Detecting Phishing Email: Hybrid Feature selection

Manohar Kumar Kushwaha[1] Mr. S. Madhu[2]
[1]M.Tech. Student [2]Assistant Professor
[1,2] Computer Science & Engineering Department
[1,2] Galgotias University, Greater Noida, U.P.

Abstract— Phishing is a technique for acquiring your personal information; e.g. credit card or bank account numbers and passwords, and subsequently committing fraud in your name. E-mail has become an invaluable communication tool, both for business and personal use. Finally we can say that, Phishing email is a fraud method in which the perpetrator sends out legitimate-looking email in an attempt to gather personal and financial information from recipients. In other words, Phishing email will direct the user to visit a website where they are asked to update personal information, such as a password, credit card, social security, or bank account numbers, that the legitimate organization already has. The website, however, is bogus and set up only to steal the information the user enters on the page. In This paper we focused on hybrid feature selection approach that based on the combination of content-based and behavior–based. When hybrid features selections are used in Email phishing, which is able to achieve 97% accuracy rate.

Keywords: Behavior-based, Feature Selection, Phishing.

I. INTRODUCTION

Phishing is the criminally fraudulent process. That attempt to acquire sensitive information that means usernames, passwords and credit card details by playact as a legal trusted by customers in an electronic communication. Communications purporting to be from banks, online organizations, internet services providers, online retailers, and insurance agencies and so on. Popular social web sites such as YouTube, Facebook, MySpace, and Windows Live Messenger, auction sites (eBay), online banks (Wells Fargo, Bank of America, and Chase), online payment processors (PayPal), or IT Administrators (Yahoo, ISPs, corporate) are commonly used to lure the users. Phishing Email is big problems for internet. Most of the phishers carried out a three step. First as the, phishers crop the email address of their appreciable victims from social engineering attacks, WebPages and forums. Second as the when first step is completed then large volumes of phishing emails impersonating legal banking domains are sent out using anonymous SMTP (Simple mail transfer protocol) servers or compromise machine. These emails contain hyperlinks to lure the recipients into a masqueraded website with appearance similar to the legitimate domain. And third and last step is fake website contains input forms requesting personal critical information such as credit card , social security numbers and so on. Hybrid feature selection is one of the best methods to detecting phishing emails. Hybrid feature selection approach is based on the combination of content-based and behavior–based. We discuss behavior-based features in phishing emails which cannot be disguised by an attacker. We analyzed email header which is usually neglected by others. We considered examining the email's header specifically the message-ID tag and sender email in order to mine the attacker's behavior. This study applies the proposed hybrid feature selection to 6923 datasets which come from Nazario [14] phishing email collection ranging from 2004 to 2007 and SpamAssassin [21] as ham emails. When hybrid features selections are used in Email phishing, which is able to achieve 97% accuracy rate.

II. RELATED WORK

Phishing detection classified into two techniques first as the server based techniques and second as the client based techniques. Server based techniques are implemented by service providers such as, financial institutions, e-commerce stores or ISP. While client-based techniques are implemented on users end point through browser plug-ins or e-mail analysis. Several anti-phishing techniques have been proposed in recent years to detect and prevent the increasing number of phishing attacks. In general, phishing detection can be classified into server based techniques and client based techniques. Server based techniques typically are implemented by service providers such as ISP, e-comrnerce stores or other financial institutions. On the other hand, client-based techniques are implemented on users' end point through browser plug-ins or e-mail analysis. Many feature selection approach that recently introduced to assist phishing detection mechanism. Most of previous researches [6], [11], [16] were focusing on email content in order to classify the emails as either abnormal or normal. Previous attempt by [11] presents an approach based on natural structural characteristics in emails. The features included number of words in the email, the vocabulary, the structure of the subject line, and the presence of 18 keywords. They tested on 400 data which then divided into five sets with different type of feature selection. Their result shows the best when more features used to classify phishing email using Support Vector Machine classifier. However, the significance of the results is difficult to assess because of the small size of the email collection. Behavior-based phishing detection approaches has been proposed by [8], [5], [14]. Zhang et. al. [8] detects abnormal mass mailing host in network layer by mining the traffic in session layer. Toolan et. al. [5] investigated 40 features and proposed behavioral features: the number of words in send field, the total number of characters in the sender field, differences between the sender's domain and the reply-to domain and the differences between the sender's domains from the email's modal domain. Ahmed Syed at. al. [14] proposed behavioral blacklisting using 4 features which is log-based on live data. Generally, Email messages divided into two parts that are the header and body parts. The header contains information about who the message was sent from, the recipients date and the route which contains optional fields such as received, reply-to, subject and message-ID. This is then

followed by the body of the message. In our analysis, we considered the "message-ID" and the "from tag" in email header. We experimented with five features belong to email structure and additional two features which are extracted based on sender behavior. The features are Domain email sender (DES), Subject blacklist words (SBW), URL IP, (URLIP), URL dots (URLD), URL symbol (URLS).

III. PHISHING EMAIL FEATURE SELECTION

In this section, we describe System model, Hybrid feature selection, Feature Defines in Email, and Mining Sender Behavior. In its, email filtering approaches can be divided into two parts. Origin-based filtering and content based filtering. Origin-based filtering is based on the black verification list and white verification list. That means Origin-based filtering that focuses on source of the e-mail and verifies whether this source is on a white verification list or black verification list. Content based filters that based on the subject and body of the email. That means content based filters focus on the subject and body of the email. Phishing emails detected by filtering that based on structural feature, text feature or linguistic feature. Structural features focus on identifying the presence of obvious sign present in the email body, which implicate it to be fake. While the textual features and linguistic features identify phishing e-mails that based on the word composition and grammatical construction.

A. System model

Email message consists of following three components, first as the message envelope, second as the message header, and third and last as the message body. Message header contains control information, including, one or more recipient addresses and sender's email address. There are other descriptive information is also added, such as a subject header field, a message submission date/time stamp, message-id, and other information about the email. When an email is sent, the message is routed from sender's server to the recipient's email server through MTA (Mail Transport Agent). MT A handles message transportation and acts as sorting area and mail carrier. This is where every email messages is stamped with email header information including message-id. This part of email header is not visible to most users but it is a useful indicator in determining phishing email. After that, MTAs communicates with one another using the SMTP protocol. The recipient's MTA then delivers the email to the MDA (Mail Delivery Agent) that acts as an incoming mail server. MDA is a mailbox where it stores the email as it waits for the user to accept it. User then retrieve email using a software program called an MUA (Mail User Agent) such as Mozilla Thunderbird, Microsoft Outlook or Eudora Mail.

B. Hybrid feature selection

The Hybrid feature selection algorithm aims to determine feature matrix for predicting an email message is a phishing message or not. In Hybrid feature selection, we describe the proposed hybrid feature selection (HFS) algorithm. In this algorithm, we use domain email sender (DES), subject blacklist word (SBW), URL dots (URLD), URL symbol (URLS), URL IP (URLIP), Unique sender (US), Unique domain (UD) and DMID valid (DMID) feature values. We then developed a methodology to extract seven features from each email [7]. First, the email messages is partitioned into four components containing ES, SE, MID and URL. The inputs to HFS algorithm are DES, SBW, URLD, URLS, URLIP, US, UD and DMID as shown in Algorithm of HSF. In step 1, reading count for each email is done. For step 2 to 5, each incoming emails will run functions to verify sender domain, identify email's subject blacklist word, URL feature matching and identify sender behavior to extract features and finally construct the feature matrix.

Algorithm HSF

(1) FOR (each incoming EMAIL) DO
(2) FOR (i=1 to K) Do
(3) Verify sender domain;
(4) Identify blacklist word;
(5) Perform URL feature matching;
(6) Identify sender behavior;
(7) Identify Message-id validity;
(8) Constructing feature matrix;
(9) ENDFOR
(10) FNDFOR

END HSF.

In hybrid feature selection system, we used Bayes Net algorithm as our classifier as it is a powerful knowledge representation and reasoning mechanism. Moreover, it is the simplest and most widely used classification method because of its manipulating capabilities of tokens and associated probabilities according to the user's classification decisions and empirical performance. We used open source software: Mbox2xml as a disassembly tool. A python module mbox2xml exported the information from mbox format to xrnl (Extensible markup language) format. We modified some scheme in order to extract all features and store in the database. The next step in the process is to generate components of a feature vector by analyzing the database.

C. Feature Defines in Email

Email messages contain two basic parts first is the header and second as the body parts. The header contains information about who the message was sent from, recipients date and the route which contains optional fields such as received, reply-to, subject and message-ID. Five features belong to email structure and additional two features that are extracted based on sender behavior. Features in Email that is following:

1) Domain sender:
It is binary feature that represents the similarity of domain name extracted from email sender with domain message-ID. We think the email is normal if it is similar and set the value 0. If it is not similar, we set the value 1 to indicate the email is abnormal. This feature has been proposed by [5].

2) Subject blacklist words:
It is binary feature that represents the appearance of blacklist words in the subject of an email which included in bags of words in [11]. If the email subject contains the blacklist word, the email is abnormal and set the value 1. This feature has been used in [8].

3) URL IP:
It is shows the number of links that are using IP address.

4) URL dots:
Its represent a number of links in email that contains dots more than 5. This feature has been used in [11] but they calculate maximum number of dots in every link.

5) URL symbol:
Its represent the occurrence of links in emails that present symbol. This has been used in [18 0] but we incorporate other symbol such as "%" and "&"to detect obfuscation URL.

Two behavior features, first as the unique sender and second as the unique domain.

Unique sender (US): It is a binary data that represent sender behavior whether the sender sends emails from more than a one domain. If it is more than 1, we think the sender is phisher and set value is 1 or else the value is 0 to indicate that the sender is not phisher. (2) Unique domain (UD): This binary data denotes if the domain names is used by more than one sender domain email. If it is more than 1, we think the email is abnormal or else the email is normal and set the value to 0.

D. Mining Sender Behavior

In mining sender behavior, data mining for sender behavior is analyzed from email header. The dataset we selected from the email header has a structure as shown in Table I. Then all the features are defined, we extracted all 7 possible features from each email. The values of all features are in various types. Sender domain, subject blacklist word, unique sender and unique domain are in binary. All URL based features are in numerical however in vastly different ranges. For example, the URL dots could number of links under five. In order to treat all the original features as equally important, the value of each feature needs to be normalized before the classification process. Features with numerical values are normalized using the quotient of the actual value over the maximum value of that feature so that numerical values are limited to the range [0, 1].

IV. Conclusion and Future Direction

In this paper, we propose behavior-based features to detect phishing emails by observing sender nature. We fetch the all features using Mbox2xml as a disassembly tool. Then we sender behavior identify whether the email came from legitimate sender or not. We take into account behavior of sender who tends to send email from more than a single domain and a domain that handle different kind of email sender domain. Other than that, the attacker also used to forge the message-id field information to cover their tracks. Future works, we would like to investigate further on message-id field to understand the attacker strategies to cover their tracks.

V. References

[1] Isredza Rahmi A. Hamid and Jemal Abawajy" Phishing Email Feature Selection Approach" .
[2] The Anti-Phishing work Group, http://www.apwg.org/
[3] http://en.wikipedia.org/wiki/Phishing.
[4] Isredza Rahmi A HAMID, Jemal ABA W AJY Tai-hoon KIM
[5] TOOLAN, F., J. CARTHY, Feature Selection for Spam and Phishing Detection, In eCrime Researchers Summit (eCrime), 2010, pp. 1-12.
[6] FETTE, 1. , N. SADEH, A. TOMASIC, Learning to Detect Phishing Emails, Proceedings of the 16th International Conference on World Wide Web (WWW'07), ACM, New York, USA, 2006, pp. 649-656.
[7] ZHANG, J., Z. DU, W. LIU, A Behaviourbased Detection Approach to MassMailing Host, In Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, vol. 4, 2007, pp. 2140-2144.
[8] MA, L., B. OFOGHANI, P. WATTERS, S. BROWN, Detecting Phishing Emails using Hybrid Features,Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, 2009, pp. 493-497.
[9] Ammar ALmomani,Tat-Chee Wan, Ahmad Manasrah,Altyeb Altaher,2Eman Almomani, Karim Al-Saedi, Ahmad ALnajjar and Sureswaran Ramadass "A survey of Learning Based Techniques of Phishing Email Filtering"
[10] ZHOU, L., Y. SHI, D. ZHANG, A Statistical Language Modelling Approach to Online Deception Detection, IEEE Transactions on Knowledge and Data Engineering, vol. 20, No.8, 2007, pp. 1077-1081.
[11] CHANDRASEKARAN, M., K. NARAYANAN, S. UPADYAYA, Phishing Email Detection Based on Structural Properties, Proceeding of the Cyber Security Conference, 2006.
[12] CHANDRASEKARAN, M., V.SHANKARANARA Y ANAN, S.UP ADHY A Y A, CUSP: Customizable and Usable Spam Filters for Detecting Phishing Emails, Proceeding 3r Annual Symposium on Information Assurance (ASIA '08), Albany, NY, 2008, pp. 10-17.
[13] SYED, N. A., N. FEAMSTER, A. GRAY, Learning To Predict Bad Behaviour, NIPS 2007 Workshop on Machine Learning in Adversarial Environments for Computer Security, 2008.
[14] NAZARIO, J., Phishing Corpus, Available: http://www .monkey .org/jose/wikil doku. Ph p?id=phishingcorpus.
[15] Saeed Abu-Nimeh1, Dario Nappa2, Xinlei Wang2, and Suku Nair" Comparison of Machine Learning Techniques for Phishing Detection"
[16] ABU-NIMEH, S., D. NAPPA, X. WANG, S. NAIR, Comparison of Machine Learning Techniques for Phishing Detection, Proceeding of APWG eCrime Researchers Summit, Pittsburgh, ACM, New York, USA, 2007, pp. 60-69.
[17] Spamassassin public corpus, Available: http:/ /spamassassin.apache.org/publiccorpus.
[18] GANSTERER, W. N. D. POLZ, E-Mail Classification for Phishing Defence, in LNCS Advances, Volume 5478, 2009, pp 449-460.
[19] FERCHICHI, S. E., K. LAABIDI, S. ZIDI, Genetic Algorithm and Tabu Search for Feature Selection,Studies in Informatics and Control, ISSN 1220-1766, vol. 18 (2), 2009, pp. 181-187

[20] ISLAM, R., J. H. ABA W AJY, A Multitier Phishing Detection and Filtering Approach, Journal of Network and Computer Applications, vol. 36 (1), 2013, pp. 324-336.
[21] Spamassassin public corpus. Available: http:// spamassassin. apache. org/ publiccorpus.

