Phishing Websites Detection Based on Web Source Code and URL in the Webpage

Satish.S[1], Suresh Babu.K[2],
Assistant Professor, Department of CSE,
Arulmigu Meenakshi Amman College of Engineering, Tiruvannamalai, India.
PG Schloar, Department of ECE, Arunai Engineering College, Tiruvannamalai, India.
satishstudy2012@gmail.com[1] , sureshstudy2010@gmail.com[2]

***Abstract*—** Major security issues for banking and financial institutions are Phishing. Phishing is a webpage attack, it pretends a customer web services using tactics and mimics from unauthorized persons or organization. It is an illegitimate act to steals user personal information such as bank details, social security numbers and credit card details, by showcasing itself as a truthful object, in the public network. When users provide confidential information, they are not aware of the fact that the websites they are using are phishing websites. This paper presents a technique for detecting phishing website attacks and also spotting phishing websites by combines source code and URL in the webpage.

Keywords*—Phishing, Website attacks, Source Code, URL.*

I.INTRODUCTION

Phishing is a type of practice done on the Internet where personal details are obtained by unlawful methods. It is an online kind of pretexting (rewriting or changing the original information) where fraud can take place by an attacker who appears to be someone else to get the most sensitive details from users [1]. Fraudsters looking to gather financial information have developed a new way to lure unsuspecting victims: they go phishing. In the first half of 2012, the RSA Anti-Fraud Command Center identified 195,487 unique phishing attacks – an increase of 19% as compared to the second half of 2011[3]. The word "phishing" originally comes from the analogy that early Internet criminals used e-mail lures to "phish" for passwords and financial data from a sea of Internet users. The use of "ph" in the terminology is partly lost in the annals of time, but most likely linked to popular hacker naming conventions such as "phreaks" which traces back to early hackers who were involved in "phreaking" – the hacking of telephone systems. The term was coined in the 1996 timeframe by hackers who were stealing America Online (AOL) accounts by scamming passwords from unsuspecting AOL users[2]. The most common purpose of phishing scams include:

• **Theft of login credentials** – typically credentials for accessing online services such as eBay, Hotmail, etc. More recently, the increase in online share trading services has meant that a customer's trading credentials provide an easy route for international money transfers.

Paper Type: Article
Corresponding Author: Satish.S, email: satishstudy2012@gmail.com

- **Theft of banking credentials** – typically the online login credentials of popular high-street banking organizations and subsequent access to funds ready for transfer.

- **Observation of Credit Card details** – access to a steady stream of credit card details (i.e. card number, expiry and issue dates, cardholder's name and credit card validation (CCV) number) has immediate value to most criminals.
- **Capture of address and other personal information** – any personal information, particularly address information, is a highly saleable and in constant demand by direct marketing companies.
- **Distribution of botnet and DDoS agents** – criminals use phishing scams to install special bot and DDoS agents on unsuspecting computers and add them to their distributed networks. These agents can be rented to other criminals.

Attack Propagation – Through a mixture of spear phishing and bot agent installations, phishers can use a single compromised host as an internal "jump point" within the organization for future attack. The proposed phishing website detection system will detect threats and indicate that e-mails, websites or the URL's are not secured and help the user avoid the hacker's trap. Such a type of detection builds confidence in both the users and the Internet community. The phishing website detection system will guide users by providing knowledge of Internet threats. In phishing detection, there are two types of techniques: the white list technique and the heuristic based mechanism. These two techniques act as filters in detecting phishing websites. In white list technique, a few anti-phishing websites are listed. If the user accessed websites are not in the white list, then these will be concluded as phishing websites. The heuristic based mechanism works with various aspects like keywords and domain name to decide whether the website is a phishing website or not [1].The rest of the paper is as follows: Section II discusses about the background, section III presents the design and implementation of the system, section IV describes the evaluation procedure and results and final conclusions are made in section V.

II. BACKGROUND AND RELATED WORK

A. Classification of Anti-Phishing Solutions

Phishing solutions can be broadly classified into five categories [11]. They are:

CANTINA:

A novel content-based approach for detecting phishing web sites. CANTINA takes Robust Hyperlinks, an idea for overcoming page not found problems using the well-known Term Frequency / Inverse Document Frequency (TF-IDF) algorithm, and applies it to anti-phishing. We described our implementation of CANTINA, and discussed some simple heuristics that can be applied to reduce false positives. We also presented an evaluation of CANTINA, showing that the pure TF-IDF approach can catch about 97% phishing sites with about 6% false positives, and after combining some simple heuristics we are able to catch about 90% of phishing sites with only 1% false positives [3].

PILFER:

We propose a new method for detecting these malicious emails called PILFER. By incorporating features specifically designed to highlight the deceptive methods used to fool users, we are able to accurately classify over 92% of phishing emails, while maintaining a false positive rate on the order of 0.1%. These results are obtained on a dataset of approximately 860 phishing emails and 6950 non-phishing emails. The accuracy of PILFER on this dataset is significantly better than that of Spam Assassin, a widely-used spam filter.[4].

Malicious Web site URLs:

An approach to this problem based on automated URL classification, using statistical methods to discover the tell-tale lexical and host-based properties of malicious Web site URLs. These methods are able to learn highly predictive models by extracting and automatically analyzing tens of thousands of features potentially indicative of suspicious URLs. The resulting classifiers obtain 95–99% accuracy, detecting large numbers of malicious Web sites from their URLs, with only modest false positives [5].

Page Rank:

This work uses the PageRank value and other features to classify phishing sites from normal sites. We have collected a dataset of 100 phishing sites and 100 legitimate sites for our use. By using this Google PageRank technique 98% of the sites are correctly classified, showing only 0.02 false positive rate and 0.02 false negative rate. [6].

Lexical Analysis

This paper presents a lexical URL analysis (LUA) technique to enhance the classification accuracy of anti-phishing email filters. Although the LUA feature is primarily focused to classify phishing websites, it proved to be effective to classify email messages due to the fact that most phishing email messages contain URLs. According to the performance evaluation, the LUA feature proved to be effective in enhancing the classifier's accuracy in all features subsets [7].

Detecting Webpage Source Code

We propose a phishing detection approach based on checking the webpage source code, we extract some phishing characteristics out of the W3C standards to evaluate the security of the websites, and check each character in the webpage source code, if we find a phishing character, and we will decrease from the initial secure weight. Finally we calculate the security percentage based on the final weight, the high percentage indicates secure website and others indicates the website is most likely to be a phishing website. We check two webpage source codes for legitimate and phishing websites and compare the security percentages between them, we find the phishing website is less security percentage than the legitimate website; our approach can detect the phishing website based on checking phishing characteristics in the webpage source code.[8]

Behavior based Detection:

A novel approach to detect phishing websites based on analysis of users' online behaviors – i.e., the websites users have visited, and the data users have submitted to those websites. Such user behaviors cannot be manipulated freely by attackers; detection based on those data can not only achieve high accuracy, but also is fundamentally resilient against changing deception methods [9].

III. Evaluation Procedure and Algorithm

Phase I: ***Blacklist:***

When user enters into the web browser and type a URL in web page. Check whether the site is phishing or not in the black listing. It is holding a phishing urls in the list .If any illegitimate site will appears, it will alert user web browser. Otherwise it goes to web parsing and heuristics terms.

Phase II: ***Scripting in the source code:***

A normal web user does not have knowledge whether a website is a malware. In the following steps are;

a) Web parsing:

Web parsing is a process in which every HTML code from the source of the web page is parsed. Tags such as <>, html, br, textbox, regular expressions, etc., will be eliminated in this method each and every HTML tag in the source of the webpage are parsed.

b) Separating the Required Tokens:

After parsing is done on the source of the webpage only the data and information other than the unwanted links and tags will be displayed. After parsing the web page, the required tokens are separated. A token could be a keyword, an operator, or a punctuation mark.

c) Classification of Scripting Tokens

If any external tokens are found while parsing, must be classified.These external tokens are created by hackers generally known as man-in-the-middle. Finally we text identification from the scripting and weight based find out phish site or legitimate site.

Phase III: ***Classification of Heuristics:*** In this phase classification a url by using heuristics based. We refer before, finally obtain a phishing or legitimate site. The contributions of this paper are: 1) to show how PageRank value can be useful to detect phishing. 2) An implementation to show high accuracy of classification of phished web sites. 3) Considering other features like age of the domain, suspicious URL, whether the domain contains IP address or not, number of dots and whether it is taking user personal information as input or not.

CONCLUSION

Thus, Phishing has become a major threat to information security and personal privacy. This paper represents new anti-phishing technique based on URL domain identity and scripting mechanism. It first identifies the related authorized URL. We used approximate classification algorithm. Two techniques i.e. URL domain identity and scripting are combined, so this proposed work performs better than other existing tools. This will reduce latency period of detection of phishing URLs.

REFERENCES

[1] Checking the Security of a Website Using Phishing Website Detector, The Department of Computing Sciences ,Texas A&M University-Corpus Christi Corpus Christi, TX.

[2] The Phishing Guide Understanding & Preventing Phishing Attacks,By: Gunter Ollmann, Director of Security Strategy,IBM Internet Security Systems

[3] Y. Zhang, J. Hong, and L. Cranor. "CANTINA: A Content-Based Approach to Detecting Phishing Web Sites". In *Proceedings of theInternational World Wide Web Conference (WWW)*, Banff, Alberta,Canada, May 2007.

[4] I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," Proceedings of the 16th international conference on World Wide Web, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 649–656. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242660.

[5] Justin Ma,Lawrence K. Saul, Stefan Savage, Geoffrey M. Voelker"Beyond Blacklists: Learning to Detecting Malicious websites from suspicious URL"Department of CSE,University of California,

[6] Anjali Sardana and A.Naga Venkata Sunil, IIT Roorkee ,Roorkee, India"A PageRank Based Detection Technique for Phishing Web Sites",2012 symposium IEEE

[7] Mahmoud Khonji and Youssef Iraqi, Andrew Jones, Computer Engineering, Khalifa University,Sharjah, UAE." "Lexical URL Analysis for Discriminating Phishing and Legit. imate E-Mail Messages",6th international conference on Internet technology and secured transactions, UAE

[8] Mona Ghotaish Alkhozae and Omar Abdullah Batarfi," "Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code**"** Volume 1 International Journal of Information and Communication Technology Research

[9] Xun Dong and John A. Clark, Jeremy L. Jacob *"User Behaviour Based Phishing Websites Detection"* Proceedings of the International Multiconference , Computer Science and Information Technology pp