LITERATURE SURVEY

Project Domain : Appiled Data Science

Project Name : Web Phising Detection

Team Lead : Vidhyalakshmi R

Team Members : Sharmila D, Sneha A, Yazhini S

ABSTRACT

Phishing is one of the many types of cybercrime targeting internet users. A phishing message is sent with the aim to obtain information from a potential victim. One of the reasons phishing is popular has to do with the connectivity that the internet provides. A message can be spread to thousands of recipients with little eort and at negligible cost. A successful phishing attack can lead to identity theft and loss of money for the victims. When an organisation is targeted, phishing can lead to, among other things, compromised network security and stolen intellectual property.Phishing is highly scalable. On the other side of the scalability spec- trum are less scalable modus operandi. We categorise less scalable methods as “shing for information”. In this thesis, we aim to explore the spectrum of scalability. This thesis uses a socio-technical approach by describing both experiments and technical perspectives to “shing” and phishing. Finally, we performed a large-scale analysis of phishing emails in the Netherlands. We discuss patterns in terms of both attacker behaviour as well as recipient behaviour. Our results demonstrate the eectiveness of phishing with dierent degrees of scalability. Less scalable methods of attack require more eort on the part of the attacker, but provide higher eectiveness. More scalable attacks provide lower success rates, but require less eort than scalable attacks. The contributions in this thesis allow researchers and security professionals to better understand the dynamic nature of phishing.

INTRODUCTION

Phishing attacks have become a significant concern owing to an increase in their numbers. It is one of the most widely used, effective, and destructive attacks, in which attackers try to trick users into revealing sensitive personal information, such as their passwords and credit card information. A typical phishing attack technique involves using a phishing website, where the attacker lures users to access fake websites by imitating the names and appearances of legitimate websites, such as eBay, Facebook, and Amazon. As shown . it is difficult for the average person to distinguish phishing websites from normal websites because phishing websites appear similar to the websites they imitate. In many cases, users do not check the entire website URL, and, once they visit

a phishing website, the attacker can access sensitive and personal information.

With the growth in the field of e-commerce, phishing attack and cybercrimes are rapidly growing. Attackers use websites, emails, and malware to conduct phishing attacks. According to the Anti-Phishing Working Group (APWG) Q4 2020 report, in 2020, there was an average of 225,759 phishing attacks per month, an increase of 220% compared to 2016 . The country most affected by phishing sites is China, with 47.9% of machines infected. Phishing has become one of the biggest threats in cybersecurity. According to the FBI Internet Crime Center data records, the economic loss due to phishing crimes can reach $3.5 billion in 2019

Phishing crimes are usually underreported. New phishing detection techniques have been developed to mitigate phishing attacks. A detailed review of the methodologies of various anti-phishing papers is given by Mohammad et al. Phishing website detection techniques are categorized into four types, whitelist/blacklist-based techniques, deep learning-based detection, machine learning-based detection, and heuristic-based detection techniques, as described

LITERATURE SURVEYS

Phishing Activity Trends Report: 4rd Quarter 2020. *Anti-Phishing Work. Group. Retrieved April* **2021**, *30*, 2020.

1. Almomani, A.; Wan, T.C.; Altaher, A.; Manasrah, A.; ALmomani, E.; Anbar, M.; ALomari, E.; Ramadass, S. Evolving fuzzy neural network for phishing emails detection. *J. Comput. Sci.* **2012**, *8*, 1099.
2. Prakash, P.; Kumar, M.; Kompella, R.R.; Gupta, M. Phishnet: Predictive blacklisting to detect phishing attacks. In Proceedings of the 2010 Proceedings IEEE INFOCOM, San Diego, CA, USA, 14–19 March 2010; pp. 1–5.
3. Zhang, J.; Porras, P.A.; Ullrich, J. Highly Predictive Blacklisting. In Proceedings of the USENIX Security Symposium, San Jose, CA, USA, 28 July–1 August 2008; pp. 107–122.
4. Cao, Y.; Han, W.; Le, Y. Anti-phishing based on automated individual white-list. In Proceedings of the 4th ACM Workshop on Digital Identity Management, Alexandria, VA, USA, 31 October 2008; pp. 51–60.
5. Srinivasa Rao, R.; Pais, A.R. Detecting phishing websites using automation of human behavior. In Proceedings of the 3rd ACM Workshop on Cyber-Physical System Security, Abu Dhabi, United Arab Emirates, 2–4 April 2017; pp. 33–42.
6. Rao, R.S.; Ali, S.T. Phishshield: A desktop application to detect phishing webpages through heuristic approach. *Procedia Comput. Sci.* **2015**, *54*, 147–156. [CrossRef]
7. Joshi, Y.; Saklikar, S.; Das, D.; Saha, S. PhishGuard: A browser plug-in for protection from phishing. In Proceedings of the 2008 2nd International Conference on Internet Multimedia Services Architecture and Applications, Las Vegas, NV, USA, 14–17 July 2008; pp. 1–6.
8. Teraguchi, N.C.R.L.Y.; Mitchell, J.C. Client-side defense against web-based identity theft. In Proceedings of the Network and Distributed System Security Symposium, San Diego, CA, USA, 5 February 2004; pp. 5–18.
9. Sahingoz, O.K.; Buber, E.; Demir, O.; Diri, B. Machine learning based phishing detection from URLs. *Expert Syst. Appl.* **2019**,
10. Rao, R.S.; Pais, A.R. Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Comput. Appl.* **2019**, *31*, 3851–3873. [CrossRef]
11. Le, H.; Pham, Q.; Sahoo, D.; Hoi, S.C. URLNet: Learning a URL representation with deep learning for malicious URL detection.
12. Xiang, G.; Hong, J.; Rose, C.P.; Cranor, L. Cantina+ a feature-rich machine learning framework for detecting phishing web sites.
13. Huh, J.H.; Kim, H. Phishing detection with popular search engines: Simple and effective. In Proceedings of the International Symposium on Foundations and Practice of Security, Paris, France, 12–13 May 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 194–207.
14. Whittaker, C.; Ryner, B.; Nazif, M. Large-scale automatic classification of phishing pages. In Proceedings of the Network and Distributed System Security Symposium, NDSS 2010, San Diego, CA, USA, 28 February–3 March 2010.
15. Miyamoto, D.; Hazeyama, H.; Kadobayashi, Y. An evaluation of machine learning-based methods for detection of phishing sites. In Proceedings of the International Conference on Neural Information Processing, Vancouver, BC, Canada, 8–11 December 2008; Springer: Berlin/Heidelberg, Germany, 2008

16. Zhang, Y.; Hong, J.I.; Cranor, L.F. Cantina: A content-based approach to detecting phishing web sites. In Proceedings of the 16th International Conference on World Wide Web, Banff, AB, Canada, 8–12 May 2007; pp. 639–648.
17. Pan, Y.; Ding, X. Anomaly based web phishing page detection. In Proceedings of the 2006 22nd Annual Computer Security Applications Conference (ACSAC'06), Miami Beach, FL, USA, 11–15 December 2006; pp. 381–392.
18. Bouvrie, J. Notes on Convolutional Neural Networks. *Neural Nets* **2006**. Available online: http://cogprints.org/5869/ (accessed

on 22 October 2021).

19. Somesha, M.; Pais, A.R.; Rao, R.S.; Rathour, V.S. Efficient deep learning techniques for the detection of phishing websites. *Sa¯dhana¯*

2020, *45*, 1–18. [CrossRef]

20. Parra, G.D.L.T.; Rad, P.; Choo, K.K.R.; Beebe, N. Detecting Internet of Things attacks using distributed deep learning. *J. Netw. Comput. Appl.* **2020**, *163*, 102662. [CrossRef]
21. Aljofey, A.; Jiang, Q.; Qu, Q.; Huang, M.; Niyigena, J.P. An effective phishing detection model based on character level convolutional neural network from URL. *Electronics* **2020**, *9*, 1514. [CrossRef]
22. Vrbanc˘ic˘, G.; Fister Jr, I.; Podgorelec, V. Datasets for phishing websites detection. *Data Brief* **2020**, *33*, 106438. [CrossRef] [PubMed]
23. Wang, W.; Zhang, F.; Luo, X.; Zhang, S. Pdrcnn: Precise phishing detection with recurrent convolutional neural networks. *Secur. Commun. Networks* **2019**, *2019*, 2595794. [CrossRef]
24. Jain, A.K.; Gupta, B. Comparative analysis of features based machine learning approaches for phishing detection. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 March 2016;

CONCLUSION

In this paper, we proposed a multi-level feature phishing website classification method based on character embedding CNN and RF. The main features of this model is as follows.

Character embedding of URLs is performed to convert URLs into normalized matrices, containing much important phishing website classification information in the URL characters. This information helps classify phishing websites. URLs are transformed into uniform signals by the character embedding technique, more suitable for CNN networks' input.

Automatic phishing web feature extractor using CNN. The CNN model is pre-trained using the converted URL data to optimize and improve the CNN model parameters. The pre-trained model can extract multi-level features from the URL data. The ex- tracted multi-level features contain sensitive information that can classify phishing websites and provide knowledge for phishing website classification.

Using multiple RF classifiers and a winner-take-all strategy improves the model's accuracy and generalization. Extracting multi-level features for low latitude can be used to classify phishing websites. The RF classifier is trained using the extracted features of each layer, outputting the results of each RF, and, finally, choosing the one with the best results, improving the classification results.The proposed method in this paper is validated by the dataset from PhishTank and Alex. A 99.35% correct classification rate of phishing websites was obtained on the dataset. Experiments were conducted on the test set and training set,

t