A STUDY ON WEBSITE PHISHING DETECTION USING DIFFERENT METHODS

Monika Bansal, Dr.Dinesh Kumar

Abstract— **Phishing websites is a semantic attack which targets the user rather than the computer. It is a relatively new Internet crime in comparison with other forms, e.g., virus and hacking. The phishing problem is a hard problem because of the fact that it is very easy for an attacker to create an exact replica of a good banking site, which looks very convincing to users. The most significant problem, is particularly relevant with the phishing corpus. The phishing problem is a hard problem because of the fact that it is very easy for an attacker to create an exact replica of a good banking site, which looks very convincing to users. Phishing is a threat in which users are sent fake emails that urge them to click a link (URL) which takes to a phisher's website to hack the secret information. There is Effectiveness Metric (EM) values of email classification features.**

Index Terms—**Effectiveness Metric (EM), Phishing, fake calls, mails etc**

I. INTRODUCTION

Phishing is an attack that makes Internet users reveal their personal information to unauthorised party. Most phishing attacks start when users receive fake emails asking them to click a URL (link) to update their accounts' information. Once clicked, this URL will deliver the user to a fake website where he/she will most probably lose control over its account information. According to Anti-Phishing Working Group report, the number of URLs which were used to host phishing attacks has increased from 164,023 in the first quarter of 2012 to 175,229 in the second quarter of the same year [1]. To detect phishing emails, it is important to choose the right detection feature(s). Among the available various antiphishing solutions, there is a considerable number of features which have been suggested to best classify ham (legitimate) and phishing emails. However, in many cases, these features are inappropriately chosen. This is because they are selected based on the author's intuition about their effectiveness in email classification process [2]. This work presents a method to choose the most efficient feature in detecting phishing emails. The importance of the selected feature is determined by calculating its Effectiveness Metric (EM) value based on three criteria which derived based on, and related to three general aspects of email. These three aspects of email are, email's sender, email's content, and email's receiver. Phishing websites is a semantic attack which targets the user rather than the computer. It is a relatively new Internet crime in comparison with other forms, e.g., virus and hacking. The phishing problem is a hard problem because of the fact that it is very easy for an attacker to create an exact replica of a good banking site, which looks very convincing to users. The word phishing from the phrase "website phishing" is a variation on the word "fishing". The idea is that bait is thrown out with the hopes that a user will grab it and bite into it just like the fish. In most cases, bait is either an e-mail or an instant messaging site, which will take the user to hostile phishing websites [7]. The motivation behind this study is to create a resilient and effective method that uses Data Mining algorithms and tools to detect e-banking phishing websites in an Artificial Intelligent technique. Associative and classification algorithms can be very useful in predicting Phishing websites. It can give us answers about what are the most important e-banking phishing website characteristics and indicators and how they relate with each other. Comparing between different Data Mining classification and association methods and techniques is also a goal of this investigation since there are only few studies that compares different data mining techniques in predicting phishing websites.

1.1. Feature Selection Process

The process of calculating the EM values of the Keywords and URLs features. EM values of these two features were calculated in order to compare their efficiency in detecting phishing emails. Since the email's body is the foremost part that users are concerning about and paying attention to, the features extracted from this part of the email are assumed to have higher importance in detecting phishing attempts than the features extracted from email's header part, and many of cues that influence user's decision about email(s) in question can be found in the email's body part [7]. The Body_no_FunctionWords feature (used in [2], and which is called the Keywords feature in this study) is a content-based feature which has not listed in Table II above. However, this feature has shown its importance in the experiment conducted in [2], it was ranked as the 1st, 16th, and 13th best amongst 40 features in three combinations of the three analyzed datasets in that experiment. In this work, we have focused on the Keywords and the URLs features which are extracted from the email's body part because these two features have a considerable importance . The Keywords feature was used to count occurrences of the selected 18 keywords in the two types of analyzed emails, whereas the URLs feature was used to count the presence and absence occurrences of fake URLs' indications in these emails. A. Feature's Effectiveness Criteria By considering email's sender, email's content, and email's receiver aspects, we have derived three effectiveness criteria which used in calculating the EM values of the Keywords and URLs features and hence to compare their efficiency in

Manuscript received April 25, 2015

Monika Bansal, M.Tech CSE -Research scholar, Department of CSE, Guru Kashi University, Bathinda(pb)

Dr.Dinesh Kumar, Associate Professor, Department of CSE, Guru Kashi University, Bathinda(pb)

detecting phishing emails. Each of these three criteria has given a 1 3⁄ of the effectiveness weight (effectiveness/3). Table III shows these effectiveness criteria and to which aspect of the email each criterion is relate to.

1.2. Phone Phishing Experiment

For our testing specimen, and after taking all the necessary authorization and approval from the management, a group of 50 employees were contacted by female colleges assigned to lure them into giving away their personal ebanking accounts user name and password (through social and friendly phone conversation with a deceiving purpose in mind). The results were beyond expectations; many of the employees fell for the trick. After conducting friendly conversation with them for some time, our team managed to seduce them into giving away their internet banking credentials for fake reasons. Some of these lame reasons included checking their privileges and accessibility, or for checking its integrity and connectivity with the web server for maintenance purposes, account security and privacy assurance…etc. To assure the authenticity of our request and to give it a social dimensional trend, our team had to contact them repeatedly for about three or four time. As shown in table 1, our team managed to deceive 16 out of the 50 employees to give away their full e-banking credentials which represented 32% of the sample. This percentage is considered a high one especially when we know that the victims were staff members of Jordan Ahli Bank, who are supposed to be highly educated with regard to the risks of electronic banking services. A total of 16% (8 employees) agreed to give their user name only and refrained from giving away their passwords under any circumstances or excuses what so ever. The remaining 52% (26 employees) were very cautious and declined to reveal any information regarding Response to Phone Phishing No. of Emp. Giving away their full ebanking credentials(user name & Password) 16 Giving away only their ebanking user name without password 8 Refused to reveal their credentials 26 Total 50 their credentials over the phone. An overview of the results reveals the high risk of social engineering security factor. Social engineering constitutes a direct internal threat to e-banking web services since its hacks directly into the accounts of e-bank customers. The results also show the direct need to increase the awareness of customers not to fall victims of this kind of threat that can lead to devastating results.

Table 1. Phone Phishing Experiment

Response to Phone Phishing	No. of Emp.
Giving away their full ebanking credentials(user name & Password)	16
Giving away only their ebanking user name without password	8
Refused to reveal their credentials	26
Total	**50**

II. LITERATURE SURVEY

Melad Mohamed et.al [2013] have proposed Phishing is a threat in which users are sent fake emails that urge them to click a link (URL) which takes to a phisher's website. At that site, users' accounts information could be lost. Many technical and non-technical solutions have been proposed to fight phishing attacks. To stop such attacks, it is important to select the correct feature(s) to detect phishing emails. Thus, the current work presents a new method to selecting more efficient feature in detecting phishing emails. Best features can be extracted from email's body (content) part. Keywords and URLs are known features that can be extracted from email's body part. These two features are very relevant to the three general aspects of email, these aspects are, email's sender, email's content, and email's receiver. In this work, three effectiveness criteria were derived based on these aspects of email. Such criteria were used to evaluate the efficiency of Keywords and URLs features in detecting phishing emails by measuring their Effectiveness Metric (EM) values. The experimental results obtained from analyzing more than 8000 ham (legitimate) and phishing emails from two different datasets show that, relying upon the URLs feature in detecting phishing emails will predominantly give more precise results than relying upon the Keywords feature in a such task.[1]

Rashid Chowdury et. al [2013] proposed a People now feel more comfortable socializing over the internet through popular social networking and media websites than face to face. Thus, the social media websites are thriving more and more nowadays. Like others YouTube is a vastly popular social media site which is expanding at very fast pace. YouTube depends mostly on user created contents and sharing and spreading. Business entities and public figures are taking advantage of this popularity by creating their own page and shared information among the large number of visitors. However, due to this popularity, YouTube has become more susceptible to different types of unwanted and malicious spammer. Currently, YouTube does not have any way to handle its video spammers. It only considers mass comments or messages to be part of spamming. To increase the popularity of a video, malicious users post video response spam, where the video content is
not related to the topic being discussed in the particular video or does not contain the media it is supposed to. In this research, we explore different attributes that could lead to video spammers. We first collect data of YouTube videos and manually classify them as either legitimate videos or spams. We then devise a number of attributes of videos which could potentially be used to detect spams. We apply Microsoft SQL Server Data Mining Tools (SSDT) to provide a heuristic for classifying an arbitrary video as either spam or legitimate. Our result demonstrates that in the long run we could successfully classify videos as spam or legitimate videos for most of the cases. [2].

Noor Ghazi et.al [2013], presented Phishing emails are messages designed to fool the recipient into handing over personal information, such as login names, passwords, credit card numbers, account credentials, social security numbers etc. Fraudulent emails harm their victims through loss of funds and identity theft. They also hurt Internet business, because people lose their trust in Internet transactions for fear that they will become victims of fraud. This paper deals with the phishing detection problem and how to auto detect phishing emails. The proposed phishing detection model is based on the extracted email features to detect phishing emails, these features appeared in the header and HTML body

of email. The developed model introduces statistical based parameters called features existence weight to decide whether the tested email is phishing or not. The results of the conducted testes indicated good identification rate (97.79%) with short required processing time (0.0004 msec.). [3]

Maher Aburrous et. al [2012] proposed Classification Data Mining (DM) Techniques can be a very useful tool in detecting and identifying e-banking phishing websites. In this paper, we present a novel approach to overcome the difficulty and complexity in detecting and predicting e-banking phishing website. We proposed an intelligent resilient and effective model that is based on using association and classification Data Mining algorithms. These algorithms were used to characterize and identify all the factors and rules in order to classify the phishing website and the relationship that correlate them with each other. We implemented six different classification algorithm and techniques to extract the phishing training data sets criteria to classify their legitimacy. We also compared their performances, accuracy, number of rules generated and speed. A Phishing Case study was applied to illustrate the website phishing process. The rules generated from the associative classification model showed the relationship between some important characteristics like URL and Domain Identity, and Security and Encryption criteria in the final phishing detection rate. The experimental results demonstrated the feasibility of using Associative Classification techniques in real applications and its better performance as compared to other traditional classifications algorithms. [4]

III. PROBLEM FORMULATION

There are different problems are faced from different literature survey that are given below:

- The most significant problem, which is particularly relevant with the phishing corpus.
- The phishing problem is a hard problem because of the fact that it is very easy for an attacker to create an exact replica of a good banking site, which looks very convincing to users.
- Phishing is a threat in which users are sent fake emails that urge them to click a link (URL) which takes to a phisher's website to hack the secret information.
- There is Effectiveness Metric (EM) values of email classification features.

IV. OBJECTIVE

- Phishing websites is a semantic attack which targets the user rather than the computer. It is a relatively new Internet crime in comparison with other forms, e.g., virus and hacking. The phishing problem is a hard problem because of the fact that it is very easy for an attacker to create an exact replica of a good banking site, which looks very convincing to users. Our Objectives are as follows:
- Data mining tool is used to implement the email phishing detection from websites.
- The implemented work to extract the phishing training data sets criteria to classify their legitimacy with six different classification algorithm and techniques.
- we also compared their performances, accuracy, number of rules generated and speed.
- The proposed work is to selecting more efficient feature in detecting phishing emails.
- The Effectiveness Metric (EM) values of email classification features are implemented.

V. METHODOLOGY

This work is to detect the intrusion from network. It is based upon weka tool. There are the programmable files containing the information about the dataset. From the previous study of phishing detection we managed to gather 27 phishing features and indicators and clustered them into six Criteria (URL & Domain Identity, Security & Encryption, Source Code & Java script, Page Style & Contents, Web Address Bar and Social Human Factor), and each criteria has its own phishing components. For example, URL & Domain Identity Criteria has five phishing indicator components (Using IP address, abnormal request URL, abnormal URL of anchor, abnormal DNS record and abnormal URL). Particularly, we used a number of different existing data mining association and classification techniques including JRip , PART , PRISM and C4.5 , CBA , MCAR algorithms to learn and to compare the relationships of the different phishing classification features and rules. The experiments of C4.5, RIPPER, PART and PRISM algorithms were conducted using the WEKA software system , which is an open java source code for the data mining community that includes implementations of different methods for several different data mining tasks such as classification, association rule and regression.

Rule 1: Social_Human_Factor = Fraud Web_Address_Bar = Fraud Page_Style_&_Contents = Doubtful -> class = Phishing

Rule 16: Web_Address_Bar = Genuine Security_&_Encryption = Doubtful URL_Domain_Identity = Doubtful -> class = Legitimate

Rule 22: Social_Human_Factor = Genuine Page_Style_&_Contents = Doubtful -> class = Suspicious

CONCLUSION

Phishing sites is a semantic assault which focuses on the client as opposed to the PC. It is a moderately new Internet wrongdoing in correlation with different structures, e.g., infection and hacking. The phishing issue is a hard issue due to the way that it is simple for an assailant to make an accurate imitation of a decent keeping money site, which looks exceptionally persuading to clients. The word phishing from the expression "site phishing" is a variety on "angling". The thought is that draw is tossed out with the trusts that a client will get it and nibble into it simply like the fish. In this paper different phishing methods are reviewed from the different researchers. In the future work different methods are implemented to get the better results.

REFERENCES

[1] Melad Mohamed et.al "A method to Measure the Efficiency of Phishing Emails Detection Features" IEEE- 2014.
[2] Rashid Chowdury et.al "A Data Mining Based Spam Detection System for YouTube" IEEE-2013.
[3] Noor Ghazi et al, "Detection Phishing Emails Using Features Decisive Values" International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 7, July 2013.

[4] Maher Aburrous et.al "Predicting Phishing Websites using Classification Mining Techniques with Experimental Case Studies" Seventh International Conference on Information Technology, 2010.

[5] AL Momani et.al "An Online Model on Evolving Phishing E-mail Detection and Classification Method", Journal of Applied Sciences, vol. 11, Issue 18, pp. 3301-3307, 2011.

[6] C. Castillo et.al , "Know your neighbors: Web spam detection using the web topology", In Int'l ACM SIGIR, pp. 423–430, 2007.

[7]] Z. Gy¨ongyi et.al "Combating web spam with trustrank", In International Conference on Very Large Data Bases (VLDB), pp. 576–587, 2004.

[8] P. Heymann et.al "Fighting spam on social web sites: A survey of approaches and future challenges", IEEE Internet Computing, 11(6):36–45, 2007.

[9]] C. Wu et.al "Using visual features for antispam filtering", In Proc. of 4th IEEE Int'l Conf. on Image Processing (ICIP), 2005.

[10] C. Shah et.al "Supporting Research Data Collection from YouTube with TubeKit". Proceedings of YouTube and 2008 Election Cycle in the United States, Amherst, MA: April 16-17, 2009.