Literature Survey on Web Phishing Detection

ABSTRACT

This article surveys the literature on the detection of phishing attacks. Phishing attacks target vulnerabilities that exist in systems due to the human factor. Many cyber attacks are spread via mechanisms that exploit weaknesses found in end-users, which makes users the weakest element in the security chain. The phishing problem is broad and no single silver-bullet solution exists to mitigate all the vulnerabilities effectively, thus multiple techniques are often implemented to mitigate specific attacks. This paper aims at surveying many of the recently proposed phishing mitigation techniques. A high-level overview of various categories of phishing mitigation techniques is also presented, such as: detection, offensive defense, correction, and prevention, which we belief is critical to present where the phishing detection techniques fit in the overall mitigation process.

I.INTRODUCTION

Phishing is a type of social engineering attack often used to steal user data, including login credentials and credit card numbers. It occurs when an attacker, masquerading as a trusted entity, dupes a victim into opening an email, instant message, or text message. The recipient is then tricked into clicking a malicious link, which can lead to the installation of malware, the freezing of the system as part of a ransomware attack or the revealing of sensitive information.

An attack can have devastating results. For individuals, this includes unauthorized purchases, the stealing of funds, or identify theft.

Moreover, phishing is often used to gain a foothold in corporate or governmental networks as a part of a larger attack, such as an advanced persistent threat (APT) event. In this latter scenario, employees

are compromised in order to bypass security perimeters, distribute malware inside a closed environment, or gain privileged access to secured data.

An organization succumbing to such an attack typically sustains severe financial losses in addition to declining market share, reputation, and consumer trust. Depending on scope, a phishing attempt might escalate into a security incident from which a business will have a difficult time recovering.

II. LITERATURE SURVEY

This section of the literature survey eventually reveals some facts based on thoughtful analysis of many authors' work as follows:

In [1] JIAN MAO, WENQIAN TIAN and ZHENKAI LIANG has proposed a system which detect the phishing using page component similarity which analyzes URL tokens to increase prediction accuracy phishing pages typically keep its CSS style similar to their target pages. Based on the observation, a straightforward approach to detect phishing pages is to compare all CSS rules of two web pages, It prototyped Phishing-Alarm as an extension to the Google Chrome browser and demonstrated its effectiveness in evaluation using real-world phishing samples.

ZOU FUTAI, PEI BEI and PAN LI [2] Uses Graph Mining technique for web Phishing Detection. It can detect some potential phishing which can't be detected by URL analysis. It utilize the visiting relation between user and website. To get dataset from the real traffic of a Large ISP. After anonymizing these data, they have cleansing dataset and each record includes eight fields: User node number (AD), User SRC IP(SRC-IP) access time (TS), Visiting URL (URL), Reference URL(REF), User Agent(UA), access server IP (DST-IP), User cookie (cookie). For a client user, he is assigned a unique AD but a variable IP selected from ISP own IP pool. Therefore, we build the visiting relation

graph with AD and URL, called AD-URL Graph and the Phishing website is detected through the Mutual behavior of the graph.

In [3] NICK WILLIAMS and SHUJUN LI proposed a system which analysis ACT-R cognitive behavior architecture model. Simulate the cognitive processes involved in judging the validity of a representative webpage based primarily around the characteristics of the HTTPS padlock security indicator. ACT-R possesses strong capabilities which map well onto the phishing use case and that further work to more fully represent the range of human security knowledge and behaviors in an ACT-R model could lead to improved insights into how best to combine technical and human defenses to reduce the risk to users from phishing attacks

XIN MEI CHOO, KANG LENG CHIEW and NADIANATRA MUSA [4] this system is based on utilizing support vector machine to perform the classification. This method will extract and form the feature set for a webpage. It uses a SVM machine as a classifier which has two phase training phase and testing phase during training phase it extracts feature set and while testing it predict the website is legitimate or a phishing

In [5] GIOVANNI ARMANO, SAMUEL MARCHAL and N.ASOKAN proposed a use of add on in the browser which is Real-Time Client-Side Phishing Prevention. It uses information extracted from website visited by the user to detect if it is a phish and warn the user. It also determines the target of the phish and offers to redirect the user there. A warning message is displayed in the foreground while the background displays the phishing webpage darkened by a black semi-transparent layer preventing interactions with the website.

TRUPATI KUMBHARE and SANTOSH CHOBE [6] have discussed various Association Rule Mining Algorithm. Association rule learning searches for relationships among variables. Various Association algorithm discussed are AIS algorithm, SETM algorithm,

Apriori algorithm, Aprioritid algorithm, Apriorihybrid algorithm, and FP-growth algorithm.

In [7] S.NEELAMEGAM and DR.E.RAMARAJ discussed various Classification Algorithm used in data mining. Data Classification is a data mining technique used to predict group membership for data instances Various Classification Algorithm discussed are decision tree, Bayesian networks, k-nearest neighbor classifier, Neural Network, Support vector machine.

VARSHARANI RAMDAS, V.Y. KULKARNI and R.A.RANE[8] proposed a system to detect a phishing website using Novel Algorithm This detection algorithm can find out the maximum number of phishing URLs because it executes multiple tests such as Blacklist search Test, Alexa ranking test, and different URL features test. But this solution is effective only for HTTP URLs.

In [9]JUN HU,YUCHUN JI and HANBING YAN-This method to detect Phishing website is based on the analysis of legitimate website server log information. every time a victim opens the phishing website, the phishing website will refer to the legal website by asking for resources. Then, there will be a log, which is recorded by the legitimate website server and from this logs Phishing site can be Detected.

SAMUEL NARCHAL, GIOVANNI ARMANO and NIDHI SINGH[10] propose a application Off-the-Hook application for detection of phishing website. Off-the-Hook, exhibits several notable properties including high accuracy, brand-independence and good language-independence, speed of decision, resilience to dynamic phish and resilience to evolution in phishing techniques.

[11] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. “Machine Learning-Based Phishing Detection from URLs,” Expert Systems with Applications, 117:345-357, January 2019:The dataset used is self-constructed. Where phishing websites belong to PhishTank and legitimate URLS are from Yandex Search API. The main purpose was

to detect the word which is similar to brand names, to detect keywords, the words, which are formed from random characters. Various classification algorithms such as Naive Bayes, Random Forest, kNN(n=3), Adaboost, K-star, SMO, and Decision Tree including some feature extraction types such as NLP-based features, Word Vectors and Hybrid are used. This system gets high accuracy throughout the test.

[12] J. James, Sandhya L. and C. Thomas, "Detection of phishing URLs using machine learning techniques," International Conference on Control Communication and Computing (ICCC), December 2013:The system proposed used a method based on lexical features, host properties, and properties related to the page for the detection of phishing websites. For getting a proper understanding of the pattern of URLs, various data mining algorithms are used. The classification algorithms such as Naïve Bayes, J48 Decision Tree, K-NN, and SVM were considered for the detection of phishing websites. Decision Tree had better accuracy of 91.08% compared to other algorithms. So Tree-based classifiers are best suited for phishing URL classification.

[13] Pradeepthi, K. V., & Kannan, A. "Performance study of classification techniques for phishing URL detection," Sixth International Conference on Advanced Computing (IcoAC), December 2014. : The system recognizes Phishing URLs, by examining the URL structure without attending the Phishing URL using classification algorithms. The data collected is first passed through the training state where it undergoes feature selection and classification. The dataset used here contains 4500 URL records, on which classification is performed. Out of which 2500 URLs are genuine and the other 2000 are the phishing ones. The 2500 URLs were collected from the DMOZ repository. The 2000 phishing URLs have been picked from PHISHTANK. Data classification after extraction of the relevant features was performed by Naive Bayes, Random Forest, Random Tree, Multi-layer Perceptron, C-RT, J 48 Tree, LMT, C 4.5, ID 3, and

K-Nearest Neighbour. The Random Forest Algorithm had the highest classification accuracy.

[14] Dipayan Sinha, Dr. Minal Moharir, Prof. Anitha Sandeep, "Phishing Website URL Detection using Machine Learning," International Journal of Advanced Science and Technology, 29(3):2495-2504, 2020. : Detection of phishing websites is performed by using machine learning techniques like Logistic Regression, Decision tree, Random Forest, Adaboost, Gradient Boosting, Gaussian NB, and Fuzzy pattern tree classifier. Data collection involves phishing and legitimate websites. Extracting useful features has two steps: URL-based involves IP Address, '@' symbol in URL, dashes in URL, long URL, presence of unusual number, dot count, sub-domains in URL, etc. Domain-based includes Page Rank of the website, age of the Domain, and Validity of the Website. The dataset is split into training and testing set in the ratio 80:20. The Random Forest algorithm shows 96% of precision and recalls along with the highest F1 score of 95%.

[15] R. Kiruthiga, D. Akila, "Phishing Websites Detection Using Machine Learning," International Journal of Recent Technology and Engineering (IJRTE), 8(2S11):11-114, September 2019:2019 A total of 15 research papers have been studied in this research paper. In this research, one method was discussed which uses five different algorithms that are Decision Tree, Generalized Linear Model, Gradient Boosting, Generalized Additive Model, and Random Forest. On comparing the results, the Random Forest algorithm had the highest accuracy of 98.4%, 98.59% recall, and precision of 97.70%. Dataset used is from the UCI machine learning repository

III.CONCLUSION

Education awareness is the most significant strategy to protect users from phishing attacks. Internet users should be aware of all security recommendations made by professionals. Every user should also be taught not to mindlessly follow links to websites where sensitive

information must be entered. Before visiting a website, make sure to check the URL. In the future, the system could be upgraded to automatically detect the web page and the application's compatibility with the web browser. Additional work can be done to distinguish fraudulent web pages from authentic web pages by adding certain additional characteristics.

IV.REFERENCES

[1] JIAN MAO1,WENQIAN TIAN1, PEI LI1, TAO WEI2, AND ZHENKAI LIANG3 Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity.

[2] Zou Futai, Gang Yuxiang, Pei Bei, Pan Li, Li Linsen Web Phishing Detection Based on Graph Mining.

[3] Nick Williams, Shujun Li Simulating human detection of phishing websites: An investigation into the applicability of ACT-R cognitive behaviour architecture model.

[4] XIN MEI CHOO, KANG LENG CHIEW,DAYANG HANANI ABANG IBRAHIM,NADIANATRA MUSA, SAN NAH SZE, WEI KING TIONG Feature-based Phishing Detection Technique.

[5] Giovanni Armano, Samuel Marchal and N. Asokan RealTime Client-Side Phishing Prevention Add-on.

[6] Trupti A. Kumbhare and Prof. Santosh V. Chobe An Overview of Association Rule Mining Algorithms.

[7] S.Neelamegam, Dr.E.Ramaraj Classification algorithm in Data mining: An Overview

[8] Varsharani Ramdas Hawanna, V. Y. Kulkarni and R. A. Rane A Novel Algorithm to Detect Phishing URLs.

[9] Jun Hu, Xiangzhu Zhang,Yuchun Ji, Hanbing Yan, Li Ding, Jia Li and Huiming Meng Detecting Phishing Websites Based on the Study of the Financial Industry Webserver Logs.

[10] Samuel Marchal, Giovanni Armano and Nidhi Singh Offthe-Hook: An Efficient and Usable.

[11] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. "Machine Learning-Based Phishing Detection from URLs," Expert Systems with Applications, vol. 117, pp. 345-357, January 2019.

[12] J. James, Sandhya L. and C. Thomas, "Detection of phishing URLs using machine learning techniques," International Conference on Control Communication and Computing (ICCC), December 2013.

[13] Pradeepthi, K. V., & Kannan, A. "Performance study of classification techniques for phishing URL detection," Sixth International Conference on Advanced Computing (IcoAC), December 2014.

[14] Dipayan Sinha, Dr. Minal Moharir, Prof. Anitha Sandeep, "Phishing Website URL Detection using Machine Learning," International Journal of Advanced Science and Technology, vol. 29, no. 3, pp. 2495-2504, 2020.

[15] R. Kiruthiga, D. Akila, "Phishing Websites Detection Using Machine Learning," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2S11, pp. 11-114, September 2019.