Familiarity Analysis and Phishing Website Detection using PhiKitA Dataset

Felipe Castaño*‡, Alicia Martínez-Mendoza*†, Eduardo Fidalgo*†, Rocío Alaiz-Rodríguez*†, Enrique Alegre*†

*Department of Electrical, Systems and Automation Engineering, Universidad de León, León, ES
†Researcher at INCIBE (Spanish National Cybersecurity Institute), León, ES
‡ VICOMTECH Research Center, Bilbao, ES
Email: *{felipe.castano, alicia.martinez, eduardo.fidalgo, rocio.alaiz, enrique.alegre}@unileon.es

***Abstract*—Phishing kits are tools used by phishers to deploy phishing attacks faster, more easily and on a larger scale. Detecting phishing kits could aid in the early detection of phishing campaigns by recognizing patterns resulting from the use of phishing kits in the creation of the attack. In this paper, we proposed a methodology to collect phishing kit data and created PhiKitA, a novel dataset that contains phishing kits and websites generated with them. Using PhiKitA, we performed three experiments (familiarity analysis, phishing website detection, and multiclass classification of phishing kits) and evaluated three algorithms: MD5 hashes, fingerprints, and graph representation DOM. The first experiment shows evidence of different phishing kits, the second indicates that the algorithms retrieve useful information to detect phishing with an accuracy of 92.50%, and the third experiment indicates that the algorithms do not retrieve enough information to classify phishing.**

***Index Terms*—Cybersecurity, cybercrime, cyber threats, phishing, social engineering, phishing kits**

Type of contribution: *Already published – PhiKitA: Phishing Kit Attacks Dataset for Phishing Websites Identification [1]*

I. Introduction

Over the past few decades the Internet has grown from 20% of the population with Internet access in 2007 to 67% in 2023[1], making the protection of Internet users and their data a major concern. Researchers have contributed to different topics related to cybersecurity, such as phishing detection [2], spam detection and classification [3] or detecting vulnerabilities in critical infrastructures [4], among others. Our proposal aims to detect phishing trying to reduce the impact of this type of cyberattack, which is categorised as "Fraud" according to the Classification of Cybersecurity Incidents[2].

Phishing is a cybercrime that uses social engineering to deceive people and steal their financial account credentials or other sensitive information. Phishers create replicas of legitimate websites to deceive victims into believing they are accessing the legitimate website. This type of cyberattacks increased significantly in 2022 when $1,097,811$ phishing attacks where reported [5]. To create phishing attacks fast and on a large scale, cybercriminals use phishing kit tools, which set up a server where the attack will be deployed and define the URL and HTML source code of the illegitimate websites. The use of phishing kit data would enable the early detection of phishing campaigns, reducing the impact of these attacks. Previous research [7], [8], [9] has not clearly established the relationship between phishing kits and their associated phishing websites. This is a summary of an already published research [1], where we propose a methodology to collect phishing kit data and create PhiKitA, a phishing kit dataset that guarantees the relationship between phishing kits and phishing websites. For the first time, we use MD5 hashes, fingerprints and the graph representation Document Object Model (DOM) to evaluate three tasks: familiarity analysis, phishing website detection and classification of phishing kits.

[1]https://www.itu.int/itu-d/reports/statistics/2023/10/10/ff23-internet-use/
[2]https://www.incibe.es/incibe-cert/incidentes/taxonomia

II. Related work

Works related to phishing kits can be studied from two perspectives: their behaviour and the support for phishing identification. To analyse the behaviour of the phishing kits, one can consider the destination where the users' stolen information is stored [6], the time responses of anti-phishing blocklists, and the life cycle of the phishing websites. Other researchers have studied the support of phishing kits for phishing identification, using MD5 hashes [7], HTML source code, URL source, phishing kit information [8], and fingerprint representations based on filenames and paths [9]. However, the phishing kit datasets used in these works do not include a relationship between the phishing kits and the phishing websites. PhiKitA aims to overcome these limitations by providing information about which websites are related to the different phishing kits analysed.

III. PhiKitA dataset

We collected phishing websites from PhiskTank[3], OpenPhish[4], Phishing.Database[5], and PhishStats[6]; and we created a script to gather new phishing reports every hour. We obtained the legitimate website samples from Quantcast Top Sites[7] and The Majestic Million[8]. Then, we used Kitphishr[9], a phishing kit collector, to save phishing kits, their domain and the URL where they were found. We used this information and the

[3]https://phishtank.org/
[4]https://openphish.com/index.html
[5]https://github.com/mitchellkrogza/Phishing.Database
[6]https://phishstats.info/
[7]https://www.quantcast.com/products/measure-audience-insights/
[8]https://majestic.com/reports/majestic-million
[9]https://github.com/cybercdh/kitphishr

ISBN:978-84-09-62140-8

previously collected phishing websites to create a new list of reporting websites where a phishing kit was found. We then used a crawler, as in [11], to collect the URL, HTML content and technologies present in the target.

As a post-processing step, we discarded images or other files and removed duplicates. We verified that phishing kits and their corresponding phishing websites matched, to avoid introducing incorrect websites into our dataset due to the cloaking techniques used by phishers. Finally, the dataset contains 510 phishing kits, 859 phishing websites and 1141 legitimate sites, and is available on our website[10].

IV. Experimentation and Results

To evaluate the usefulness of the dataset, we implemented three algorithms: MD5 hash [7], fingerprint representation using path files [12], and HTML DOM analysis [10], and performed three experiments.

For the *familiarity analysis*, we used the MD5 hash to compare the files of the phishing kits. We considered that two kits are related if they shared above 75% of files, following the approach proposed in [12]. This experiment strongly suggests that PhiKitA contains 50 familiarity groups, identifying two types of phishing kit families. The first includes families that share the same functionalities and file distribution, belonging to this class a group with 37 samples. The second type includes families with different functionalities, but they all attack the same target. This class includes another significant group that targets Standard Bank, an international bank and financial services provider.

For *phishing detection*, we performed a binary classification into legitimate or phishing websites based on the similarity to the phishing kits. We set a threshold of 0.46 to determine whether a website belongs to the phishing or legitimate class. Graph representation, MD5 hash and fingerprint representation achieved accuracies of 92.50%, 91.69% and 83.25%, respectively. This could indicate that the information from the phishing kits is useful for detecting phishing websites.

Finally, for the *phishing kits classification* experiment, we discarded samples of phishing kits without phishing websites and considered phishing kits deemed familiarity-related in the first experiment as belonging to the same class. Results show that the algorithms do not extract enough information to distinguish between phishing websites and their phishing kit sources. As shown in Table I, the MD5 hash algorithm achieves the best performance, with an F1-score of 39.54, while the fingerprint and graph representation algorithms achieve F1-scores of 9.03 and 31.11, respectively.

TABLE I
Performance of phishing detection and kits classification.

		Graph	MD5	Fingerprint
Phishing detection	(Accuracy)	**92.50**	91.69	83.25
	(F1-score)	**91.16**	89.60	81.17
Phishing kits classification	(F1-score)	31.11	**39.54**	9.09

V. Conclusions and Future Work

This paper presents a novel methodology to collect data for phishing kits and corresponding phishing websites, including information about their relationship. We have created the PhiKitA dataset, a publicly available dataset that researchers can use to evaluate their proposals in three different tasks: familiarity analysis, phishing binary classification and multi-class classification.

The familiarity analysis on the PhiKitA dataset revealed the presence of 50 phishing kit groups. Moreover, the binary classification experiment for phishing detection showed that the algorithms extract relevant information to detect phishing websites based on the phishing kit information, with the graph representation algorithm achieving an accuracy of 92.5%. Finally, the multi-class classification experiment showed the opposite results in distinguishing between phishing websites and their phishing kit sources. The highest F1-score, 34.92%, was obtained with the MD5 hash algorithm.

In future work, we will extend the PhiKitA dataset by adding more samples and including data that could be used in other approaches, such as screenshots of the samples. We will also modify the collection process to consider the cloaking techniques employed by phishers. Due to the presence of these techniques, 235 out of the 510 samples that we collected do not contain phishing websites related to them.

[10]https://gvis.unileon.es/dataset/phikita-500/

Acknowledgements

This work has been funded by the Recovery, Transformation, and Resilience Plan, financed by the European Union (Next Generation), thanks to the LUCIA project (Fight against Cybercrime by applying Artificial Intelligence) granted by INCIBE to the University of León.

References

[1] Castaño, F., Fernández, E. F., Alaiz-Rodríguez, R., & Alegre, E.: "PhiKitA: Phishing Kit Attacks dataset for Phishing Websites Identification", in *IEEE Access*, 2023.

[2] Sánchez-Paniagua, M., Fernández, E. F., Alegre, E., Al-Nabki, W., & Gonzalez-Castro, V.: "Phishing URL detection: A real-case scenario through login URLs", in *IEEE Access*, vol. 10, pp. 42949-42960, 2022.

[3] Jánez-Martino, F., Fidalgo, E., González-Martínez, S., & Velasco-Mata, J.: "Classification of spam emails through hierarchical clustering and supervised learning", in *arXiv preprint arXiv:2005.08773*, 2020.

[4] Chaves, D., Fidalgo, E., Alegre, E., Alaiz-Rodríguez, R., Jáñez-Martino, F., & Azzopardi, G.: "Assessment and estimation of face detection performance based on deep learning for forensic applications", in *Sensors*, vol. 20, n. 16, pp. 4491, 2020.

[5] Anti-Phishing Working Group: "Phishing Activity Trends Report 2 Quarter", 2022. [Online]. Available: https://apwg.org/trendsreports

[6] Castaño, F., Fidalgo-Fernández, E., and Jañez-Martino, F.: "Creation of a Phishing Kit Dataset for Phishing Websites Identification". León, Spain: TFM, Univ. León, 2022.

[7] Britt, J., Wardman, B., Sprague, A., & Warner, G.: "Clustering Potential Phishing Websites Using DeepMD5", in *5th USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 12)* 2021.

[8] Orunsolu, A. A., Sodiya, A. S., Akinwale, A. T., & Olajuwon, B. I.: "An Anti-Phishing Kit Scheme for Secure Web Transactions", in *ICISSP*, pp. 15-24, 2017

[9] Tanaka, S., Matsunaka, T., Yamada, A., & Kubota, A.: "Phishing site detection using similarity of website structure", *n 2021 IEEE conference on dependable and secure computing (DSC)*, pp. 1-8, IEEE, 2021.

[10] Feng, J., Qiao, Y., Ye, O., & Zhang, Y.: "Detecting phishing webpages via homology analysis of webpage structure", in *PeerJ Computer Science*, vol. 8, pp. e868, 2022.

[11] Sánchez-Paniagua, M., Fidalgo, E., Alegre, E., & Alaiz-Rodríguez, R.: "Phishing websites detection using a novel multipurpose dataset and web technologies features", in *Expert Systems with Applications*, vol. 207, pp. 118010, 2022.

[12] Bijmans, H., Booij, T., Schwedersky, A., Nedgabat, A., & van Wegberg, R.: "Catching phishers by their bait: Investigating the dutch phishing landscape through phishing kit detection", in *30th USENIX Security Symposium (USENIX Security 21)*, pp. 3757-3774, 2021.