The Role of Fuzzy Logic in Improving Accuracy of Phishing Detection System

Neelam Badi, Mayank Patel, Amit Sinhal

Abstract*: Considering the ease of implementation, the wide vector of targets it addresses and the flexibility to adapt and breach sophisticated technologies, phishing attacks are becoming widely popular among cyber attackers to gain sensitive user data and credentials. Despite the humongous mass of this form of cyber attack, McAfee Labs Threats Report 2018 [1] suggests that 97% of the total number of email users were not able to identify sophisticated phishing content (URL, email, links, etc). Hence, a number of anti-phishing and phishing detection tools based on various techniques such as the blacklist, heuristic, visual, data mining, machine learning, etc are available.However, the subjective ambiguity and the vastness in methodologies used for phishing attacks makes real-time detection and identification of such malicious URLs to be a very complicated and dynamic process, that even machine learning cannot address. Considering the vagueness of phishy URLs, the human cognition-like natural approach of fuzzy logic makes it a suitable choice for tackling the quality factors of the findings rather than just numerical values. In this paper, we will show how usage of fuzzy logic can improve the accuracy of machine learning based phishing detection systems and provide a resilient and smart model for better cyber security.

***Index Terms*: Anti-Phishing, Cyber Security, Fuzzy Logic, Hacking, Machine Learning, Phishing Attack, Phishing Detection**

I. INTRODUCTION

Phishing is the practice of illegally trying to lure a victim on digital internet-based platforms by impersonating as some reputed website or organization - and then fraudulently acquiring sensitive data like passwords, bank details and personal information from the victim through duplicate forms [2]. The form of cyber attack originated in early 90s, and has evolved along with technology to a much more sophisticated form in current times [3].

Over the years, this form of cyber attack has become by far the most exploited attack vector ever, and yet the motives of the attackers remain similar to what it were at the time of its inception. Following are the core motives behind majority of phishing attacks:

- **Piracy of sensitive data, credentials:**

The cheapness, easiness and success ratio of phishing attacks make them a perfect choice for cyber attackers if and when their purpose is to steal credentials, user-sensitive data or gain access to a network.

Revised Manuscript Received on June 05, 2019

Neelam Badi, student of Geetanjali Institute of Technical Studies, Udaipur, Rajasthan

Dr. Mayank Patel, Associate Professor in the Department of Computer Science and Engineering at Geetanjali Institute of Technical Studies, Udaipur, Rajasthan.

Dr Amit Sinhal, Department of Computer Science and Engineering at Geetanjali Institute of Technical Studies, Udaipur, Rajasthan.

- **Stealing PII or PHI data:**

The cyber attacker uses spear phishing technique to target a single entity, community or organization and gain illegal access to Personally Identifiable Information (PII) or Protected Health Information (PHI) using malware, instead of sending out mass blind emails to targets.

- **Getting financial details using fake transactions:**

In contrast with the spear phishing technique, cyber attackers target lower-level employees with administrative rights instead of top level executives and request them to reveal sensitive data via fake transactions or fraudulent C-suite requests. The seriousness of phishing attack problem in modern day can be understood from FBI's 2017 Internet Crime Report [4] that "phishing and phishing-related scams were the third most common type of scam reported by victims regardless of company size, industry, or location". Furthermore, statistical data for the first two quarters of 2018 indicated that as many as 5,00,000 phishy websites were detected, which is almost double compared to the last two quadrant of 2017 [5].

286 major international brands were suffered from phishing attacks in September 2018, of which, most of the targets belonged to online banking, finance and payment sector, followed by SAAS/webmail and financial institutions. Since the targets of most phishing attacks are financial, banking and data sensitive sources or individuals – the phishing attacks comes with an average loss of $1.6 million for a medium sized organization. Proofpoint's 2019 State of the Phish Report [6] supported by which multiple phishing data sources, shows that over 50% of phishing sites are now using HTTPS encryption. Usage of webpage redirection on the phishing website to mask the maliciousness is becoming a common practice among latest cyber attackers too. Hence, relying on a single parameter or methodology compromises the functionality and effectiveness of most existing anti-phishing and phishing detection techniques, at some point. The ease of execution and the success ratio of the form of cyber attack make phishing to be extremely profitable for attackers, and the problem is not going away anytime soon. In fact, as per the aforementioned statistics, increasing in frequency and sophistication, and without proper protection it can be the death blow to organizations. Hence, finding suitable and effective solutions to address the virtually omnipresent problem of phishing has become an urgent necessity. A number of anti-phishing and phishing detection tools based on blacklist, heuristic, visual and data mining approaches are available, but the subjective and ever-changing characteristic of phishing attacks make each of these

tools to be ineffective at some degree when implemented in real-time. Also, majority of the currently available phishing detection techniques rely third parties for extracting features and giving out final results - making them complicated, dependent and sluggish for practical usage.

II. Existing Phishing Detection Methodologies

The model proposed by Chunlin et al. [7] is aimed to find effective classifiers for detecting malicious URLs and hence emphasizes on character-related features. The author have used six machine learning algorithms and improved the accuracy by adding statistical analysis of the input URLs.

In another research paper by Sudhanshu et al. [8] kept the concepts of data mining and combined it with classification algorithm to achieve an accuracy of 92.67%. The authors have mentioned that their proposed model consisted of 16 rules, and the accuracy can be increased by defining better rules since the simplicity of rule-based classification technique in combination with data mining algorithms is better than any other method for phishing detection.

In the phishing detection model proposed by Luong et al. [9], heuristic approach was used six features to determine the authenticity of the input URLs. The model calls for weights and values of these six features and based on the outcome, the legitimacy of the URL is determined. The authors have mentioned that defining better heuristics can improvise the results. Jain and Gupta et al. [10] presented an anti-phishing approach, which uses machine learning by extracting 19 features in the client side to distinguish phishing websites from legitimate ones. They used the 2141 phishing pages from PhishTank (2018) and Openfish (2018), and 1918 legitimate web pages from Alexa popular websites, some online payment gateways, and some top banking websites. With the use of machine learning, their proposed approach reached 99.39% true positive rate. Although most of the researchers focus on the phishing detection through URLs, some researchers tried to detect phishing emails by checking the data stored in the email packets. To detect phishing attacks, Smadi et al. [11] combined the neural network approach with reinforcement learning for classification. The proposed system contains 50 features, which are grouped into four different categories as mail headers, URLs is the content, HTML content and main text and reached 98.6% of accuracy rate and 1.8% false positive rate.

III. Problem Statement

Huge amount of research has been made in anti-phishing area, but as observed in the literature survey, we can conclude that each method has some drawback there is not any single technique that proves to be sufficient for detection of all types of phishing attacks. Phishing attacks evolve and propagate a step further than cyber security tools, which becomes a motive for us to find a universal phishing detection method.

Of all existing methods, machine learning approach seems to be best suitable way, considering its closeness with human cognitive characteristics that help the computers in learning things and doing tasks like humans. While this is beneficial for the versatility of phishing, "machine learning focuses on the development of computer programs that can access data and use it learn for themselves". The outputs of machine learning algorithms are extremely precise and detailed, which is great for achieving precision. However, most of the users who are potential victims of phishing attacks are non-technical and might be unable to interpret detailed outcomes of machine learning systems.

Statistically we absorb 75% our data around us (viewing, hearing, talking, smelling and tasting) and the rest comes from the web or other media. So it can be concluded that brief and clear expression of final outcome will be more effective than detailed verbose explanation or description. That is when the fuzziness of the system comes into picture where we will use fuzzy logic to amalgamate the results of machine learning system and give a short, clear and accurate final result.

IV. Fuzzy Logic

Fuzzy logic is a decision-making method in the world of computers, and is based on the “degrees of truth” instead of straight “true or false”. So unlike other computational logics that are based on literal results, fuzzy logic involves humanly characteristics knowledge and reasoning to control what decision is made and how the computer processes.

For example, for any face detection system, the system can define characteristics like “XYZ’s face is pale”, but the degree of paleness can only be defined by fuzzy logic principles. So rather than having hard-line conclusions like a pale face or not, fuzzy logic can tell us how pale the face is in form of multi valued number from 0 to 1. The mileages we can receive out of fuzziness of any model are naturally more fine-tuned than what we might get with naive binary logic. Also, fuzzy logic is comparatively more forgiving and responsive to system behavior than naive binary logic that only knows how to determine “too many”, and throttle completely, or “not too many”.

V. Fuzzy Logic for Improved Phishing Detection

The use of fuzzy logic for developing more human-like systems is not new as engineers have been utilizing it in a wide range of computer-related applications since decades. As for phishing detection, fuzzy logic gives out informative details of the given URL, helping better assessment and ranking of malicious URLs based on the real-time qualitative analysis. When combined with the precision of machine learning results, the processing of vaguely defined values and variables can be given out in a rather humanly format. Inputs for fuzzy systems can be anything between 0 and 1 representing the degree of uncertainty of the parameter, where 0 represents false values and 1 represents true values. These values similar to the human thinking and is closer to the cognitive decision making power of human brain. Just how we aggregate data to decide the degree of truth which we aggregate further into higher truths instead of declaring any aspect to be 100% true or 100% false - fuzzy logic too aggregates data to check the degree of truth and finally come to a conclusion by combining all these outcomes. The Figure 1 shows the hierarchical flow of the phishing detection system based on machine learning, and enhanced by fuzzy logic. As it can be seen in Figure 1, the entire process is divided into three layers, each consisting 9 parametres that are to be checked; after which, the results of each layer are defuzzified using fuzzy principles to get a singular, clear and accurate result valuing either legitimate, or phishy.

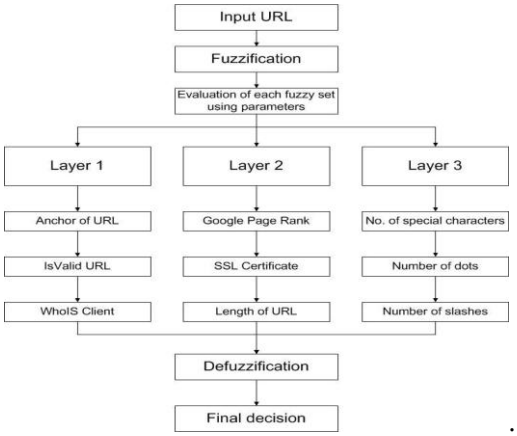

Fig.1. Phishing detection system flowchart

Since this approach provides word-based computational scope while offering support in tackling the design uncertainties - fuzzy logic has become an integral part of machine learning based systems. The accuracy of machine learning systems and the decisive nature of fuzzy logic take perfect care of imprecise and uncertain situations. Also, its close similarity with the functional patterns of human neural networks makes fuzzy logic to be omnipresent in expert systems and other artificial intelligence applications.

VI. CONCLUSION AND FUTURE WORK

The primary purpose of machine learning algorithms is to process the data and collect knowledge for improving the system's functionality with each input, for which traditional methods of clustering, classification and associations are used. And when fuzzy logic is are applied in conjunction with machine learning principles, more accurate, reliable and flexible outcomes are generated. Fuzzy principles and reasoning is widely incorporated in machine learning systems because its flexibility and knowledge based representation gave a new life to the scientific facts of machine learning. Fuzzy sets are used when data is being processed and are used for data representation due to their capability to showcase incomplete and unspecified content. Hence, a combination of fuzzy logic and machine learning approach using various parameters and rules make the system to be an optimum choice in phishing detection. The proposed system can be upgraded by adding more defined fuzzy rules and parameters. Also, the machine learning section of the system can be better trained with bigger and better datasets that are currently unavailable for free.

REFERENCES

1. McAfee Labs Threats Report for Fourth Quarter of 2018, December 2018.https://www.mcafee.com/enterprise/en-us/assets/reports/rp-quarterly-threats-dec-2018.pdf
2. Montazer, Gholam Ali, and Sara ArabYarmohammadi. "Detection of phishing attacks in Iranian e-banking using a fuzzy–rough hybrid system". Applied Soft Computing 35 ELSEVIER (2015): pp 482-492.
3. Sahingoz OK, Buber E, Demir O, Diri B (2018) Machine learning based phishing detection from URLs. In: Expert Systems with Applications, ElSEVIER, Volume 11 (March 2019): 345-357. https://www.sciencedirect.com/science/article/pii/S0957417418306067
4. Federal Bureau of Investigation Internet Crime Report 2017, May 2018. https://www.fbi.gov/news/stories/2017-internet-crime-report-released-050718
5. APWG Phishing Activity Trends Report, APWG, October 2018. http://docs.apwg.org/reports/apwg_trends_report_q2_2018.pdf
6. 2019 State of the Phish: Attack Rates Rise, Account Compromise Soars, Proof Point, https://www.proofpoint.com/us/security-awareness/post/2019-state-phish-attack-rates-rise-account-compromise-soars
7. Chunlin Liu, Bo Lang : Finding effective classifier for malicious URL detection : In ACM,2018
8. Sudhanshu Gautam, Kritika Rani and Bansidhar Joshi: Detecting Phishing Websites Using Rule-Based Classification Algorithm: A Comparison : In Springer,2018.
9. L. A. T. Nguyen, B. L. To, H. K. Nguyen and M. H. Nguyen, "Detecting phishing web sites: A heuristic URL-based approach," 2013 International Conference on Advanced Technologies for Communications (ATC 2013), Ho Chi Minh City, 2013, pp. 597-602.
10. Jain, Ankit Kumar and Brij B. Gupta. "A novel approach to protect against phishing attacks at client side using auto-updated white-list". EURASIP J. Information Security 2016 (2016): Page: 9.
11. Smadi, Sami & Aslam, Nauman & Zhang, Li. (2018). Detection of Online Phishing Email using Dynamic Evolving Neural Network Based on Reinforcement Learning. Decision Support Systems. 107. 10.1016/j.dss.2018.01.001.

AUTHORS PROFILE

Neelam Badi is a student of Geetanjali Institute of Technical Studies, Udaipur, Rajasthan (Affiliated to RTU, Kota and AICTE, New Delhi), studying Masters of Technology in Computer Science Engineering. She has completed her Bachelors in Engineering from Gujarat Technological University. She has served as the first-female executive member of for Institution of Engineers (India)'s SLC centre. She is working on developing an advanced anti-phishing system using a combination of fuzzy logic and machine learning. She has also worked on developing a practically useful campus recruitment automation system using C#.NET and holds expertise in Search Engine Optimization through her practical experience while working with an IT company in Gujarat.

Dr. Mayank Patel is working as an Associate Professor in the Department of Computer Science and Engineering at Geetanjali Institute of Technical Studies, Udaipur, Rajasthan. (Affiliated to RTU, Kota and AICTE, New Delhi). He completed Ph.D. in the domain of Multimedia Services over Wireless LAN from College of Technology and Engineering (MPUAT, Udaipur). His area of interest includes Programming in C, Data Structure and Algorithms, OOPS, Programming in JAVA, Web Application Development through J2EE, Web Application Development through .Net Framework, Computer Networks and Wireless Networks, Principles of Programming Languages and many more. He has also published large number of quality research papers in prominent international journals. A project in under the edge of AIESEC, Indonesia has been carried out under his supervision. He also publishes various textbooks in the programming domain.

Dr Amit Sinhal is HOD of the Department of Computer Science and Engineering at Geetanjali Institute of Technical Studies, Udaipur, Rajasthan. (Affiliated to RTU, Kota and AICTE, New Delhi). He holds 10 years of IT Industry experience as software developer & Project Manager; and 11 years of teaching & research experience. He has published 1 patent, 1 international book, 4 book chapters, and more than 50 research papers in international journals. He has presented in more than 10 papers in International and National Conferences, organized 4 International/National Conferences and attended more than 10 FDPs & Workshops. Dr Sinhal has also delivered numerous lectures and key-note addresses in conferences and training programs as the conference chair, a conference paper reviewer, and an invited speaker. He has supervised 45 M. Tech. scholars and currently supervising 3 Ph.D. scholars. He has also received the award of the Best Academician of the Year, Best Head of the Department & Best Faculty. He is a member of ACM, life member of ISTE, CSI, IAENG, AMLE and CSTA; and is also a reviewer with 4 international journals.