International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 6 Issue: 4 214 - 219

A Review on Malicious URL Detection using Machine Learning Systems

Dipali K. Karnase
Department of Computer Sci. & Engg.
P. R. Pote College of Engg. & Mng., Amravati

Megha G. Mishra
Department of Computer Sci. & Engg.
P. R. Pote College of Engg. & Mng., Amravati

Snehal H. Dighole
Department of Computer Sci. & Engg.
P. R. Pote College of Engg. & Mng., Amravati

Snehal R. Shelke
Department of Computer Sci. & Engg.
P. R. Pote College of Engg. & Mng., Amravati

Mr. D. C. Dhanwani
Asst. Prof.,
Department of Computer Sci. & Engg.
P. R. Pote College of Engg. & Mng., Amravati

Abstract – Malicious web sites pretendsignificant danger to desktop security and privacy.These links become instrumental in giving partial or full system control to the attackers. This results in victim systems, which get easily infected and, attackers can utilize systems for various cyber-crimes such as stealing credentials, spamming, phishing, denial-of-service and many more such attack. Detection of such website is difficult because of thephishing campaigns and the efforts to avoid blacklists.To look for malicious URLs, the first step is usually to gather URLs that are liveon the Internet. There are various stages to detect this URLs such as collection of dataset, extracting feature using different feature extraction techniques and Classification of extracted feature. This paper focus on comparative analysis of malicious URL detection techniques.

Keywords – *Malicious web sites, credentials, phishing.*

*****

I. INTRODUCTION

Malicious URL consist of some malware or spyware that are harmful to the system. This malware can enter to the user system unknowingly and can also still some legal information from the system. Machine learning is field of computer science that uses statistical techniques to give computer science the ability to "learn" with data, without being explicitly programmed. Machine learning is classified into three types i.e supervised machine learning, unsupervised machine learning and semi supervised learning.

A Phishing is an attempt by an individual or a group to steal personal confidential information such as passwords, credit card information from unsuspecting victims for identity theft, financial gain and other fraudulent activities. In the current scenario, when the end user wants to access his confidential information online (in the form of money transfer or payment gateway) by logging into his bank account or secure mail account, the person enters information like username, password, credit card no. etc. on the login page. But quite often, this information can be captured by attackers using phishing techniques (for instance, a phishing website can collect the login information the user enters and redirect him to the original site). There is no such information that cannot be directly obtained from the user at the time of his login input.

Phishing web page as "any web page that, without permission, alleges to act on behalf of a third party with the intention of confusing viewers into performing an action with which the viewers would only trust a true agent of the third party." This definition, which is similar to the definition of "web forgery", covers a wide range of phishing pages from typical ones – displaying graphics relating to a financial company and requesting a viewer's personal credentials – to sites which claim to be able to perform actions through a third party once provided with the viewer's login credentials. Thus, a phishing URL is a URL that leads user to a phishing web page. Our study, by this definition, is therefore independent of the attack vector by which a phishing URL is distributed.

Phishing is a generally new internet crime in correlation with different forms such as hacking and virus attacks. A phishing site as demonstrated in Figure 1 is an extensively dispatched social engineering attack that endeavors to cheat individuals of their own data including Visa number, bank account data, standardized savings number, and their own certifications with a specific end goal to utilize these points of interest falsely against them. Phishing has a tremendous negative effect on associations' incomes, client connections, advertising endeavors, and general corporate picture. Phishing attacks can cost organizations keep an eye on a huge number of money per attack in fraud-related misfortunes and personnel

time. Far more terrible, expenses connected with the degradation of brand image and consumer confidence can keep running into a huge number of dollars

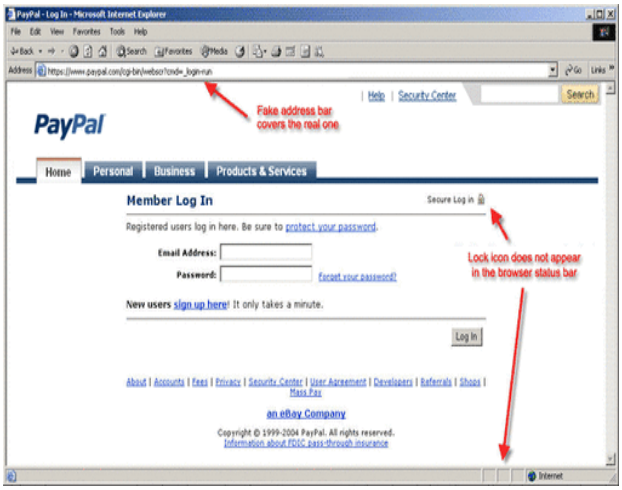

Figure 1: Screenshot of a phishing website

There are many definitions of phishing website; that wants to be very careful how they define the term, since it is constantly evolving. One of these definitions comes according to the Anti-Phishing Working Group (APWG)'s definition (APWG, 2005), "Phishing attacks use both social engineering and technical subterfuge to steal consumers' personal identity data and financial account credentials". Typically, a phishing attack is a combination of fraudulent emails, spoofed websites, and identity theft. Internet users or customers of many banks and financial institutions are the targets of phishing attacks. Nevertheless, there are lots of definitions of a phishing website from different perspectives. Hereunder they mention some of these definitions to get better understanding of its features and attack tactics.

Phishing web pages are forged web pages that are created by malicious people to mimic Web pages of real web sites. Most of these kinds of web pages have high visual similarities to scam their victims. Some of these kinds of web pages look exactly like the real ones. Victims of phishing web pages may expose their bank account, password, credit card number, or other important information to the phishing web page owners. It includes techniques such as tricking customers through email and spam messages, man in the middle attacks, installation of key loggers and screen captures. These popular technologies have several drawbacks:

- Blacklist-based technique with low falsealarm probability, but it cannot detect the websites that are not in the blacklist database. Because the life cycle of phishing websites is too short and the establishment of blacklist has a long lag time, the accuracy of blacklist is not too high.
- Heuristic basedantiphishingtechnique, with a high probability of false and failed alarm, and it is easy for the attacker to use technical means to avoid the heuristic characteristics detection.
- Similarity assessment based technique is time-consuming. It needs too long time tocalculate a pair of pages, so using the method to detect phishing websites on the client terminal is not suitable. And there is low accuracy rate for this method depends on many factors, such as the text, images, and similarity measurement technique. However, this technique (in particular, image similarity identification technique) is not perfect enough yet.

II. LITERATURE REVIEW

Dhamija and Tygar's (2005) approach involves the use of a so-called dynamic security skin on the user's browser [1]. This technique uses a shared secret image that allows a remote server to prove its identity to a user in a way that supports easy verification by humans but which is difficult for the phishers to spoof. The disadvantage of this approach is that it requires effort by the user. That is, the user needs to be aware of the phishing threat and check for signs that the site he/she is visiting is being spoofed. This approach requires changes to the entire web infrastructure (both servers and clients), so it can succeed only if the entire industry supports it. Also this technique does not provide security for situations where the user login is from a public terminal.

Dhamija et al. (2006) analyzed 200 phishing attacks from the AntiPhishing Work Group database and identified several factors, ranging from pure lack of computer system knowledge, to visual deception tricks used by adversaries, due to which users fall for phishing attacks [2]. They further conducted a usability study with 22 participants. The participants were asked to study 20 different websites to see if they could tell whether they were fraudulent or authentic. The result showed that age, sex and computer habits didn't make much difference. They even noticed that pop-up warnings of invalid signature of the sites and visual signs of SSL (Secure Sockets Layer), padlocks etc. were very inefficient and were overlooked. They found that 23% of the participants failed to look at security indicators warning about phishing attacks and, as a result, 40% of the time they were susceptible to a phishing attack. Based on their analysis, the authors suggest that it is important to re-think the design of security systems, particularly by taking usability issues into consideration.

Wu et al. (2006) proposed methods that require web page creators to follow certain rules to create web pages, by adding sensitive information location attributes to HTML code [3].

However, it is difficult to persuade all web page creators to follow the rules.

Liu et al. (2005) analyzed and compared legitimate and phishing web pages to define metrics that can be used to detect a phishing page on visual similarity (i.e. block level similarity, layout similarity and overall style similarity) [4].

The DOM -based (Wood, 2005) visual similarity of web pages is oriented, and the concept of visual approach to phishing detection was first introduced [5]. Through this approach, a phishing web page can be detected and reported in an automatic way rather than involving too many human efforts. Their method first decomposes the web pages (in HTML) into salient (visually distinguishable) block regions. The visual similarity between two web pages is then evaluated in three metrics: block level similarity, layout similarity, and overall style similarity, which are based on the matching of the salient block regions. A web page is classified as a phishing page if its visual similarity value is above a predefined threshold.

Fu, et al. (2006) proposed a phishing web page detection method using the EMD-based visual similarity assessment [6]. This approach works at the pixel level of web pages rather than at the text level, which can detect phishing web pages only if they are "visually similar" to the protected ones without considering the similarity of the source codes. The phishing filter in IE8 is a toolbar approach with more features such as blocking the user's activity on a detected phishing site. The most popular and widely-deployed techniques, however, are based on the use of blacklists of phishing domains that the browser refuses to visit. For example, Microsoft has recently integrated a blacklist based anti-phishing solution into its Internet Explorer (IE8). The browser queries lists of blacklisted and whitelisted domains from Microsoft servers and makes sure that the user is not accessing any phishing sites. Microsoft's solution is also known to use some heuristics to detect phishing symptoms in web pages (Sharif, 2005). Obviously, to date, the company has not released any detailed public information on how its anti-phishing techniques function [7].

Chandrasekaran et al. (2006) proposed anapproach to classify phishing based on phishing emails' structural properties.25 features, comprising style markers (e.g. the words suspended, account, and security) and structural attributes, such as the structure of the subject line of the email and the structure of the greeting in the body, were used in the study. 200 emails (100 phishing and 100 legitimate) were tested. Simulated annealing was applied as an algorithm for feature selection. After a feature set was chosen, information gain (IG) was used to rank these features based on their relevance. Thus, they applied one-class SVM to classify phishing emails based on the selected features. The results demonstrated a detection rate of 95% of phishing emails with a low [8].

Fette et al. (2007) compared a number of commonly used learning methods through their performance in phishing detection on a past phishing data set, and finally Random Forests were implemented in their algorithm PILFER. The authors claim that the methods can be used in the detection of phishing websites as well. 860 phishing emails and 6950 legitimate emails were tested. The proposed method correctly detected 96% of the phishing emails with a false positive rate of 0.1%. Ten handpicked features were selected for training using a phishing dataset that was collected in 2002 and 2003. As pointed out by the authors themselves, their implementation is not optimal and further work in this area is warranted [9].

Abu-Nimeh et al. (2007) compared six machine learning techniques to classify phishing emails. Their phishing corpus consisted of a total of 2889 emails and they used 43 features (variables). They used a bag-of-words as their feature set and the results demonstrated that merely using a spam detection mechanism, i.e. bag-of-words only, achieves high predictive accuracy. However, relying on textual features results in high false positive rates, as phishing emails are very similar to legitimate ones. The studied classifiers could successfully predict more than 92% of the phishing emails [10].

Pan and Ding (2006) examined the anomalies in web pages, in particular, the discrepancy between a web site's identity and its structural features and HTTP transactions [11].

Herzberg and Gbara (2004) proposed a solution to combine the technique of standard certificates with a visual indication of correct certification; a site-dependent logo indicating that the certificate was valid would be displayed in a trusted credentials area of the browser [12]. Another approach detects certain common attack instances, such as attacks in which the images are supplied from one domain while the text resides with another domain, and attacks corresponding to misspellings of URLs of common targets. "The Phishing Guide" by Ollmann (2004) gives a detailed understanding of the different techniques often included in phishing attacks [13]. The phenomenon that started as simple emails persuading the receiver to reply with the information the attacker required has evolved into more advanced ways to deceive the victim. Links in email and false advertisements sends the victim to more and more advanced fraudulent websites designed to persuade the victim to type in the information the attacker wants, for example to log into the fraudulent site mimicking the company's original. Ollmann also presents different ways to check whether websites are fraudulent or not. Apart from inspecting whether the visited site really is secure through SSL (Secure Sockets Layer), the user should also check that the certificate added to the website really is from the company it claims to be from and that it is signed by

a trusted third party. Focusing more attention on the URL can also often reveal fraudulent sites. There are a number of ways for the attackers to manipulate the URL to look like the original, and if the users are aware of this they can more easily check the authentication of the visited site.

Watson et al. (2005) describe in their White Paper, "Know your enemy: Phishing", different real-world phishing attacks collected in German and United Kingdom honeynets [14]. Honeynets are open computer networks designed to collect information about different attacks out in the real world, for further forensic analysis. They noticed that phishing attacks using vulnerable web servers as hosts for predesigned phishing sites are by far the most common, compared to using self-compiled servers. A compromised server is often host for several different phishing sites. These sites are often only active for a few hours or days after being downloaded to the server.

Garera et al. [15] focus on studying the structure of URLs employed in various phishing attacks. They find that it is often possible to tell whether or not a URL belongs to a phishing attack without requiring any knowledge of the corresponding page data. It describe several features that can be used to distinguish a phishing URL from a benign one. These features are used to model a logistic regression filter that is efficient and has a high accuracy. use filter to perform thorough measurements on several million URLs and quantify the prevalence of phishing on the Internet today [15].

Ma et al. [16] propose a method to classify malicious URLs using variable number of lexical and Hostbased properties of the URLs. They describe an approach for 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 [16].

Whittaker et al. [17] describe the design and performance characteristics of a scalable machine learning classifier that has been used in maintaining Google's phishing blacklist automatically. Their proprietary classifier analyzes millions of pages a day, examining the URL and the contents of a page to determine whether or not a page is phishing. Their system classifies web pages submitted by end users and URLs collected from Gmail's spam filters. Though some URL based features are similar, they propose several new features and evaluate our approach with publicly available machine learning algorithms and public data sets. Unlike their approach, they do not use any proprietary and page content based features.

Zhang et al. [18] present CANTINA, content-based approach to detect phishing websites, based on the TF-IDF information retrieval algorithm and the Robust Hyperlinks algorithm. By using a weighted sum of 8 features (4 content related, 3 lexical, and 1 WHOIS-related) they show that CANTINA can correctly detect approximately 95% of phishing sites. The goal of their approach is to avoid downloading the actual web pages and thus reduce the potential risk of analyzing the malicious content on user's system. In order to achieve this goal, they evaluate only the features related to URLs. A number of machine learning-based studies can be found in related contexts such as in detecting phishing emails.

Fette et al. [19] use a set of 10 features extracted from email headers, WHOIS information on sender's domain, email contents, URL structures, etc. and apply Support Vector Machines (SVMs) to classify phishing emails from legitimate ham emails. We further improve the accuracy of Fette et al. by introducing groups of keyword based features from the email contents [20]. Using different classification modelsthey achieve classification accuracy of 98%, while maintaining low false positive and negative rates.

Fette et al. [19] hypothesized that phishing email classification appears to be simple text classification problem but, the classification is confounded by the fact that the class of "phishing" emails is nearly identical to the class of real emails. Motivated by the hypothesis, base the phishing email classification problem as the text classification problem in previous work [21]. Using Confidence Weighted linear classifier, an online algorithm, and using only the email text contents as "bag-of-words" representation, they achieve a classification accuracy of 99%, maintaining false positive and false negative rates of less than 1% on public benchmark data sets. Besides machine learning (ML) based techniques, there exist many other approaches in phishing detection. Perhaps, the most widely used antiphishing technology is the URL blacklist technique that most modern browsers are equipped with [22] and [23]. Other popular methods are browser based plug-in or add-in toolbars.

SpoofGuard [24] uses domain name, URL, link, and images to evaluate the spoof probability on a webpage. The plug-in applies a series of tests, each resulting in a number in the range from 0 to 1. The total score is a weighted average of the individual test results. There has been an attempt to detect phishing attack using user generated rules [25]. Other anti-phishing tools include SpoofStick [26], SiteAdvisor [27], Netcraft antiphishing toolbar [28], AVG Security Toolbar [29], etc.

Table 1: comparative study of different techniques of malicious URLs detection

Sr no.	Authors	Paper Title	Technique used	conclusion
1	Rachna Dhamija, J.D. Tygar	The battle against phishing: Dynamic security skins	Secret image sharing using remote server	Easy to verify the remote server identity. Difficult to for phisher spoof.
2	Min Wu	Fighting Phishing at the User Interface	HTML code location attributes	It is difficult to persuade all web page creator to follow the rules.
3	Madhusudhanan chandrasheran, Krishnan narayanan.	Phishing E-mail detection based on structural properties.	25 features, comprising style markers, structural attributes, Apply SVM classifier	Give 95% accurate result.
4	Sujata Garera, Niels Provos, Monica Chew	A Framework for Detection and Measurement of Phishing Attacks	Extraction based on URL structure	logistic regression filter.
5	J. Ma, L.K. Saul, S. Savage, G.M. Voelker	Beyond blacklists: Learning to detect malicious web sites from suspicious URLs	Used Lexical and Host based properties of URLs	Highly predictive with 95.99% accuracy
6	C. Whittaker, B. Ryner, M. Nazif,	Large-scale automatic classification of phishing pages	Used scalable machine learning classifier	Contents of page to determine phishing websites
7	Y. Zhang, J. Hong, L. Cranor	CANTINA: a content based approach to detecting phishing web sites	8 features are used 4 content related, 3 lexical and 1 WHOIS	Reduce the potential risk of analyzation.

III. CONCLUSION

we have studied different techniques of malicious URL detection. From this survey we analyzed that the various techniques used in different existing system such as, web page content feature, Host based feature and feature extraction based on URL structure and some system used hybrid feature by combine different feature. So hybrid feature are more effective as compare to distributed features. machine learning approach can easily train the system and gives possible true positive result.

REFERENCES

[1]. S. Dhamija, R., and Tygar, J., "The battle against phishing: Dynamic security skins", *In Proc. ACM Symposium on Usable Security and Privacy (SOUPS 2005)*, pp. 77–88, 2005.

[2]. Dhamija, R., Tygar, J., and Marti, H. "Why phishing works", *In CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, ACM Press*, pp. 581-590, New York, NY, USA, 2006.

[3]. Wu, M., "Fighting Phishing at the User Interface", *PhD Thesis in Computer Science and Engineering*, 2006.

[4]. Liu, W., Guanglin, H., Liu, X., Xiaotie, D. and Zhang, M. "Phishing Webpage Detection", *Proceedings of the 2005 Eight InternationalConference on Document Analysis and Recognition (ICDAR'05)*, pp. 560-564, 2005.

[5]. *http://www.w3.org. 2005.*

[6]. Fu, A., Wenyin, L., and Deng, X. "Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD)", *IEEE transactions on dependable and secure computing*, Vol. 3, No. 4, pp. 301-311, 2006.

[7]. *http://blogs.msdn.com/ie/archive/2005/09/09/463204.aspx*

[8]. Chandrasekaran, M., Narayanan, K., and Upadhyaya, S., "Phishing email detection based on structural properties", *Proceedings of the NYS Cyber Security Conference*, 2006.

[9]. Fette, I., Sadeh, N., and Tomasic, A., "Learning to detect phishing emails", In *WWW'07: Proceedings of the 16th international conference on World Wide Web*, pp. 649–656, New York, NY, USA, ACM Press, 2007.

[10]. Abu-Nimeh, S., Nappa, D., Wang, X., and Nair, S. "A comparison of machine learning techniques for phishing detection", *In eCrime'07: Proceedings of the antiphishing working groups 2nd annual eCrime researchers summit*, pp. 60–69, New York, NY, USA, ACM, 2007.

[11]. Pan, Y., and Ding, X. "Anomaly Based Web Phishing Page Detection", *Proceedings of the 22nd Annual Computer Security Applications Conference (ACSAC'06)*, pp.381-392, 2006.

[12]. Herzberg, A., and Gbara, A. "Protecting Naive Web Users", Draft of July 18, 2004.

[13]. Ollmann, G., "The Phishing Guide, Understanding and Preventing Phishing Attacks", Online Available: http://www.nextgenss.com/papers/NISR-WPPhishing.pdf 2004.

[14]. Watson, D., Holz, T., and Mueller, S. "Know your enemy: Phishing, behind the scenes of Phishing attacks", The Honeynet Project & Research Alliance, 2005.

[15]. S. Garera, N. Provos, M. Chew, A.D. Rubin, "A framework for detection and measurement of phishing attacks", In: Proc. 5th ACM Workshop on Recurring Malcode, WORM‟07, ACM, New York, NY, USA, 2007, pp. 1-8.

[16]. J. Ma, L.K. Saul, S. Savage, G.M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs", In: Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Paris, France, 2009, pp. 1245-1254.

[17]. C. Whittaker, B. Ryner, M. Nazif, "Large-scale automatic classification of phishing pages", In: Proc. 17th Annual Network and Distributed System Security Symposium, NDSS‟10, San Diego, CA, USA, 2010.

[18]. Y. Zhang, J. Hong, L. Cranor, "CANTINA: a content based approach to detecting phishing web sites", In Proc. 16th Int. Conf. World Wide Web, WWW‟07, Banff, Alberta, Canada, 2007, pp. 639-648.

[19]. I. Fette, N. Sadeh, A. Tomasic, "Learning to detect phishing emails", In: Proc. Int. Conf. World Wide Web, WWW‟07, Banff, Alberta, Canada, 2007, pp. 649-656.

[20]. R.B. Basnet, S. Mukkamala, A.H. Sung, "Detection of phishing attacks: a machine learning approach", In: Bhanu Prasad (Ed.), Studies in Fuzziness and Soft Computing, Springer, 2008, pp. 373-383.

[21]. R. B. Basnet, A.H. Sung, "Classifying phishing emails using confidence-weighted linear classifiers", In: Proc. Int. Conf. Information Security and Artificial Intelligence, ISAI,,10, Chengdu, China, 2010, pp. 108- 112.

[22]. Google Safe Browsing API - Google Code,http://code.google.com/apis/safebrowsing/

[23]. SmartScreen Filter – Microsoft Windows, Online available at: http://windows.microsoft.com/enUS/internetexplorer/products/ie-9/features/smartscreenfilter, 2011.

[24]. N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, J. Mitchell, "Client-side defense against web-based identity theft", In: Proc. 11th Network and Distributed System Security Symposium, NDSS‟04, San Diego, CA, USA, 2004.

[25]. R. B. Basnet, A.H. Sung, Q. Liu, Rule-based phishing attack detection, In: Proc. Int. Conf. Security and Management, SAM‟11, Las Vegas, NV, USA, 2011.

[26]. SpoofStick Home, Online available at: http://www.spoofstick.com

[27]. McAfee Site Advisor Software – Website Safety Ratings and Secure Search, Online available athttp://www.siteadvisor.com

[28]. Netcraft Anti-Phishing Toolbar, Online available at: http://toolbar.netcraft.com

[29]. AVG Security Toolbar, Online available at: http://www.avg.com/product-avg-toolbar-tlbrc#tba2