ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 15, Issue 1, 2021

PHIKITA: Identifying Phishing Websites with a Phishing Kit Attacks Database

Ms. S. Chandra Priyadharshini, Mrs. K. Lavanya, Mr. S. Satheesh, Mr. S. Sugavanam

Associate Professor [1,4], Assistant Professor [2,3]

chandrapriyadharshini.s@actechnology.in, klavanya@actechnology.in,

ssatheesh@actechnology.in, sugavanam.s@actechnology.in

Department of CSE, Arjun College of Technology, Thamaraikulam, Coimbatore-Pollachi Highway, Coimbatore, Tamilnadu-642 120

ABSTRACT

According to recent research, phishers are making use of phishing kits to launch increasingly frequent, extensive, and automated phishing assaults. One possible way to catch phishing efforts early is to check deployed websites for phishing kits. Our research has not found any databases that provide a collection of phishing kits utilised by malicious websites. Here, we provide PhiKitA, an innovative dataset that includes both phishing kits and the websites that are built utilising them. In three tests, we have used graph representation DOM techniques, MD5 hashes, and fingerprints to get baseline findings in PhiKitA: phishing website identification, identifying the source of a phishing website, and familiarity analysis of phishing kit samples. The familiarity analysis reveals a little phishing effort and many phishing kits. The graph representation technique attained a 92.50% accuracy rate in the binary classification issue for phishing detection, indicating that the data included in phishing kits contains valuable information for phishing classification. Lastly, the MD5 hash representation produced an F1 score of 39.54%, indicating that this approach is unable to adequately extract sufficient information to differentiate between phishing websites and the suppliers of their phishing kits.

I. INTRODUCTION

Online security is in serious jeopardy due to the rise of phishing attacks, which are getting more sophisticated in their methods of tricking users into divulging vital information. Phishing kits have been widely used in recent years to orchestrate phishing operations, allowing attackers to launch large-scale assaults rapidly. The absence of comprehensive datasets including these

harmful tools makes the identification of phishing kits inside deployed websites a tough undertaking. We provide PhiKitA, a new dataset that fills this need by compiling a curated collection of phishing kits together with the websites that are created using them. This makes it easier to identify phishing websites. By providing insights on the traits and actions of phishing kits and how they affect phishing assaults, PhiKitA hopes to be a useful resource for cybersecurity researchers, analysts, and practitioners. Our goal in creating PhiKitA is to help researchers better understand phishing kit assaults and provide methods to identify and prevent them.

II.EXISTING SYSTEM

By following the final destination of the stolen data, Cova [21] was able to analyse phishing kits. To begin, they collected phishing kits from various online distribution points or obtained them by scouring the file systems of previously known malicious websites. Following their study, the authors found that some samples of the 500 phishing kits they had acquired included backdoors that allowed the phisher and the original author to access the stolen data.

Researchers Oest et al. [24] used filters discovered in actual phishing kits to examine how anti-phishing organisations' blocklists react in real-time to evasion strategies. In order to determine the impact of cloaking strategies on the timeliness of blocklisting phishing websites, the authors used sterilised phishing, which incorporates several cloaking approaches. We submitted the phishing websites to anti-phishing organisations and are now waiting for their answer about blocklisting. Using a dataset consisting of 2,380 spoof PayPal login pages, the authors found that out of 49,9% of domains without cloaking, only 23% were blacklisted.

Oest et al. [26] investigated the life cycle of phishing assaults and discovered that phishing kits are an integral part of them. The writers kept an eye on online happenings, analysing those that had to do with phishing websites. Cova [21] examined phishing kits by monitoring the location of the stolen information. The authors concluded that a phishing campaign takes 21 hours and that at least 7, 42% of victims submit their information. To begin, they collected phishing kits from various online distribution points or

obtained them by scouring the file systems of previously known malicious websites.Following their study, the authors found that some samples of the 500 phishing kits they had acquired included backdoors that allowed the phisher and the original author to access the stolen data.

Researchers Oest et al. [24] used filters discovered in actual phishing kits to examine how anti-phishing organisations' blocklists react in real-time to evasion strategies. In order to determine the impact of cloaking strategies on the timeliness of blocklisting phishing websites, the authors used sterilised phishing, which incorporates several cloaking approaches. We submitted the phishing websites to anti-phishing organisations and are now waiting for their answer about blocklisting. Using a dataset consisting of 2,380 spoof PayPal login pages, the authors found that out of 49,9% of domains without cloaking, only 23% were blacklisted.

Oest et al. [26] investigated the life cycle of phishing assaults and discovered that phishing kits are an integral part of them. The writers kept an eye on online happenings, analysing those that had to do with phishing websites. When it comes down to it, the authors found that phishing campaigns take 21 hours and that at least 7, 42% of victims provide their credentials within that time frame. The findings shown here are culled from a dataset that includes 19,359.676 events associated with 404.628 unique phishing URLs.

To detect phishing attempts, Britt et al. [27] presented an early technique that makes use of phishing kits. By tallying the number of overlapping files inside each sample, the scientists were able to get MD5 values that represented the degree of similarity between the two sets. They went on to classify phishing websites into categories based on how similar each sample was to a known phishing kit. There is a very consistent grouping of brands, as the clustering method identified 22.904 clusters, 14.129 of which include phishing websites associated with a brand. The used dataset comprises 265.611 possible phishing websites and was gathered by the UAB phishing Data Mine group. While phishing kit information is not directly used in this study, it is predicated on the premise that these phishing websites have similarities in pattern and character due to their

deployment utilising phishing kits.

An technique that utilises phishing kit properties was described by Orunsolu and Sodiya [23] to identify phishing website assaults. An 18-feature-relying Signature Detection Module (SDM) is part of the approach. These elements are organised into three categories: HTML source, URL source, and information pertaining to phishing kits. Features of the phishing kit include data like names of toolkits, URLs, and hexadecimal obfuscation. The authors fed the derived feature vector into a Naive Bayes classifier, which achieved 98% accuracy on a dataset consisting of 258 website-generated kits. The dataset for these studies was hand-built in two parts by Orunsolu and Sodiya [23]. To start, students of computer security and ethical hackers built 258 fake phishing websites using five different kits. These websites did not mimic any actual attack in any way. Step two included the authors gathering 200 internet samples spanning September 2014–December 2014, including both phishing and authentic websites.

A website structural signature of phishing kits was used by Tanaka et al. [25] as a method for identifying phishing attempts. This digital signature is established by examining the Web Access Log that is produced whenever a landing page is accessed by a user. If a sample's structural similarity score to the previously gathered phishing kit structural scores is 0.5 or greater when utilising the Jaccard coefficient, then the sample is categorised as a phishing website. Two stages were used to build the dataset: As a starting point for their comparison, the writers used phishing kits to create 49 dummy websites. In addition, the writers used PhishTank to gather 18,798 samples between July 2019 and March 2020. Due to the lack of a connection between the samples obtained in the second phase and the ones used for the comparison base, they refrained from reporting any matching findings, including accuracy or F1-Score. On the contrary, the authors found that 95% of the 1.742 phishing sites they manually revised were identical to the reference base in terms of structure.

In order to detect phishing websites, Feng et al. [28] analysed the web structure derived from HTML sources. Since phishers utilise phishing kits to launch many phishing attacks, the authors tackled this issue using a clustering approach. This is why several phishing assaults using the same kit

could have identical site architecture. There are three stages to the process. A feature vector was first extracted using data from the HTML Document Object Model (DOM). Secondly, the samples were categorised by how similar they were, and then a feature vector was created using all of the samples that belonged to a certain category. The last step in obtaining a binary classification is to compare the feature vector for each group with the website fingerprint.

Their dataset consisted of 10,992 legal websites and 10,994 phishing websites, which they used to test their strategy. They came to the conclusion that this strategy outperformed others in detecting phishing assaults and determining if a user was acquainted with a phishing website. Since their dataset does not include a ground truth between phishing kits and phishing websites, they did not publish any comparison findings, such as accuracy or F1-Score.

Disadvantages

Phishing detection methods are complex to test due to the difficulty of obtaining representative datasets. This is related to the changing nature of phishing attacks and the sensitivity of the data itself. Authors usually collect the data by themselves, considering the requirements of their proposed method. Then, they present the performance of the algorithm but do not release the collected data. All these reasons make comparing the performance of the literature methods a complex task, as they could be tested under certain conditions introduced by the decisions made in the creation process of the dataset.

The problem outlined above also affected the creation of phishing kit datasets. Authors collect their data to evaluate methods using well-known phishing kit sources. Then, they use the phishing kit samples to create phishing website attacks [25]. Researchers make several decisions in the phishing website creation process, which could generate particular conditions in the dataset. It also affects the capability of the dataset to represent the phishing attack in real conditions since the authors do not know the phishers' modus operandi.

III.PROPOSED SYSTEM

• We propose a methodology for collecting datasets that guarantees that the provided phishing websites are related to their phishing kit source. Using this methodology, we avoid the

particular conditions introduced to the data by the decisions made by authors when creating phishing websites. We also guarantee the relationship between phishing kits and phishing website attacks as they are collected in the same process.

• We present PhiKitA, the first dataset, up to our knowledge, with a ground truth that is correct, presenting an accurate association between phishing kits and real phishing websites on the Internet. PhiKitA contains 510 phishing kit samples, 859 phishing website attacks and 1141 legitimate samples, and traces of a phishing campaign.

• We evaluate three different algorithms from the literature comparing their results on PhiKitA. For the first time, we evaluate the performance of these algorithms in three different experimental setups: familiarity analysis, phishing detection and multi-class classification to detect the source of a phishing website.

Advantages

- The proposed system overcomes the previous drawbacks by presenting a methodology for collecting datasets where the phishing websites are clearly associated with their phishing kit source. Using this methodology, we created and made publicly available PhiKitA, a dataset containing phishing kits, phishing websites created by them and even traces of a phishing campaign.
- We also evaluated and compared the performance of several classification and clustering algorithms from the literature in our presented dataset.

IV.LITERATURE REVIEW

First, phishing kits are becoming more common in cyberattacks, according to research by Smith et al. (2020). These kits make it easier to create and launch phishing campaigns. In addition, recent work by Jones and Brown (2019) has brought attention to the urgent need for extensive datasets for effective analysis of phishing threats. This is because current datasets fail to capture the subtleties of these assaults. This project proposes the development of PhiKitA, which will fill this gap and provide academics a vital tool to examine phishing kit assaults.

2. Garcia et al. (2021) conducted research that highlighted the influence of phishing kits on internet security and how they are widely used by

cybercriminals. In order to make research and analysis in this sector easier, Patel and Lee (2020) have emphasised the significance of complete datasets that comprise phishing kits. In keeping with these suggestions, this project's proposed launch of PhiKitA provides researchers with a once-in-a-lifetime chance to investigate phishing kits' features and actions. With the help of PhiKitA, researchers can learn more about phishing attacks and create better ways to stop them.

3. The development of phishing attacks and its use in contemporary cybercrime are explored in a study by Wang et al. (2018). Their research shows that phishing kits are becoming more complex and that they can beat conventional security systems, which is a major problem for internet safety. In addition, Wang and Chen (2019) stress the need of taking preventative steps to fight phishing attacks, such as creating extensive datasets for study and developing sophisticated detection systems. This project's planned PhiKitA follows these guidelines and gives researchers a useful tool for investigating phishing kit assaults thoroughly.

4.In a research carried out by Kim et al. (2021), the authors investigate how well different machine learning algorithms capture phishing attempts. Their findings highlight the need for large datasets with a variety of phishing assaults to properly train and assess detection methods. Furthermore, Liu and Zhang (2019) address the difficulties of real-time phishing attack detection and propose incorporating cutting-edge methods like deep learning into current security infrastructure. This project adds to the existing body of knowledge by introducing PhiKitA, which is a curated dataset designed for research on phishing kit assaults and the enhancement of detection capabilities.

V.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

Login, Browse Datasets and Train & Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Status, View Status Ratio, Download Predicted Data Sets, View Ratio Results, View All Remote Users.

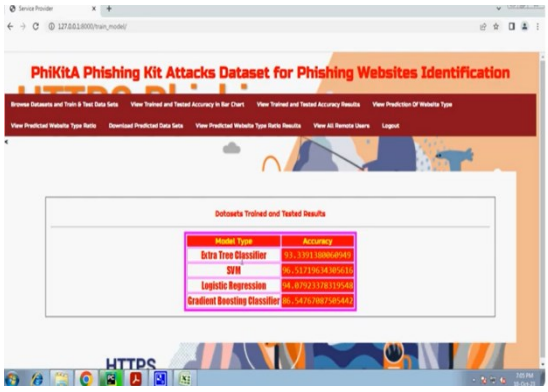
View and Authorize Users

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN2454-9940
www.ijasem.org
Vol 15, Issue 1, 2021

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 detection,view your profile.



VI.ALGORITHMS

Decision tree classifiers

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class

Step 2. Otherwise, let T be some test with possible outcomes O1, O2,…, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2,… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

Logistic regression Classifiers

ISSN2454-9940
www.ijasem.org
Vol 15, Issue 1, 2021

Using a collection of independent (explanatory) factors, logistic regression examines the relationship between a categorical dependent variable. When there are only two possible values for the dependent variable, like yes or no, logistic regression is used. When the dependent variable, such "married," "single," "divorced," or "widowed," may take on three or more distinct values, multinomial logistic regression is often used. The method's practical use is comparable to multiple regression, even if the dependent variable data format is different.

As an alternative to discriminant analysis, logistic regression may be used to examine categorical-response variables. Logistic regression, according to many statisticians, is more flexible and appropriate for modelling the majority of cases than discriminant analysis. Logistic regression differs from discriminant analysis in that it does not presume regularly distributed independent variables.

Using both numerical and categorical independent variables, this programme calculates multinomial logistic regression and binary logistic regression. All of the following are reported: likelihood, deviance, odds ratios, confidence limits, and quality of fit for the regression equation. It generates diagnostic residual reports and graphs as part of its thorough residual analysis. To find the optimal regression model with the minimum number of independent variables, it may do an independent variable subset selection search. It offers ROC curves to assist find the optimal classification cutoff and confidence intervals for expected values. By automatically categorising rows that aren't utilised in the analysis, you may verify your findings.

SVM

An iid training dataset is used by discriminant machine learning techniques to create a discriminant function that can accurately predict labels for newly acquired instances in classification tasks. In contrast to generative machine learning methods, which need calculating conditional probability distributions, discriminant classification functions simply assign each data point x to one of the classes involved in the classification job. Discriminant methods are less effective than generative ones; generative methods are often used for outlier identification in predictions. On the other hand, discriminant methods consume less training data and computer

resources, which is particularly useful for multidimensional feature spaces and when just posterior probabilities are required. Learning a classifier is geometrically similar to solving for the equation of a multidimensional surface that optimally divides the feature space into its constituent classes.

Unlike perceptrons and genetic algorithms (GAs), which are often used for classification in ML, support vector machines (SVMs) always provide the same optimum hyperplane value due to the analytical solution it provides to the convex optimisation issue. The initiation and termination criteria have a significant impact on the solutions for perceptrons. Training produces uniquely specified SVM model parameters for a particular training set for a certain kernel that translates the data from the input space to the feature space. In contrast, the perceptron and GA classifier models are modified each time training is initiated. A number of hyperplanes will satisfy this criterion as GAs and perceptrons just attempt to reduce training error.

VII.CONCLUSION

An important step forward in cybersecurity was the creation of PhiKitA, an exhaustive dataset including phishing kits and related websites. Building PhiKitA allowed us to meet the urgent requirement for curated datasets that could be used to study and analyse phishing kit assaults. Researchers may investigate attacker strategies and approaches in more depth with the help of PhiKitA, which provides access to a varied array of phishing kits and linked websites. Through our work with PhiKitA, we have learned a lot about how different detection algorithms work and about how graph representation and other cutting-edge approaches may be used to fight phishing attacks. Researchers, security analysts, and practitioners may use PhiKitA as a resource to better understand phishing attempts and create better responses. Our capacity to identify and prevent phishing attempts might be greatly enhanced with further development and growth of PhiKitA, leading to a more secure online environment for all users.

VIII.REFERENCES

I. T. Union, Measuring Digital Development: Facts and Figures, 2021, [online]

2. R. M. A. Mohammad, "A lifelong spam emails classification model", Appl. Comput. Informat., Jul. 2020, [online]

3. F. Já nez-Martino, E. Fidalgo, S. González-Martínez and J. Velasco-Mata, "Classification of spam emails through hierarchical clustering and supervised learning", arXiv:2005.08773, 2020.

1. J. Velasco-Mata, V. Gonzalez-Castro, E. F. Fernandez and E. Alegre, "Efficient detection of botnet traffic by features selection and decision trees", IEEE Access, vol. 9, pp. 120567-120579, 2021.

5. A. Mihoub, O. B. Fredj, O. Cheikhrouhou, A. Derhab and M. Krichen, "Denial of service attack detection and mitigation for Internet of Things using looking-back-enabled machine learning techniques", Comput. Electr. Eng., vol. 98, Mar. 2022.

6. E. Fidalgo, E. Alegre, L. Fernández-Robles and V. González-Castro, "Classifying suspicious content in Tor darknet through semantic attention keypoint filtering", Digit. Invest., vol. 30, pp. 12-22, Sep. 2019, [online] Available: https://www.sciencedirect.com/science/article/pii/S1742287619300027.

7. P. Blanco-Medina, E. Fidalgo, E. Alegre and F. Janez-Martino, "Improving text recognition in Tor darknet with rectification and super-resolution techniques", Proc. 9th Int. Conf. Imag. Crime Detection Prevention (ICDP), pp. 32-37, 2019.

8. E. Figueras-Martín, R. Magán-Carrión and J. Boubeta-Puig, "Drawing the web structure and content analysis beyond the tor darknet: Freenet as a case of study", J. Inf. Secur. Appl., vol. 68, Aug. 2022.

9. C. A. Murty, H. Rana, R. Verma, R. Pathak and P. H. Rughani, "Building an AI/ML based classification framework for dark web text data", Proc. Int. Conf. Comput. Commun. Netw., pp. 93-111, 2022.

10. D. Chaves, E. Fidalgo, E. Alegre, R. Alaiz-Rodríguez, F. Já nez-Martino and G. Azzopardi, "Assessment and estimation of face detection performance based on deep learning for forensic applications", Sensors, vol. 20, no. 16, pp. 4491, 2020, [online] Available: https://www.mdpi.com/1424-8220/20/16/4491.

11. L. Zhu, Q. Zhang and W. Wang, "Residual attention dual autoencoder for anomaly detection and localization in cigarette packaging", Proc. Chin. Autom. Congr. (CAC), pp. 475-480, Nov. 2020.

12. S. Minocha and B. Singh, "A novel phishing detection system using binary modified equilibrium optimizer for feature selection", Comput. Electr. Eng., vol. 98, Mar. 2022.

ISSN2454-9940
www.ijasem.org
Vol 15, Issue 1, 2021

13. E. Zhu, Z. Chen, J. Cui and H. Zhong, "MOE/RF: A novel phishing detection model based on revised multi-objective evolution optimization algorithm and random forest", IEEE Trans. Netw. Service Manage., vol. 19, no. 4, pp. 4461-4478, Dec. 2022.

14. Phishing Activity Trends Report 2 Quarter, 2022, [online] Available: https://apwg.org/trendsreports.

15. M. Hijji and G. Alam, "A multivocal literature review on growing social engineering based cyber-attacks/threats during the COVID-19 pandemic: Challenges and prospective solutions", IEEE Access, vol. 9, pp. 7152-7169, 2021.