>. When this value is empty, the web page is consider Phishing.
F_{26}	Foreign Request tokens	The Foreign Request (Requested URL) is tokenized and its tokens are matched with words in identity set. If all its tokens are in this set, then the related website is legitimate otherwise it is Phishing. The identity set is a set of legitimate words extracted from Meta title, meta description, content attributes and ”href” attributes of tag <a>.
F_{27}	Cookies	Cookies maintain the state of a user in the client and server side. The server can send the state information of a user to the client(browser) and the client does the same action as well. When the domain attribute of the cookie is different from the web page’s domain, then the related web page is Phishing.

F_{28}	SSL Certificate	SSL certificate assures a secure communication between clients and servers. Hence, most bogus websites do not have this certificate. If a certificate is generated, then the web page is legitimate otherwise it is a Phishing web page.
F_{29}	Search Engine	Search Engine results could be a means of ranking websites. Hence, a requested website from a search engine, should appear among the five first results otherwise the website is suspicious.
F_{30}	Whois Lookup	Whois Lookup is a set of databases in which most legitimate website's domains are registered. Hence, the absence of a domain name in Whois databases is suspicious.
F_{31}	BlackList	BlackList is a huge database of Phishing URLs that could help in classifying URLs.
F_{32}	Domain name Properties	Domain name Properties contain DNS record's Time To Live (TTL) value associated with the hostname. The DNS record's TTL value associated to a Phishing website is different from that of a legitimate website.
F_{33}	Geographic Properties	Geographic Properties reveal the geographical situation of an IP address and the uplink connection can greatly help in detecting Phishing sites.
F_{34}	Length of the Host name	The length of the Host name in a Phishing website's URL is different in length from that of a legitimate website because obfuscating a URL demands to insert some characters in their ASCII format.
F_{35}	Length of the Entire URL	Same reason as in F_{34}.
F_{36}	Lexical Features or Textual Properties	Lexical Features or Textual Properties consists of URL tokens. The website's URL is tokenized to form a bag of words for classification purposes.
F_{37}	Known Images	Known Images refers to the process of checking for any similarity between images found in bogus and legitimate websites.
F_{38}	Input Forms	Input Forms are all areas where users could input information. They check for block off <input> tags mostly those labelled as Pin code or password.

F_{39}	TF-IDF	TF-IDF stands for Term Frequency-Inverse Document Frequency and mainly focuses on website's content. Five words are retain from the suspicious page by using information retrieval techniques and combined with its URL domain name and inserted in the search engine to check if the engine will return any real page from this set of words.
F_{40}	Domain Top-Page Similarity	The Domain Top-Page Similarity Known as Web Category Comparison(WCC), checks whether a suspicious page is suited to be in its domain.
F_{41}	HTTPS Protocol	HTTPS Protocol is the secured protocol that should confirm the establishment of secure communication between a server and a client. However, hackers still find a way to lure users by using this protocol in webpage's body source code.
F_{42}	Iframes	Iframes are small frames in which a website could be loaded in another website. Phishers use this technique to load legitimate websites in bogus ones while hiding their borders so as to give the impression the bogus sites are authentic.
F_{43}	Script Tags	Script tags are used to include external files(JQuery, CSS) to website's source code. Hence, when a code is found in the place of a link between the tag <Script>then the entire webpage is suspicious.
F_{44}	Popup Windows	Popup Windows are small size windows that occasionally appear on websites and might ask users to input credentials to confirm their account information. The use of these features is suspicious.
F_{45}	Hostname length	Hostname length determines its number of characters. The greater than 22 this feature is, the more suspicious it becomes.
F_{46}	Number of dash in URL HostName	Number of dash in URL HostName represents the number of time the hyphens "-" is used in the host name. Excess usage of dashes in hostname is suspicious.
F_{47}	HTTP Protocol	HTTP Protocol does not guarantee secure connectivity between a client and a sever. Hence, all website's links or links found in e-mails with the protocol HTTP are suspicious.

F_{48}-F_{53}	Keywords	Keywords as define in F_{11}-F_{16} give us another set of words consider as a feature(F11, Banking, secure, ebayisapi, webscr, Log in, Sign in).
F_{54}-F_{63}	Brand Name	Brand Name defines a name by which a product maker is uniquely recognized. This feature checks if a brand name is correctly written. Example of brand names (PayPal, sulake, facebook, orkut, santander, mastercard, warcraft, visa, bradesco).
F_{64}	Domain Token Count	Domain Token Count Is The number of token a URL contains in its domain. The importance of this feature is due to the fact that hyphens are used to insert many tokens in the domain space that could lure users.
F_{65}	Average domain token length	Average domain token length represents the average length of tokens found in domain name of a given URL.
F_{66}	Longest Domain Token Length	Longest Domain Token Length determines the maximum length of a token in a domain name.
F_{67}	Path Token Count	Path Token Count is the number of token in the path portion of the entire URL.
F_{68}	Average Path Token Length	Average Path Token Length Represents the average number of tokens in the Path portion of the URL.
F_{69}	Domain Brand Name Distance	Domain Brand Name Distance Represents the distance between ”B” and ”S” where ”B” is the set of brand names of one or multiple websites and ”S” is the set of string(tokens) in the given domain. As an example, to protect Taobao website from Phishing attacks, B=(”taobao”, ”alibaba”, ”alipay”) is a set of three SLD names used by Taobao.
F_{70}-F_{75}	Search Engines	Search Engines are use to capture information on a website such as its link popularity. Eg: (”Domain Google links”, ”Domain Baidu links”, ”Domain bing links”, ”Domain yahoo! links”, ”SLD Google links”, ”SLD baidu links”). Most legitimate websites have high link popularity values compare to fake websites.
F_{76}	Domain Page Rank	Domain Page Rank is used by google search engine to rank pages according to their level of importance.
F_{77}	Domain Confidence Level	Domain Confidence Level tries to detect new Phishing websites that are sharing same domain with other old and known Phishing websites.

F_{78}	Path brand name distance	Path brand name distance is Partly similar to F36 but the set of words here comes from the entire path.
F_{79}	Domain Alexa Rank	Domain Alexa Rank provides domain ranking based on data collected over some period of time. That gives the ability to identify Phishing websites with low rank value.
F_{80}	Longest Path Token length	Longest Path Token length represents the number of tokens in a path. The higher this number is, the more probable the corresponding website is suspicious.
F_{81}	Primary Domain Name Token	Primary Domain Name Token refers to the name given By a website registrar. Collected tokens from this name constitute a bag of red flag words.
F_{82}	Top-Level Domain	Top-Level Domain refers to the last segment of a domain name. Example ”com” in ”example.com”, some times the top level domain ”.com” is fake. Hence, tokens from the top level domain portion of the domain name are collected for validation.
F_{83}	Last Path Token	Last Path Token represents the file extension part from the path.
F_{84}	Location	Location refers to the host's geographical location, IP address prefix and autonomous system number. An IP prefix is a pattern that matches the first 'n' binary of an IP address. Eg: 128.8.0.0 16 or 128.8 16 means to match the first sixteen bits of IP address 128.8.0.0 therefore, the matching is represented as 10000000.00001000.xxxxxxxx.xxxxxxxx. Hence, if malicious URLs tend to be hosted in a specific IP prefix of an ISP, then when classifying the URLs, the ISP should be taken into account.
F_{85}	Connection Speed	Connection Speed of a host is recorded when ever there are some evidences that malicious sites reside on machines.

F_{86}	Team Cymru	Team Cymru is a server that contains network information and geographical location associated to each URL. Some information obtained after querying this server are for instance: The network Border Gate Protocol (BGP) prefix, the Autonomous System number(AS), the country code and so forth. BGP prefix is composed of path of (AS) numbers, indicating which networks the packet must pass through and the IP block that is being routed. Eg of BGP prefix: 701 1239 42 206.24 14.0 24.
F_{87}	URL length and blacklist of Words	A Phishing URL is always long in length and could contains 54 tokens, with more than three dots. In most cases, at least one token of the URL belongs to the blacklist of words.
F_{88}	Port Numbers	Port Numbers are use by computers to communicate. Just as IP Address can be used to obfuscate domain name, port numbers are used as well.
F_{89}	Hyphens	Hyphens link words and are used also for other purposes. Eg:”-” in the domain name are used a luring technique helping phishers to associate brand names to their domain names so as to fool users.
F_{90}	Length of the directory	The Length of the directory is the number of characters of the website directory. The directory is extracted from the URL representing the folder in which the websites reside. This directory is suspicious when its length is greater than 6.
F_{91}	Number Of Sub-Directory’s tokens	The number of sub-directory represents number of directories found in the website’s directory folder. Bogus websites probably have a high number of sub-directories regarding of the huge data the store for similarity achievement.
F_{92}	Length Of File Name and Number Of Dots	From the entire path of an URL, the active page name is determined which is the file name. The length of the file name is obtained, its number of dots and delimiters as well. The objective of this feature is to make sure the file name is not replaced with an obfuscated host name.

F_{93}	Arguments Features	Arguments Features are extracted from server side Scripting language pages such as PhP, AsP pages. These arguments (length of argument, number of variable, length of longest variable value, maximum number of delimiters used in a value) are collected and compared to their corresponding threshold value before any further classification conclusion is taken.
F_{94}	Number of Underscore in URL	An excess Number of Underscore in URL is an attempt of linking brand names to Phishing URL in order to lure users.
F_{95}	Attribute Method In Form Tag	Attribute Method In Form Tag represents the method by which a From is submitted. It could be "get" or "Post". When the method in tag <Form> is "get", then the objective is to obtain all information entered by users into the form fields or password field. Hence when this attribute takes the value "get", the web page is suspicious.
F_{96}	Destination URL Refresh Properties and Meta-tag	Meta tag contains an attribute "http-equiv". When this later takes the value "refresh", and has an URL attribute which isn't in the same domain with the web page, then it's flag as Phishing. As an example, <meta http-equiv="refresh" content="0;url=http:/ example.com >. If example.com is in an external domain compare to that of the web page, then the web page is suspicious.
F_{97}	Password Field And Bad HTML markups	Password Field And Bad HTML markups is when a web page contains password field and possess some spelling mistakes that do not comply with the W3C standards. These types of websites are suspicious.
F_{98}	Obfuscating a Host With Another Domain	Obfuscating a Host With Another Domain is when the host portion of the URL has a valid look, but the URL path has the name of the targeted organization. However, redirection is the main purpose of this obfuscation type though the host part of the URL has a redirection URL containing the name of targeted organisation. For example: http:/ 21photo.cn https:/ cgi3.ca.ebay.com ebayISAPI.dllSignIn.php.

F_{99}	Domain Misspelled	Domain Misspelled is a domain name is unknown or misspelled. for instance, http:/ www.PayPa1.com. Here the domain PayPa1.com is misspelled.
F_{100}	Page Rank Of URL	The Page Rank Of URL determines how important a page is. Phishing web pages tend to have a short living period. Therefore, they have a lower page rank compare to legitimate web pages. Hence, if a web page URL has a low rank value in crawl Database, then it is Phishing suspicious.
F_{101}	Host Rank	The same idea as in F100 here with host part of the URL.
F_{102}	Page Rank Present in crawl database.	Page Rank Present in crawl database determines if a page rank is present in crawl database or not. if not, then it is Phishing. It is important to note that crawl database maintain scores that quantify the quality of a page.
F_{103}	White Domain Table	White Domain Table contains a list of legitimate domains and is a great feature to determine the authenticity of a domain.
F_{104}	White List Of URL	White List Of URL contains a list of legitimate URLs and is a great feature to determine the authenticity of a URL.
F_{105}	Hash Digest Of Source Code	Hash Digest Of Source Code keeps the state of a page. Its value is save in case any inconsistency is found with its newly computed value. In case its calculated values vary over time, the corresponding web site is flag as Phishing.
F_{106}	URL Weighted System	URL Weighted System is introduced introduced to establish TF-IDF primary condition stating that fairness metric between words occurrences should be taken into consideration.

F_{107}	HTTP field	HTTP fields are: -The Server Response-Header field which contains information about software used by the origin server to handle requests(eg: Server: Apache 2.2.14(win32)). -Content-Type Entity-Header Field which indicates the media type of entity-body sent to the recipient or in the case of the HEAD method which indicates the meta type that would have been sent. Eg: Content-Type: text html;charset: ISO-8859-4. -X-Powered-By: framework a web application that produces the content of the web resource(eg: A.SP.NET, PhP, Jboss). - Age feature. These four headers values just named are all combined as a unique feature. Each value of these fields is converted to a numerical value of 0 or 1 in order to build a matrix use as input in Phishing detection algorithms.
F_{108}	N-Grams	N-Grams check the similarity on common domain names. Its importance lies where domain names can be matched with a set of legitimate domain names after the similarity has been calculated.
F_{109}	Longest Token In Parameter	Parameter is the portion of the URL determining the file on the local repository. Phishers may obfuscate this parameter in order to make available Phishing page to web visitors by introducing long tokens in this parameter value.
F_{110}	URL Longest Token	Similar reason as in the case of F109 but applied to URLs.
F_{111}	Counting Features	Counting Features count the occurrences of characters such as: (’-’, ’@’, ’?’, ’.’, ’=’, ’%’, ’http’, ’www’, digits, numbers, letters, tokens, none alpha numeric characters, directories). Phishers mostly use more than one of these characters. Hence, this feature keeps track of the number of time these characters are used to be able to classify a URL as Phishing or not.
F_{112}	Pattern Features	Pattern Features count the occurrences of specific pattern within the URL. Eg: case changes, most consecutive occurrences of character, most frequent token, similarity in the blacklist of words, blacklist words count.

Chapter 4

The Proposed Detection Framework

4.1 Proposed solution: Phishing Alerting System

Filtering email content helps to identify Phishing scams spam and many other types of deceptive attacks. Researchers used collection of features extract from e-mails contents to detect scam mails. Some of these features are not efficients enough to accurately identify Phishing. Hence, we found our motivation from the used of ad e-mail by phishers as a means to achieve deceptive Phishing attacks. Spammers simply want to advertise a product while phishers or scammers deliver a message that looks like a legitimate one in order to steal sensitive information. These existing difference between a spam and scam e-mail in their definition does not make great difference in practice in the sense that both spamming and scamming are using attractive ways to reach their goals. The detection of Phishing emails could achieve a high accuracy as stated by Chandrasekaran et al in [8] when features like the structure of greeting provided in the email body are extracted and used, equally the frequency distribution of the selected function words such as "click, confirm" etc and a feature equals to the quotient of division of the total number of words found in the mail by it number of characters. This idea has been taken into consideration in our work with a good set of words we use as feature in our set of features to efficiently detect and alert deceptive e-mails.

4.1.1 Alerting System Flow Chart

Our system is base on eighteen features selected from Table 3.2 which have proved to work well together. Each feature is seen as a rule to flag Phishing e-mails in our system. An e-mail as input to the system is checked sequentially. Each level of e-mail verification could directly lead to a classification conclusion regarding the e-mail. Figure 4.1 shows the System flow chart. It should be noted that, for testing our system, we didn't connect to an e-mail account but we used as input the dataset described above.

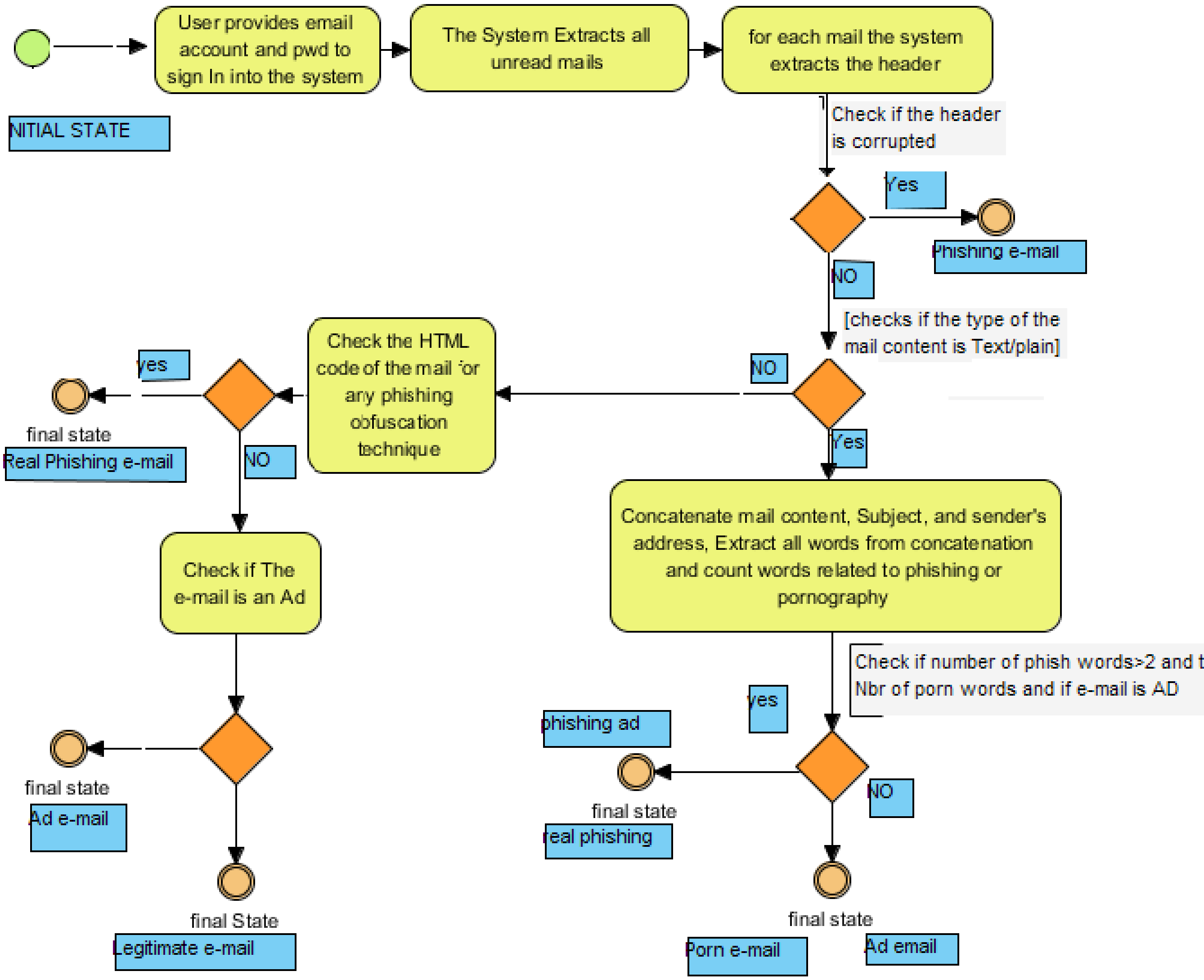

User provides email account and pwd to sign In into the system
The System Extracts all unread mails
for each mail the system extracts the header
NITIAL STATE
Check if the header is corrupted
Yes
Phishing e-mail
NO
[checks if the type of the mail content is Text/plain]
NO
Yes
Check the HTML code of the mail for any phishing obfuscation technique
yes
final state
Real Phishing e-mail
NO
Check if The e-mail is an Ad
final state
Ad e-mail
final State
Legitimate e-mail
Concatenate mail content, Subject, and sender's address, Extract all words from concatenation and count words related to phishing or pornography
Check if number of phish words>2 and t
Nbr of porn words and if e-mail is AD
yes
phishing ad
final state
real phishing
NO
final state
Porn e-mail
Ad email

Figure 4.1: Alerting System Flow Chart

4.1.2 Alerting System Classification Process

Once a new email arrives to the system, it goes respectively to the header classification process and depending on this classification result, it may equally be processed by the content-based classification process. The content-based classification process can undergo the text mail classification process or the multi-part mail classification process depending on the nature of the email content. The classification processes of our system are as follow:

Email header classification process: The objective here is to determine when an email is classify as Phishing while taking into consideration only its header. An email header is composed of the sender’s address, the subject of the mail and other fields. The sender’s address has two parts, both separated by the symbol ’@’. The header is Corrupted in any of the following cases:

If the first part of the sender’s address has more than one dash (feature Hyphens). For example: considering the e-mail address zxe234-ebay-market@yahoo.com, the first part zxe234-ebay-market has two dashes, therefore it is suspicious.

If the second part is an IP address, then it is characterized as domain obfuscation (feature IP Address). Hence, we check if this domain name is registered using the Whois lookup which is related to our feature ”Whois”.

If the domain in the email address has any of the following symbol (@; #; −; _; //) then the e-mail is Phishing. Features use are - Symbol ’@’ and Hyphens.

If the String resulting from the concatenation of the e-mail sender’s field and the subject, contains more than one word from the set of ”porn words” then we classify this e-mail as pornographic.

Email content-based classification process: The classification of an email from the content-based perspective depends on the email content type. An email content can be of two types, the Multipart/* type and Text/plain type. The type mulipart/* is the combination of either Text/plain or Text/html. Hence depending on the content type, the alerting system proceeds differently as shown in the system flow chart and explain below.

Text mail classification process: Text mail classification process has the ability to alert mails in which social engineering Phishing techniques are being apply, mail related to advertisement and those related to pornography. The process can be described as follow:

When the system starts, two features, porn words and phish words are extracted from the database in order to be used by the system. These features, which are bag-of-words are just to assist the system to take an accurate decision when the HTML code analysis seems to be legitimate. The text found in the mail is extracted and concatenated to the subject field and sender's field. We checked each word from this concatenation in our bags-of-words for any match. If a word is found in phish list or porn list, then the mail is likely to be an ad or porn. We then verify if it is an ad by checking if the e-mail generates more outbound flows than inbound flows which is related to one of our eighteen features called "More Outbound Flow" or by checking if the above concatenated text contains a word related to Phishing. If the test is positive then the mail is an ad. If the mail is an ad and has more phish words (>=2) compare to pornography words or the other way round, then this mail is classify as Phishing Ad. And the system will output the list of words found in the text which are related to Phishing or pornography when this was identified through word matching. Additional details to identify the mail are showed to users. Another situation is that we have a real Phishing mail in the sense that it is not an Ad but has more phish words (>=2) compare to pornographic words or the other way round.

Multipart mails classification process: The HTML code of each multipart mails is directly analysed in order to check the following obfuscation techniques:

Technique 1- If the HTML code has a <*form*> tag and the value of it method *"post"* has a domain which is not the same with the domain embedded in the sender's email's address, or in most of its links, then as a conclusion, this mail is suspicious. For instance, considering you received a mail from the email address roberto@yahoo.fr, this mail type is Multipart/* then it html code contains a "form" tag, with the value of the method *"post"* set to method post ="http://www.phishersdomain.com/..". In this case we conclude that the mail is Phishing because the *form* is not posted in the same domain. Again only the fact that this e-mail is link to a website that have an input "form" is suspicious. The feature "InputForm" from our set of eighteen features is used here.

Technique 2- All links of the HTML code are extracted and their domains are checked for any anomaly. If the domain name of a link is an IP address then the entire mail is seen as Phishing. For example if one of the links appears as http://www.143.45.21.0.com then the whole mail is Phishing. Feature IP Address is being used in this case.

Technique 3- If the HTML code has a broken link, then the whole mail is phish. A broken link leads no where. For example, <a href=" " ></a >or <a href="# " ></a >. Equally, each link is being checked whether any redirection technique is used, or any confusing digit or character is used in the domain space, or the link contains an excess number of dots, or more than two top level domain names are use at the same time, and or the domain name in links is not too long. However, from our set of 18 features, all features related to links are used at this level.

Technique 4- If none of the above obfuscation techniques is noticed, then we check whether the HTML code is generating more outbound flows than inbound flows. When the number of external links (links with domain different of the email address's domain) is greater than the number of internal links (Nbr external links >Nbr internal links) then this mail is Phishing.

Technique 5- If the previous step is not true, then for each link found in the HTML code of the mail, a new html code is extracted and is checked whether a link domain is an IP or whether the new html code has a form posting to a domain different from that of the sender's email address. If one of this conditions is found, then the related mail is Phishing otherwise the mail is legitimate mail or an ad mail. The mail is an ad when the number of external links is greater to the number of internal links and equally greater than 5 (Nbr external links >Nbr internal links).

4.1.3 Feature Selection

Feature Vector

We selected eighteen features for detecting Phishing deceptive attacks advertisement and pornographic e-mails. However our proposed system as shown above used them in a well organized way. Some features are extracted from links embedded in e-mails of type text/plan, text/html and the e-mail content. We are using these ad and porn features to guarantee a high detection accuracy of advertisement or pornographic e-mails for these are some new social engineering ways for attracting users to bogus websites. The following table 4.1 outline a list of our above system features. We give more details on these features below.

Features	**Features**
1) Excess Slash In Link	10) Excess Dots In Link
2) Hyphens In Link	11) More Images
3) Broken Links	12) Outbound Flow
4) Input Form	13) Phish Words
5) Number of Links	14) Porn Words
6) Redirecting Form	15) Obfuscating Top Level Domain
7) IP-Address	16) Hyphens In Sender's Address
8) Whois	17) Confusing Characters and Digits
9) Symbol @	18) Long Domain Name

Table 4.1: Lists Of 18 Selected Features

4.1.4 Features Definition

4.1.4.1 Phishing emails Detection

The alerting system helps users in decision making on the question to open or not open an email? The approach use features extract from the email sender's field, subject field, and e-mail content. When the message type is text/html, some html tags are check for any obfuscation usage. it is important to note that our system works for both live detection and off-line detection. for live detection, we directly connect to an e-mail account like yahoo account meanwhile for off-line, we had a dataset of Phishing and legitimate e-mails. The off-line detection was an efficient way to train our system features.

4.1.4.2 Advertisement emails Detection

Advertisement e-mails are known as attractive e-mails with the objective to mislead users to various bogus sites. The use of huge number of images is a way ad related websites are recognized. Equally, most of these ad websites are connected to a myriad of other sites which are frequently loaded in iframes. Advertisement e-mails use attractive words or phrases such as "promotion, new offers etc..". Once an Ad e-mail is detected, we analyse all associate links to detect any obfuscation technique. At this level the system is using the following Ad features: *More Images, Outbound Flow, Phish words.*

More Images: Ad e-mails always have a huge number of images from various websites. we flag these type of e-mails as suspicious.

Outbound Flow: Advertisement websites are known to present products from a myriad of websites in an attractive way. Those websites are not from the same domain reason why, once a website has more distinct links compare to domain related links, we conclude that the website is likely to be an ad website. This idea equally goes to e-mail messages. This feature flags e-mail messages with links having distinct domains.

Phish Words: The following are some words used to flag luring e-mails. { Sign in, log in, confirm, congratulation, bank, account, password, win, card etc..}

4.1.4.3 Pornography emails Detection

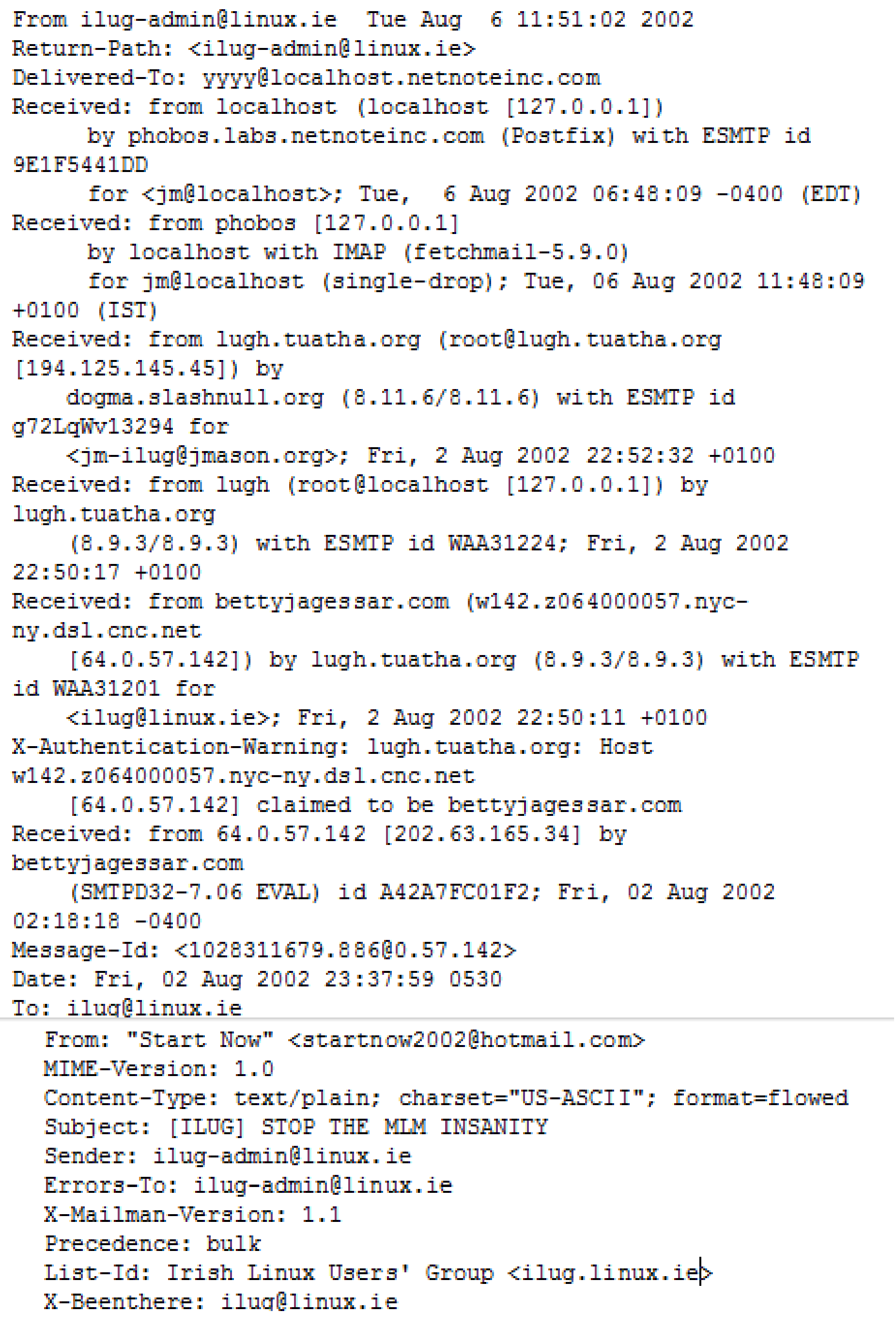
The same idea on Ad e-mails goes here. Their detection is based on attractive porn words such as "sex, smut, dirt, etc..".

Porn features are as follow: *More Images, Outbound Flow, Porn words.* The features More Images and Outbound flow are similar as in the case of Ad features.

Porn words: As example of set of porn words used by the system, we have { fuck, chick, sex, free porn, pussy, porno, F ck etc .. }

4.2 Datasets

We are using a publicly available Dataset of 9308 off-line e-mails to train our framework in which 6951 e-mails are labelled as legitimate and 2357 are phishing e-mails. The legitimate e-mails are from both the 2002 and 2003 ham collections, easy and hard from [1], and the phishing ones are from the publicly available phishing corpus [24]. All e-mails contain in The ham collection have their headers reproduced in full as shown in figure 4.2 below, some address obfuscation has been taken place for privacy purposes as well as hostnames in some cases have been replaced with ”spamassassin.taint.org” and in most cases, the e-mails headers appear as they were received. The easy-ham collection are typically quite easy to differentiate from spam, since they frequently do not contain any spam signatures such as HTML. The hard-ham are non-spam messages which are closer in many respects to typical spam: use of html, unusual html markup, coloured text and ”spammish-sounding” phrases.



```
From ilug-admin@linux.ie  Tue Aug  6 11:51:02 2002
Return-Path: <ilug-admin@linux.ie>
Delivered-To: yyyy@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
      by phobos.labs.netnoteinc.com (Postfix) with ESMTP id
9E1F5441DD
      for <jm@localhost>; Tue,  6 Aug 2002 06:48:09 -0400 (EDT)
Received: from phobos [127.0.0.1]
      by localhost with IMAP (fetchmail-5.9.0)
      for jm@localhost (single-drop); Tue, 06 Aug 2002 11:48:09
+0100 (IST)
Received: from lugh.tuatha.org (root@lugh.tuatha.org
[194.125.145.45]) by
    dogma.slashnull.org (8.11.6/8.11.6) with ESMTP id
g72LqWv13294 for
    <jm-ilug@jmason.org>; Fri, 2 Aug 2002 22:52:32 +0100
Received: from lugh (root@localhost [127.0.0.1]) by
lugh.tuatha.org
    (8.9.3/8.9.3) with ESMTP id WAA31224; Fri, 2 Aug 2002
22:50:17 +0100
Received: from bettyjagessar.com (w142.z064000057.nyc-
ny.dsl.cnc.net
    [64.0.57.142]) by lugh.tuatha.org (8.9.3/8.9.3) with ESMTP
id WAA31201 for
    <ilug@linux.ie>; Fri, 2 Aug 2002 22:50:11 +0100
X-Authentication-Warning: lugh.tuatha.org: Host
w142.z064000057.nyc-ny.dsl.cnc.net
    [64.0.57.142] claimed to be bettyjagessar.com
Received: from 64.0.57.142 [202.63.165.34] by
bettyjagessar.com
    (SMTPD32-7.06 EVAL) id A42A7FC01F2; Fri, 02 Aug 2002
02:18:18 -0400
Message-Id: <1028311679.886@0.57.142>
Date: Fri, 02 Aug 2002 23:37:59 0530
To: ilug@linux.ie
  From: "Start Now" <startnow2002@hotmail.com>
  MIME-Version: 1.0
  Content-Type: text/plain; charset="US-ASCII"; format=flowed
  Subject: [ILUG] STOP THE MLM INSANITY
  Sender: ilug-admin@linux.ie
  Errors-To: ilug-admin@linux.ie
  X-Mailman-Version: 1.1
  Precedence: bulk
  List-Id: Irish Linux Users' Group <ilug.linux.ie>
  X-Beenthere: ilug@linux.ie
```

Figure 4.2: Example of Email Header

4.2.1 Concluding Remarks

In this chapter, we proposed our detection framework called Phshing Alerting System which is our contribution in detecting phishing attacks. We equally provide the flow chart of this system as well as it classification process which is basically the process by which our system efficiently makes decisions to alert phishing e-mails. We further explained the set of 18 features used by our system and give more details on the dataset used to evaluate our framework.

Chapter 5

Experiments And Evaluations

5.1 Experiment

We evaluate our framework in two metrics. The first is to use some machine learning algorithms in Weka to test our selected features, and the second is to propose an alerting detection system to detect and alert Phishing. According to [13], Weka is a popular suite of machine learning software written in Java, and developed at the University of Waikato, New Zealand that contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to these functions. Weka supports several standard data mining tasks, more specifically: Clustering, Visualization, Feature Selection, Regression, Classification, Data Preprocessing. Two main weka tasks Among these tasks are important for us: Firstly, the classification algorithms mainly K-Nearest-

Neighbor (KNN) and Decision Tree (J48) algorithm to evaluate Phishing detection based on our eighteen features then secondly we used Feature Selection algorithms task to select best features in our set of eighteen features and re-evaluate our framework.

J48: This is an algorithm mostly used when a decision tree need to be generated for classification purposes. C4.5 is in charge of the decision tree generation and needs a valid data base to analyze.

KNN: KNN algorithm assigns each document(feature vector) to the majority of its k closest neighbors where K is a parameter. This algorithm produces good and accurate results in general mostly in the case where the value of K increases causing the complexity of the algorithm.

5.1.1 System Features Test

The below Table 5.1 gives us the experiment result when our selected features(18 features) are taken into consideration.

Algorithms	**Precision**	**Recall**
J48	0.916	0.911
KNN	0.934	0.931

Table 5.1: Experimental Result Using 18 Selected Features

The below Figure 5.1 gives more details on this experiment.

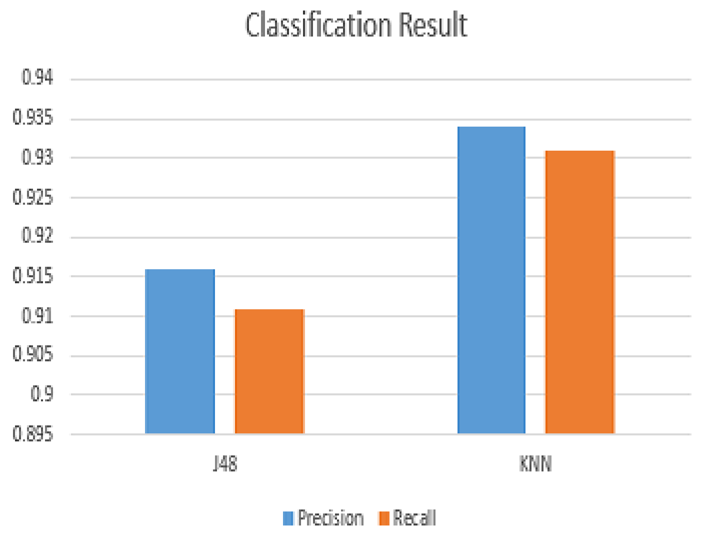


Figure 5.1: Experiment With 18 Features

5.2 Evaluations

5.2.1 Alerting System Testing

For testing, we use as input, our off-line dataset of 6951 legitimate e-mails and 2357 Phiishing e-mails. We apply the Alerting System Algorithm as explained in the Alerting System Classification Process section and found that we are able to detect all IP-based obfuscation techniques and other obfuscation techniques as well. In our case 2097 Phishing e-mails were accurately detected that gives us an accuracy of 88.96% which is closely the same with that of KNN and J48 using our eighteen selected features. The system output mostly advertisement e-mail and pornography e-mails as Phishing e-mails.

5.2.2 Concluding Remarks

This chapter put forth the evaluation of our framework which is done in two metrics. the first evaluats our set of features using machine learning techniques that has proved the efficiency of our selected features and has yielded an extremely good results for alerting phishing e-mails. The second metric build some rules to classify an email as phishing or legitimate while using our selected features for rules specifications.

Chapter 6

Conclusions and Future Work

6.1 Conclusion

Phishing has caused many losses all over the world and continue to increase its number of victims tremendously. It appears in many forms or types with distinct modes of operation. The variety of phishing operational mode gives us a hint to pay more attention on some features that could help to efficiently detect phishing attacks. Therefore to address the problem of phishing through e-mail, we proposed a successful phishing detection framework that uses features that have prove to be good in the literature and yielded high accuracy using machine learning techniques. This framework helps users to avoid being Phish through e-mails by alerting e-mails that are related to Phishing mostly appearing as ad or porn e-mails. Most of these ad mails generate excess outbound flows compare to inbound flows and are identify by our framework with high accuracy, but some ad sites do not generate more outbound flows reason why the system will only focus in this case on some ad words (bag-of-words) from the database for identification. Hence, the detection accuracy at this level can greatly increase by populating the list of ad and porn words in the database.

It is important to note that the alerting system only output Phishing mails, Ad email, and Phishing ad emails. From our experience, we noticed that by taking into consideration ad and porn e-mails, we have been able to alert most Phishing e-mails.

6.2 Future Works

Our framework known as Phishing Alerting System (PHAS) uses features extracted from e-mail content. These features can be categorized into three groups. features extracted from URLs embedded in the e-mail, features extracted from the web page URL and content. This web page can be open from a link embedded in the e-mail, and features extracted from the e-mail content itself. To extract features from the web page URL and content, our system needs sometimes to extract the web page HTML code before feature extraction. However, most e-mails from our e-mails dataset contain death links that contribute to a high false positive rate in our result. Hence, we strongly believed that, to avoid this problem, our system should be deploy using a dataset of live e-mails because in case they contain links, our system will avoid the problem of death links and reduce the false positive rate.

Our system equally considered the fact that Phishing web pages generate a high outbound flow, in order words, they contain a high number of links that are in different domain than theirs, in order to achieved similarity with the targeted legitimate website. However, we came to notice that, attacker now do not mainly use more outbound links but try to reduce their usage to avoid their sites to be suspicious. Hence, for future works, we should try to extend to number of features by adding features that could be of used to identify phishing websites with low outbound flow. Furthermore, we should try to come up with an amelioration of our strategy to select bag of words feature.

Other functionalities could be implemented at this level to automate the kind of decision a user could take regarding a mail that has been output by the system. However, for each alert, a reason why the mail has been highlighted as Phishing is being showed to users to assist them in decision making. It should be noted that advertisement emails are able to mislead a user to bogus sites because of their attractiveness that could lower users attention on some details to identify phishing activities. Overall, this system can still be extended not only with other features (functionalities) but also by improving the filtering criteria specified above. Our nearest future works would be to improve our system algorithm so as to reduce the false positive rate and re-evaluate our system using live e-mails.

Bibliography

[1] *Apache software foundation spamassassin public corpus*, 2006.

[2] Mona Ghotaish Alkhozae and Omar Abdullah Batarfi, *Phishing websites detection based on phishing characteristics in the webpage source code*, International Journal of Information and Communication Technology Research **1** (2011), no. 6.

[3] Ammar Almomani, BB Gupta, Samer Atawneh, A Meulenberg, and Eman Almomani, *A survey of phishing email filtering techniques*, 2013.

[4] APWG, *Apwg attack trends report*, 2014.

[5] Ram Basnet, Srinivas Mukkamala, and Andrew H Sung, *Detection of phishing attacks: A machine learning approach*, Soft Computing Applications in Industry, Springer, 2008, pp. 373–383.

[6] Ram B Basnet, Andrew H Sung, and Quingzhong Liu, *Rule-based phishing attack detection*, International Conference on Security and Management (SAM 2011), Las Vegas, NV, 2011.

[7] ———, *Feature selection for improved phishing detection*, Advanced Research in Applied Artificial Intelligence, Springer, 2012, pp. 252–261.

[8] Madhusudhanan Chandrasekaran, Krishnan Narayanan, and Shambhu Upadhyaya, *Phishing email detection based on structural properties*, 2006.

[9] Weibo Chu, Bin B Zhu, Feng Xue, Xiaohong Guan, and Zhongmin Cai, *Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing urls*, Communications (ICC), 2013 IEEE International Conference on, IEEE, 2013, pp. 1990–1994.

[10] Michael Darling, Greg Heileman, Gilad Gressel, Aravind Ashok, and Prabaharan Poornachandran, *A lexical approach for classifying malicious urls*, High Performance Computing & Simulation (HPCS), 2015 International Conference on, IEEE, 2015, pp. 195–202.

[11] DrJBHL, *Man-in-the-midddle attack*, 2015.

[12] Sujata Garera, Niels Provos, Monica Chew, and Aviel D Rubin, *A framework for detection and measurement of phishing attacks*, Proceedings of the 2007 ACM workshop on Recurring malcode, ACM, 2007, pp. 1–8.

[13] Stephen R Garner et al., *Weka: The waikato environment for knowledge analysis*, Proceedings of the New Zealand computer science research students conference, Citeseer, 1995, pp. 57–64.

[14] Huajun Huang, Liang Qian, and Yaojun Wang, *A svm-based technique to detect phishing urls*, Information Technology Journal **11** (2012), no. 7, 921.

[15] Huajun Huang, Junshan Tan, and Lingxi Liu, *Countermeasure techniques for deceptive phishing attack*, New Trends in Information and Service Science, 2009. NISS'09. International Conference on, IEEE, 2009, pp. 636–641.

[16] Jun Ho Huh and Hyoungshick Kim, *Phishing detection with popular search engines: Simple and effective*, Foundations and Practice of Security, Springer, 2011, pp. 194–207.

[17] Mahmoud Khonji, Andrew Jones, and Youssef Iraqi, *A novel phishing classification based on url features*, 2011 IEEE GCC Conference and Exhibition (GCC), 2011.

[18] V Santhana Lakshmi and MS Vijaya, *Efficient prediction of phishing websites using supervised learning algorithms*, Procedia Engineering **30** (2012), 798–805.

[19] Anh Le, Athina Markopoulou, and Michalis Faloutsos, *Phishdef: Url names say it all*, INFOCOM, 2011 Proceedings IEEE, IEEE, 2011, pp. 191–195.

[20] Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker, *Beyond blacklists: learning to detect malicious web sites from suspicious*

urls, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2009, pp. 1245–1254.

[21] ———, *Identifying suspicious urls: an application of large-scale online learning*, Proceedings of the 26th annual international conference on machine learning, ACM, 2009, pp. 681–688.

[22] Microsoft, *How to recognize phishing email messages*.

[23] Tyler Moore and Richard Clayton, *Discovering phishing dropboxes using email metadata*, eCrime Researchers Summit (eCrime), 2012, IEEE, 2012, pp. 1–9.

[24] J. Nazario, *Apache software foundation spamassassin public corpus*, 2006.

[25] Stephen Northcutt and Judy Novak, *Network intrusion detection*, Sams Publishing, 2002.

[26] OWASP, *Session hijacking attack*, 2014.

[27] PCworld, *2016 mobile world congress*, 2016.

[28] Ranti, *Fake paypal email*, june 22, 2015.

[29] Nuttapong Sanglerdsinlapachai and Arnon Rungsawang, *Using domain top-page similarity feature in machine learning-based web phishing detection*, Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on, IEEE, 2010, pp. 187–190.

[30] Amandeep Singh and Somanath Tripathy, *Tabsol: An efficient framework to defend tabnabbing*, Information Technology (ICIT), 2014 International Conference on, IEEE, 2014, pp. 173–178.

[31] Enrico Sorio, Alberto Bartoli, and Eric Medvet, *Detection of hidden fraudulent urls within trusted sites using lexical features*, Availability, Reliability and Security (ARES), 2013 Eighth International Conference on, IEEE, 2013, pp. 242–247.

[32] Choon Lin Tan, Kang Leng Chiew, et al., *Phishing website detection using url-assisted brand name weighting system*, Intelligent Signal Processing and Communication Systems (ISPACS), 2014 International Symposium on, IEEE, 2014, pp. 054–059.

[33] Chunyong Yin, Mian Zou, Darius Iko, and Jin Wang, *Botnet detection based on correlation of malicious behaviors*, International Journal of Hybrid Information Technology **6** (2013), no. 6, 291–300.

Vita

Candidate's full name:

KENNETH FON MBAH

Degrees and University attended

Bachelor of Science
Computer Science
Faculty of Mathematics and Computer Sciences
Department of Computer Science
University of Dschang, Cameroon
2007-2011

Master of Computer Science
University of New Brunswick
Fredericton Canada
2014-2017

Publications:

Mbah, K. F. , Lashkari, A. H. , Ghorbani, A. A. (2017). 'A Phishing Email Detection Approach using Machine Learning Techniques'. World Academy of Science, Engineering and Technology, International Science Index, Computer and Information Engineering, 3(1), 2383.