International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

PHISHCATCHER: CLIENT-SIDE DEFENSE AGAINST WEB SPOOFING ATTACKS USING MACHINE LEARNING

[1]MR.G. HARISH KUMAR, [2]B. SRIVANI, [3]B. NIKHITHA, [4]C. VARSHITHA

[1]Assistant Professor**,** Department of Electronics and Communication Engineering, Malla Reddy Engineering College For Women**,** Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

[2,3,4]Student, Department of Electronics and Communication Engineering, Malla Reddy Engineering College For Women**,** Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

ABSTRACT

Cyber security confronts a tremendous challenge of maintaining the confidentiality and integrity of user's private information such as password and PIN code. Billions of users are exposed daily to fake login pages requesting secret information. There are many ways to trick a user to visit a web page such as, phishing mails, tempting advertisements, click-jacking, malware, SQL injection, session hijacking, man-in-the-middle, denial of service and cross-site scripting attacks. Web spoofing or phishing is an electronic trick in which the attacker constructs a malicious copy of a legitimate web page and request users' private information such as password. To counter such exploits, researchers have proposed several security strategies but they face latency and accuracy issues. To overcome such issues, we propose and develop client-side defence mechanism based on machine learning techniques to detect spoofed web pages and protect users from phishing attacks. As a proof of concept, a Google Chrome extension dubbed as PhishCatcher, is developed that implements our machine learning algorithm that classifies a URL as suspicious or trustful. The algorithm takes four different types of web features as input and then random
forest classifier decides whether a login web page is spoofed or not. To assess the accuracy and precision of the extension, multiple experiments were carried on real web applications. The experimental results show remarkable accuracy of 98.5% and precision as 98.5% from the trials performed on 400 classified phished and 400 legitimate URLs. Furthermore, to measure the latency of our tool, we performed experiments over forty phished URLs. The average recorded response time of PhishCatcher was just 62.5 milliseconds.

I. INTRODUCTION

In Oct 2022,1 the members/users of the National Institute for Research in Digital Science and Technology (Inria) in France received an email in French asking the users to confirm their webmail account with the direct link https://www.educationonline. nl/Cliquez.ici.cas.inria.fr.cas.login/login.htm l. When clicked on this link, it takes to a fake but appearing genuine central authentication login page of Inria. As this fake login page resembles the real login page of Inria fromhttps://cas.inria.fr/cas/login?service=, users will mistakenly enter username and password of the Inria to a fake website which the attacker can later submit to the real Inria login page. This is a phishing attack on the Inria and users/members registered with Inria. The real and fake login pages of Inria are given in the Figures 1. Both of the web pages are exactly the same

and it is easy for the users to fall victim of this phishing attack. We have tested our tool PhishCatcher against this and few other attacks as detailed in the Section V.

With the tremendous advancement in modern technologies, there has been a great escalation in the online world, such as e-commerce, online banking, distant learning, e-health and e-governance. Since social networking applications, such as Face book and Twitter, are performing leading role in the globalization of the modern era, billions of users have adopted this increasing trend. Numerous websites provide the web-users with an opportunity to create an account for a customized experience. To obtain online specialized services from the web-sites, users are required to create a personalized account. Conventionally, users are exposed to login web pages for this purpose where they have to set up an account by creating and registering an identification (e.g., username) and secret (e.g., password). Next time, when the user needs to access the remote resource or service, she/he sends a web requests and receives a login form for submitting the identification along with the secret. At this point, the users' privacy is at high risk in terms of identity theft and confidential information. A phishing attack scenario, as described in Figure 2, begins with receiving an email with a link to malicious website [1]. The email message might contain text convincing or luring the user to click and follow the pointer. When the unsuspecting user clicks and opens the web page, it appears genuine as the honest website where the user has an account. After the victim user enters his/her secret information, such as the username and password and presses the submit/login button, they are sent to the attacker. The attacker who sat up the phishing attack receives the secret credentials and logins to the legitimate website upon submitting the credentials to it.

Identity theft, online frauds and scams have immensely increased since the advent of web spoofing or phising attacks. Web spoofing or phishing is a type of cyber crime in which a malicious intruder tries to steal valuable data from the user. Attackers have adopted many phishing and web spoofing techniques to threaten online systems. Initially, webspoofing was used for identity theft but now attackers are using it to steal sensitive information related to national security, intellectual property and organizational secrets. Current era's phishing attacks have already been entered into a new evolutionary dimension including, but not limited to, QR code phishing, spoofing application for mobile and spear phishing etc. Such attacks and scam approaches may circumvent the protections such as firewalls, digital certificates, encryption software and other mechanisms like the two factor authentication. Numerous companies are using such two-factor authentication systems to avoid monetary scams and identity theft. Sadly, the advanced scam approaches have made all these systems vulnerable [2].

To deceive the victims, the attackers normally include logos, either by storing copies or adding links to logos, from the honest site onto their spoof sites to imitate their appearance. In addition to logos, the attacker may also include HTML from the honest site and make some necessary changes. The phishing attack vectors used by the attackers for tricking the users include email, trojan horse, key loggers and man in-the-middle proxies. The favorite attack targets of the attackers are online banking sites, third party payment systems (the most targeted industry sector) and e-commerce sites [3]. As the phisers target the

non-cryptographic components, the cryptographic security protocols SSL/TLS do not provide a complete solution. To depend against spoofing attacks, these protocols must be complemented with additional protection mechanisms [4]. These mechanisms may be enforced at the server-side or client-side or both. The server-side solutions [5], [6] requires changes to the websites which is a tedious job and is often ignored by most of the developers [7]. The client side solutions, on the other hand, provide protection to users without the server support. Server-side solutions may be effective in identifying spoofed site, however, the focus of this paper is on client-side solutions. Most of the anti-spoofing tools are based on either the third party certification [8], password [9] or URLs [3].

Anti-spoofing tools are sometimes categorized as stateful or stateless. They may also be classified based on the automatic phishing detection mechanism used: blacklists and heuristics. Tools that rely on black/white lists generate almost zero false positives (accuracy) and can recognize almost 90% of the phishing sites [10], however, they miss zero-day attacks [11]. Furthermore, black-listing methodologies come with several drawbacks as they cannot control the changing domain and new attacks and can easily be fooled by the spam URLs [12]. To capture phish sites not included in the black lists, the heuristic-based techniques have been very encouraging. The heuristic (content) based tool such as CANTINA [13] and Spoof Catch [1] can identify 90% phishing sites with 1% false positives. The latency of the tool Spoof Catch is in the order of seconds and it further increases with passage of time. While the stateful anti-phish techniques are good in accuracy, they quickly fill the local storage and the performance degrades with passage of time. In Spoof Catch, the visual similarity is initially compared with few login page images, but as the user browse further websites, the number of login page images increases in the local storage. In addition, this increases the time to compare the image of a received login page with every login image in the storage. Following this line of research, we design and develop a stateless anti-phish tool based on the Machine Learning (ML) technique.

From the last decade, many renowned researchers have proposed machine learning techniques for the detection of malicious URLs to avoid any kind of scam in future. Many sets of URLs are treated as training data in the ML approaches. On the basis of the statistical properties obtained by the training sets, it is proposed that whether the requested URL is a scam or scam free. Training data is the primary concern for the URL identification using ML. Once training data is obtained then it is further processed to obtain a mathematical model. The primary concern is to collect the features from the training data because simple strings may not help to predict the status of the URL under test. At final stage, an actual model is obtained through predicted model from the training data. Machine learning techniques, such as Naïve Bayes, Support Vector Machines (SVM) and Logistic Regression (LR), are a few algorithms being used for this purpose by many scholars but there are several issues which make them vulnerable [14].

In this paper, we propose and develop a stateless client-side tool, dubbed as PhishCatcher, to protect against web spoofing attacks. The PhishCatcher, a Google Chrome extension, is based on machine learning techniques and implements the random forest algorithm to

classify whether or not a login web page is legitimate or spoofed. We have evaluated the efficiency and accuracy of the PhishCatcher on real web applications and the results were remarkable. The source code of the Google Chrome extension PhishCatcher is available online at the link https://github.com/wilstef/PhishCatcher.

The major contributions of this research work are the following.

• A client-side anti-phishing mechanism based on the machine learning is proposed.

• Design and development of a Google Chrome extension, PhishCatcher, implementing the proposed mechanism.

• Careful selection of web features for the phish classifier algorithm used in the extension, and

• Experimental analysis of the PhishCatcher.

The rest of the paper is organized as follows. A summary of the related work in the literature is given in the next section. The detailed research methodology followed during this research is discussed in the Section III. The design and development of the Google Chrome extension is described in the Section IV. The testing results of the Chrome extension are included in the Section V and evaluated in the Section VI. The paper is concluded in section VII.

II. EXISTING SYSTEM

Wilayat et al. [1] designed and developed a phish identification tool, called SpoofCatch, based on visual similarity. Using SpoofCatch, when the user first time visits a website, its login web page is identified and its screenshot is stored locally.

Zhang et al. [13] applied a content focused strategy to detect malicious phishing techniques. In the proposed methodology based on the Term Frequency-Inverse Document Frequency (TF-IDF) filter [17], 95% of the phishing URLs were detected accurately. A browser extension PWDHASH++ was proposed in the [18] for client-side protection against phishing. The authors suggested a method to identify visual similarities between the two web sites. The suggested solution, based on Gestalt philosophy [19], acknowledges a web page as a single indivisible entity. These indivisible super signals are explicitly evaluated using algorithmic complexity analysis.

Kaur and Sharma [22] implemented the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [23] algorithm for malicious e-mail detection. An interesting feature of their implementation is that, after a phished URL is detected, it automatically generates a mail and sends it to the victim server. The email message includes the IP, location and contact info of the attacker server and blocks all the traffic coming from the server with malicious intentions. The authors in [24] have combined the machine learning and Resource Description Framework (RDF) to reduce false positives and enhance accuracy of their proposed model. Several machine learning approaches have been applied by the authors in [25] such as Linear Model (LM), Decision Tree (DT), Random Forest (RF) and Neural Networks (NNs) on the test data to detect phishing and malicious sites.

Mao et al. [26] have described few attributes of web page that can be implemented to recognize phished URLs. They designed a logistic regression classifier and used it as a filter to distinguish phishing sites. It was observed that out of millions of URLs, approximately 777 phishing web sites were visited per day and almost 8.24% users were affected.

Xiang et al. [32], proposed an anti-phishing approach based on CANTINA+ model. A

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

filtering algorithm has been adopted to lower FP ratios. Moreover, the designed model was trained on linear and non-linear phishing test beds. Lakshmi and Vijaya [33] applied supervised machine learning techniques including multi-layer perceptron, Naïve Bayes classifier and decision tree classifier to classify and predict malicious websites. Different features were extracted from a collection of 200 URLs and the HTML source codes of the bogus and legal websites. The two performance standards, predictive precision and quick learning combined with 10-fold cross validation determined the efficiency of the model. Their findings showed that the decision tree classifier outperformmed the rest of the classifiers.

Disadvantages

1. In the existing work, the system did not implement Set of fake and their legitimate login page image pairs (visual similarity based) and URL parameters (URL based) and Web page content (content based), and Blacklist.

III.PROPOSED SYSTEM

In the proposed system, the system proposes and develop a stateless client-side tool, dubbed as PhishCatcher, to protect against web spoofing attacks. The PhishCatcher, a Google Chrome extension, is based on machine learning techniques and implements the random forest algorithm to classify whether or not a login web page is legitimate or spoofed. The proposed system evaluated the efficiency and accuracy of the PhishCatcher on real web applications and the results were remarkable.

Advantages

• A client-side anti-phishing mechanism based on the machine learning is proposed.

• Design and development of a Google Chrome extension, PhishCatcher, implementing the proposed mechanism.

• Careful selection of web features for the phish classifier algorithm used in the extension.

IV.MODULES

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Browse and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of Web Spoofing Attack Status, View Web Spoofing Attack Status Ratio, Download Trained Data Sets, View Web Spoofing Attack Status Ratio Results, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT ATTACK STATUS TYPE, VIEW YOUR PROFILE.

V.CONCLUSION

Users have become dependent on the online applications as they provide significant quality of service in many domains i.e.,

online banking, e-commerce, social connectivity, digital libraries, online health services, virtual education, digital marketing and multi-player gaming applications. Commonly, an authentication procedure is followed by the users for the creation of their online account to access the private web content. The security and privacy of users is at stack amid highly sophisticated web spoofing attacks. Several research and commercial tools have been developed to fight against web spoofing attacks but most of them appear with a few lapses. We have developed an optimized user-friendly browser plug-in dubbed as Phish Catcher for the smart disclosure of phishing attacks based on supervised machine learning. Contrary to the traditional approaches, our scheme offers to run the classification in the browser itself. It addresses the loopholes in the existing web applications by fixing the latency issues and improving the efficiency of the tool. The user interface of our plug-in is made simple for the better understanding of the user. When a user enters a phished URL, it displays a phishing alert on the screen and highlights the corresponding phishing features of that URL in a drop-down menu.

The feature-set contains thirty features which are categorized into four groups where each group is acknowledged as a decision tree. Random forest classifier employs the aggregated outcome of the decision trees to identify the bogus and genuine login web pages. The data-set for testing and evaluation comprises of 400 malicious and 400 legitimate URLs. The criteria for testing and evaluation is based on a confusion matrix which enlists the true positives, true negatives, false positives and false negatives. Our plug-in displayed remarkable classification results with the precision and recall, both to be 98.5% and accuracy of 98.5%. Furthermore, the average latency of the plug-in was just 62.5 milliseconds which was measured by running it over forty phished URLs.

The feature set contains thirty features, though, the addition of more automated features might be a great idea to improve the overall performance. Some other discriminative classifiers such as SVM can also be implemented for the prediction of fake or real URL by training larger data-sets. Evaluation metrics may also be evolved by using different tools for a better performance analysis.

VI. REFERENCES

[1] W. Khan, A. Ahmad, A. Qamar, M. Kamran, and M. Altaf, ‘‘SpoofCatch: A client-side protection tool against phishing attacks,’’ IT Prof., vol. 23, no. 2, pp. 65–74, Mar. 2021.

[2] B. Schneier, ‘‘Two-factor authentication: Too little, too late,’’ Commun. ACM, vol. 48, no. 4, p. 136, Apr. 2005.

[3] S. Garera, N. Provos, M. Chew, and A. D. Rubin, ‘‘A framework for detection and measurement of phishing attacks,’’ in Proc. ACM Workshop Recurring malcode, Nov. 2007, pp. 1–8.

[4] R. Oppliger and S. Gajek, ‘‘Effective protection against phishing and web spoofing,’’ in Proc. IFIP Int. Conf. Commun. Multimedia Secur. Cham, Switzerland: Springer, 2005, pp. 32–41.

[5] T. Pietraszek and C. V. Berghe, ‘‘Defending against injection attacks through context-sensitive string evaluation,’’ in Proc. Int. Workshop Recent Adv. Intrusion Detection. Cham, Switzerland: Springer, 2005, pp. 124–145.

[6] M. Johns, B. Braun, M. Schrank, and J. Posegga, ‘‘Reliable protection against

session fixation attacks,'' in Proc. ACM Symp. Appl. Comput., 2011, pp. 1531–1537.
[7] M. Bugliesi, S. Calzavara, R. Focardi, andW. Khan, ''Automatic and robust client-side protection for cookie-based sessions,'' in Proc. Int. Symp. Eng. Secure Softw. Syst. Cham, Switzerland: Springer, 2014, pp. 161–178.
[8] A. Herzberg and A. Gbara, ''Protecting (even naıve) web users from spoofing and phishing attacks,'' Cryptol. ePrint Arch., Dept. Comput. Sci. Eng., Univ. Connecticut, Storrs, CT, USA, Tech. Rep. 2004/155, 2004.
[9] N. Chou, R. Ledesma, Y. Teraguchi, and J. Mitchell, ''Client-side defense against web-based identity theft,'' in Proc. NDSS, 2004, 1–16.
[10] B. Hämmerli and R. Sommer, Detection of Intrusions and Malware, and Vulnerability Assessment: 4th International Conference, DIMVA 2007, Switzerland, July 12-13, 2007 Proceedings, vol. 4579. Cham, Switzerland: Springer, 2007.