A Review on Techniques Detection of Phishing URL

Asha Choudhary[1], Rakesh Rathi[2]
[1]*Student, Department CS & IT, Govt. Engg. College, Ajmer, India*
[2]*Professor and HoD, Department CS & IT, Govt. Engg. College, Ajmer, India*

***Abstract*: URL phishing is becoming more popular and more dangerous as more and more platforms are coming online, be it shopping over various portals, marketing for cause or business or for many other purposes. In this paper we have mentioned studies of various paper about the tested mechanisms, techniques to detect the fake websites and current proposed work which are enforced in the industries to protect such practices.**

***Keywords*: URLs, Phishing attacks, Trojans, Text Data, Image Data, Audio Data, Multimedia Data.**

1. Introduction

Phishing is a cybercrime in which a target or targets are contacted by email, telephone or text message by someone posing as a legitimate institution to lure individuals into providing sensitive data such as personally identifiable information, banking and credit card details, and passwords.

The information is then used to access important accounts and can result in identity theft and financial loss. There are two types of phishing as described in this section:

- *Hand over sensitive information*: These messages aim to trick the user into revealing important data — often a username and password that the attacker can use to breach a system or account. The classic version of this scam involves sending out an email tailored to look like a message from a major bank; by spamming out the message to millions of people, the attackers ensure that at least some of the recipients will be customers of that bank. The victim clicks on a link in the message and is taken to a malicious site designed to resemble the bank's webpage, and then hopefully enters their username and password. The attacker can now access the victim's account.
- *Download malware*: Like a lot of spam, these types of phishing emails aim to get the victim to infect their own computer with malware. Often the messages are "soft targeted" — they might be sent to an HR staffer with an attachment that purports to be a job seeker's resume, for instance. These attachments are often .zip files, or Microsoft Office documents with malicious embedded code. The most common form of malicious code is ransomware — last year it was estimated that 93 percent of phishing emails contained ransomware attachments.

There are few proactive measures too for protecting a machine or user from these phishing attacks:

- "Sandboxing" inbound email, checking the safety of each link a user click.
- Inspecting and analyzing web traffic.
- Pen-testing your organization to find weak spots and use the results to educate machine and user.
- Marking abilities of machine and a user, perhaps by showcasing a "catch of the day" if someone spots a phishing email.

2. Literature review

Fonseca et al. [1] have proposed a methodology where web application security mechanism is tested before deployment. Vulnerabilities are injected into the web applications, and their effect is reported. This way before deploying the system itself the system gets tested in real life scenarios and the vulnerable points can be fixed. A prototype has been built to automate the process of injection attack, analyzing the effect on the application and then publishing the results.

Ahmed Abbasi, et aI., [2] have researched on the topic of detecting fake medical websites. *The techniques used for detecting fake websites are graph-based classifiers and recursive trust labeling. They have analyzed the different features of the websites which will be able effectively to distinguish fake and genuine websites in the medical domain.*

Fu, Wenyin et a1.[3] proposed detection of phishing websites using similarity *detection by using Earth Mover's Distance(EMD).*The websites that are suspected to be fake are those, whose URLs are present in apparent phishing e-mails. The website is converted into an image and its various properties are extracted for the purpose of classification.

In their paper, Zhang, Liu, et al. [4] talk *about two classifiers and an algorithm to fuse the result of both of them. One is a text classifier which uses naIve Bayes rule to perform classification, and the other is an image classifier which uses Earth Mover's distance.* The algorithm that fuses the image and visual classifier uses Bayes theorem. A Bayesian approach is used to calculate the threshold of both the classifiers through offline training.

Chen, Dick, et al. [5], have applied a different theory of visual similarity between web pages. They have used the *Gestalt Theory*, which is a contemporary theory in the field of philosophy and psychology.

Xiang, Hong [6] have *worked on a feature rich machine learning framework named CANTINA+*. They concentrate on the two important features of a phishing scam. 1) A website that is a mirror image of a financial institution's site. 2) A fake login page asking for sensitive information like password, credit card number, etc. on behalf of the financial institution.

From these works, it is conspicuous that for the purpose of classification, the websites have been deconstructed. For each web page that has been considered, the corresponding features have been extracted and used for classification.

However, an alternative methodology has been put forward by Justin Ma et al. [7] which argue that without extracting the web page contents, just by looking at the URL structure alone, its authenticity can be ascertained. The *URL itself has certain features, using which the website can be classified.* By using this approach, we can avoid the malware and Trojans that would otherwise be encountered when we visit the URL without knowing its authenticity. *Lexical features considered were domain name, URL size, dots present in the URL, etc.* Ma et al. [7] have used a bag-of-words concept where every term after a delimiter is picked up and added as a feature. But we perceive that this method would increase the feature set size drastically and affect the classification time.

Blacklisting and machine learning based solutions are used mainly to detect phishing URLs till date. In this section we describe briefly the related works of such solutions.

Using blacklisting when a URL is requested, it does a pattern matching to know whether the given URL is present in their repository [8]. If it is present the request will be blocked. This has been employed in web browsers such as Phish Tank, DNS-BH and jwSpamSpy and commercial malicious URL detection systems such as Google Safe Browsing, McAfee Site Advisor, Web of Trust (WOT), Websense Threat Seeker Network, Cisco IronPort Web Reputation and Trend Micro Web Reputation Query Online System. Blacklisting mechanisms are very simple and easy-to-implement. It shows higher detection rate but fails to detect newly formed phishing URLs and it requires human feedback to update its database.

Machine learning methods rely on feature engineering to extract lexical, host-based features and a hybrid of both to distinguish between the benign and malicious URLs.

In the paper [9], *they use various machine learning classifiers with URL based features to classify the URLs as either malicious or benign.* For feature engineering, they used recursive entropy reduction-based mechanism to obtain tokens and extracted a set of features from the collected tokens. Their work inferred that the URL based features on comparison with the page content features performed relatively well.

Liang, Bin, et al. [10] *used lexical, header and time information as features to study hidden fraudulent characteristics of URL.* The detection rate of malicious URL was higher for their model based on the results they attained. They also made a statement that, the method will perform extremely well when a large sample of data about million samples of malicious and benign URLs is used in the training phase.

Garera et al. [11] proposed a method to detect phishing webpages using *lexical features* i.e. the structure of webpage URLs. They used the features like IP address, hostname length, obfuscating a host with another domain, domain misspelled, page rank of URL, host rank, page rank present in crawl database, white domain table to detect phishing attempts.

Basnet et al. [12] *used machine learning techniques such as Support Vector Machine (SVM), Biased Support Vector Machine (BSVM), Neural Network (NN) and Self-Organizing Map (SOMs) to develop their framework.* The set of features they used are HTML formatted e-mails, IP based URL, domain age, number of domains, number of sub-domains, token JavaScript, large number of links, Tag ¡Form¿, image source, matching domains, keywords.

The work [13] *compared artificial neural network (ANN) approach with static classifiers such as SVM, DT, NB and KNN to compare the efficiency of malicious web page detection by using the static feature sets from lexical in URL and page contents.* ANN approach gave the best performance by reporting highest accuracy of 95.08% in comparison to other static classifiers. Additionally, in their work they discussed in detail the importance of each feature towards identifying attacks and thereby reducing the false positive rate.

3. Conclusion

Security analysts throughout the world are constantly challenged by the phishing community as new and advanced methods are developed each day. In this evolving environment, it's every researcher's main responsibility to deceive a system that can tackle the situation. In this study, when we compare the different classification algorithms, we have identified the tree-based classifiers as best suitable for the task of phishing URL classification. As an extension of this work, we intend to enhance the system performance further by incorporating an online learning mode. This will further improve the accuracy and help to achieve better performance as the system becomes dynamic. This paper analyzed the performance of logistic regression using bigrams, CNN and CNN-LSTM models to detect phishing URLs. Deep learning methods like CNN and CNNLSTM are preferable over machine learning methods as they have the capability to obtain optimal feature representation themselves by taking the raw URLs as their input. We can claim based on the results we obtained that, the machine learning and deep learning based malicious URL detection can foreclose detection systems built using blacklisting and regular expression methods.

References

[1] Jose Fonseca, Marco Vieira, Henrique Madeira, " Evaluation of Web Security Mechanisms using Vulnerability & Attack Injection", IEEE

Transactions On Dependable And Secure Computing, Volume II, Issue 5, pp. 440-453, 2014.

[2] Ahmed Abbasi, Fatemeh Zahedi, Siddharth Kaza, " Detecting Fake Medical Web Sites Using Recursive Trust Labeling" , ACM Transactions on Information Systems, Volume 30, Issue 40, Article 22, pp. 22.1-22.36, 2012.

[3] Anthony Y. Fu, Liu Wenyin and Xiaotie Deng, "Detecting Phishing Web Pages with Visual Similarity Assessment based on Earth Mover's Distance (EMD)", IEEE Transactions on Dependable and Secure Computing, Volume 3, No. 4,pp. 301-311, 2006

[4] Haijun Zhang, Gang Liu, Tommy W. S. Chow and Wenyin Liu, "Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach", IEEE Transactions on Neural Networks, Volume 22, No. 10, pp. 1532-1546, 2011.

[5] Teh-Chung Chen, Scott Dick and James Miller, "Detecting Visually Similar Web Pages: Application to Phishing Detection", ACM Transactions on Internet Technology, Volume 10, No. 2, pp. 5.1-5.38, 2010.

[6] Guang Xiang, Jason Hong, Carolyn P.Rose and Lorrie Cranor, "CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites", ACM Transactions on Information and System Security, Volume 14, No. 2, Article 21, pp. 21: 1-21 :28, 2011.

[7] Justin Ma, Lawrence K. Saul, Stefan Savage And Geoffrey M. Voelker, "Learning to Detect Malicious URLs", ACM Transactions on Intelligent Systems and Technology, Volume 2, No. 3, Article 30, pp. 30.1-30.24,2011.

[8] A. K. Jain and B. Gupta, “A novel approach to protect against phishing attacks at client side using auto-updated white-list,” EURASIP Journal on Information Security, vol. 2016, no. 1, p. 9, 2016.

[9] M.-Y. Kan and H. O. N. Thi, “Fast webpage classification using url features,” in Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005, pp. 325–326.

[10] B. Liang, J. Huang, F. Liu, D. Wang, D. Dong, and Z. Liang, “Malicious web pages detection based on abnormal visibility recognition,” in EBusiness and Information System Security, 2009. EBISS’09. International Conference on. IEEE, 2009, pp. 1–5.

[11] S. Garera, N. Provos, M. Chew, and A. D. Rubin, “A framework for detection and measurement of phishing attacks,” in Proceedings of the 2007 ACM workshop on Recurring malcode. ACM, 2007, pp. 1–8.

[12] R. B. Basnet, S. Mukkamala, and A. H. Sung, “Detection of phishing attacks: A machine learning approach.” Soft Computing Applications in Industry, vol. 226, pp. 373–383, 2008.

[13] A. Sirageldin, B. B. Baharudin, and L. T. Jung, “Malicious web page detection: A machine learning approach,” in Advances in Computer Science and its Applications. Springer, 2014, pp. 217–224.