SURVEY ON PHISHING WEBSITES DETECTION

[1]Mrs. Vaneeta M, [2]Pratik N N, [3]Prajwal D, [4]Pradeep K S, [5]Suhas Kakade K

[1]Associate professor, Department of Computer Science and Engineering, K S Institute of Technology,
[2,3,4,5]Undergraduates, Computer Science and Engineering, K S Institute of Technology,
Bengaluru,Karnataka, India-560109, Affiliated to VTU, Belagavi.

Abstract : Phishing is a website forgery with an intention to track and steal the sensitive information of online users.It is a form of identity theft, in which criminals build replicas of target websites and lure unsuspecting victims to disclose their sensitive information like passwords, PIN, etc.It is one of the social engineering methods that gathers personal information through malicious websites and deceptive e-mail to canvass personal information from a company or an individual .Phishing is often carried out by using email as a medium to users that represents a part of a company or an institution who perform business such as financial institution, banking etc . Phishing is becoming more malicious day by day and its detection is very important. In cyberspace, phishing is motivating the researchers to develop the model through which we can develop more security towards the safe services provided by the web.

***IndexTerms* - Phishing websites,Phishing types,Phishing detection techniques, Machine learning classifiers, Cyber security.**

I. INTRODUCTION

In recent years, the web has evolved explosively due to the availability of numerous services such as online banking, entertainment, education, and social networking. Accordingly, a huge volume of information is downloaded and uploaded constantly to the Web. This gives opportunities for criminals to hack important personal or financial information, such as usernames, passwords, account numbers and national insurance numbers. This is called a Web phishing attack, which is considered as one of the major problems in Web security. Worldwide spending on cyber security is forecasted to reach $133.7 billion in 2022, 62% of businesses experienced phishing and social engineering attacks in 2018, 68% of business leaders feel their cyber security risks are increasing,Only 5% of companies' folders are properly protected, on average, Data breaches exposed 4.1 billion records in the first half of 2019[15].

The success of phishing website detection techniques mainly depends on recognizing phishing websites accurately and within an acceptable timescale. Many conventional techniques based on fixed black and white listing databases have been suggested phishing websites. However, these techniques are not efficient enough, since a new website can be launched within few seconds. Therefore, most of these techniques are not able to make an accurate decision dynamically on whether the new website is phishing or not. Hence, many new phishing websites may be classified as legitimate websites.

This paper develops an anti-web spoofing solution based on inspecting the URLs and content of fake web pages. This solution developed takes series of steps to check characteristics of websites URLs.

Here we propose a phishing website detection scheme using a Wrapper feature selection based on 14 features, to detect phishing websites with high accuracy. In addition, neural network, support vector machine, and random forest classification techniques have been employed in detection of phishing websites. The wrapper feature Selection method selects 14 features and uses these features in different classification techniques and compares the results with the feature selection method of 31 features.

II. TYPES OF PHISHING

1.Vishing is a name given to voice phishing. Here attack is done based on gathering data in the caller's details. We do not require a fake website to perform this attack. Taking the help of fake caller-ID, by giving an appearance that data is obtained from the trusted organisation.

2.Smishing is the name given to SMS phishing To reveal the personal information text messages are used as a tool for inducing people from their mobiles. This is a technique used in this SMS phishing.

3.Tab nabbing Opening multiple tabs at a time is an advantage of tab nabbing. Redirecting the user to affected site and other types. Reverse technique is method loaded here that is copying the affected sites into the original site happens here.

4.Pharming In General all attackers do normal traditional phishing, but only some attackers will use the idea of "baiting' on the selected victims entirely. Pharming, a type of attack being used where stems from domain name system (DNS) cache poisoning is done.

5.Spear Phishing Spear phishing is done by sending mail to a targeted individual. Phishers generally got the information of individuals through social media sites such as Linkedin, Facebook and use of fake addresses for sending emails that similarly happens to be the mail that was received from anyone of our co-workers.

6.Deceptive Phishing is one of the most common ways of phishing. Attacking the customers for stealing the personal information and login credentials happens here.

III. PHISHING DETECTION TECHNIQUES

1.Visual similarity based:

A user could become the victim of the phishing attack by looking the high visual resemblance of phishing website with the targeted legitimate site, such as page layouts, images, text content, font size, and font colour. The fake and genuine web pages have same visual appearance but different URLs. It is not always necessary that the people carefully notice on URL and SSL (Secure Socket Layer) certificate of websites. If an attacker does not copy the visual appearance of targeted website well, then chances of inputting credentials by Internet users are very less.

2. Machine learning based:

A real-time anti-phishing system, which uses seven different classification algorithms and natural language processing (NLP) based features. The system has the following distinguishing properties from other studies in the literature: language independence, use of a huge size of phishing and legitimate data, real-time execution, detection of new websites, independence from third-party services and use of feature-rich classifiers. For measuring the performance of the system, a new dataset is constructed, and the experimental results are tested on it. According to the experimental and comparative results from the implemented classification algorithms, Random Forest algorithm with only NLP based features gives the best performance with the 96.98% accuracy rate for detection of phishing URLs.

3. Intelligent phishing possible Detector

For detecting phishing websites and mails, efficient techniques were formed by Fuzzy logic in combination and association of classification data mining algorithms.

4. Detection of phishing E-mails using CS-SVM

To reduce damage of phishing attacks, some email detection techniques have been proposed. These can be grouped under as whitelist, blacklist, content-based approach and network-based approach.

5. Dynamic Malware Analysis

Malware is used to share a lot of characteristics with legitimate software like creating files, modifying registry keys which communicates over the network. This requires putting in place monitoring tools that captures malware activity on the machine. Malware activities information has been gathered using two approaches.One of them is static and other is dynamic analysis. Both dynamic and static analysis uses different approaches to collect data. Depending on the circumstances and available options these methods are used.

IV. LITERATURE SURVEY

Ankit Kumar Jain and B. B. Gupta [1] proposed Visual similarity based phishing detection techniques, utilise the feature set like text content, text format, HTML tags, Cascading Style Sheet (CSS), image and so forth, to make the decision. These approaches compare the suspicious website with the corresponding legitimate website by using various features and if the similarity is greater than the predefined threshold value then it is declared phishing. In order to avoid phishing detection technique, attackers usually insert images, Flash, ActiveX and Java Applet in place of HTML text. Visual similarity based detection approaches can quickly detect such embedded objects present in phishing webpage. These techniques use a signature to identify phishing webpages. The signature is created by taking common features from the whole website rather than a single webpage.Therefore, one signature is sufficient to detect various targeted webpages of a single website or different versions of a website. This approach matches the URL, SSL certificates,and webpage contents, which is an advantage over blacklist based approaches.

R. Kiruthiga and D. Akila [2] proposed a novel approach to detect phishing websites using machine learning algorithms. They also compared the accuracy of five machine learning algorithms Decision Tree (DT), Random Forest (RF), Gradient Boosting (GBM), Generalized Linear Model (GLM) and Generalized Additive Model (GAM). Top three algorithms namely Decision Tree, Random Forest and GBM performance were compared,Random Forest algorithm has given highest 98.4% accuracy, 98.59% recall and 97.70% precision. Also this paper proposed a efficient way to detect phishing URL websites by using c4.5 decision tree approach. This technique extracts features from the sites and calculates heuristic values. These values were given to the c4.5 decision tree algorithm to determine whether the site is phishing or not. Dataset is collected from PhishTank and Google. This process includes two phases namely pre-processing phase and detection phase. In which features are extracted based on rules in pre-processing phase and the features and their respected values were inputted to the c4.5 algorithm and obtained 89.40% accuracy.

JIAN MAO et al. [3] proposed a straight forward approach to detect phishing pages, that is to compare all CSS rules of two web pages and calculate the similarity rate according to the number of matched selectors. The visual appearance of a web page is decided by its page layout and contents. To achieve a consistent appearance across all variants of web browsers, web developers use Cascading Style Sheets (CSS) as the standard technique to represent the layout of web pages. CSS includes a series of rules that specifies the visual properties of web page elements. The browsers retrieve the CSS specification of the webpages and render them accordingly.

A. MahaLakshmi et al. [4] explained various phishing detection techniques such as Intelligent phishing possible Detector,Honey Pots,Detection of phishing E-mails using CS-SVM,Dynamic Malware Analysis,Anti-Phishing Simulator, and Detection of phishing URL using Artificial Neural Network.These types of detection mechanisms help to prevent phishing to an extent and this paper will help the general public for taking prevention as well as precautionary steps against the phishing attacks.

Moitrayee Chatterjee and Akbar Siami Namin [5] concentrate upon Modeling the identification of phishing websites through Reinforcement Learning (RL): where an agent learns the value function from the given input URL in order to perform the classification task. In Deep Reinforcement Learning-Based classifier they train a deep neural network as a reinforcement learning agent. To automate the problem they employ a two-step procedure: 1) Feature Extraction 2)Deep Reinforcement Learning.

S. Carolin Jeeva and Elijah Blessing Rajsingh [6] focused on discerning the significant features that discriminate between legitimate and phishing URLs. These features are then subjected to associative rule mining—a priori and predictive a priori. The rules obtained are interpreted to emphasize the features that are more prevalent in phishing URLs. Analyzing the knowledge accessible on phishing URL and considering confidence as an indicator, the features like transport layer security, unavailability of the top level domain in the URL and keyword within the path portion of the URL were found to be sensible indicators for phishing URL. In addition to this number of slashes in the URL, dot in the host portion of the URL and length of the URL are also the key factors for phishing URL.

Karim Hashim et al. [7] proposed Mobile Phishing Websites Detection and Prevention Using Data Mining Techniques.The widespread use of smart phones nowadays make users vulnerable to phishing. Mobile devices facilitate phishing attacks due to the following properties. Firstly the rapid increase of mobile users worldwide. Secondly the limited screen sizes makes it difficult for mobile users to determine legitimate web-page from phishing one.To minimize time wastage and system resources consumption, a System Data Base [SDB] have been utilized. Check if the domain name is an IP as a verification of their identities, legitimate websites use their company, institute or services names as a domain name.This work models the prediction of phishing websites on mobile devices as a classification task and demonstrate the machine learning approach to predict the websites status and take the proper action towards it.

Routhu Srinivasa Rao et al. [8] concentrates on URL and Website Content of phishing page. PhishShield takes URL as input and outputs the status of URL as phishing or legitimate website. The heuristics used to detect phishing are footer links with null value, zero links in body of html, copyright content, title content and website identity. PhishShield is able to detect zero hour phishing attacks which blacklists unable to detect and it is faster than visual based assessment techniques that are used in detecting phishing.

Hemali Sampat et al. [9] proposed a system which detects the phishing using features of URLs and WHOIS protocol. They used classification and association Data Mining algorithms to identify and characterize all rules and factors in order to classify the phishing website and relationship that correlate them with each other to detect them by their performance, accuracy, number of rules generated and speed.

Ram B.basnet et al. [10] focusses to study the anatomy of phishing URLs that are created with the specific intent of impersonating a trusted third party to trick users into divulging personal data. Unlike previous work in this area, that only use a number of publicly available features on URL alone. In addition, compares performance of different machine learning techniques and evaluates the efficacy of real-time application of methods used. Applying it on real-world data sets, they demonstrate that the proposed approach is highly effective in detecting phishing URLs. It uses a heuristic-based approach to classify phishing URLs by using the information available only on URLs. It treats the problem of detecting phishing URLs as a binary classification problem with phishing URLs belonging to the positive class and benign URLs belonging to the negative class.

Peng Yang et al. [11] proposed a multidimensional feature phishing detection approach based on a fast detection method by using deep learning (MFPD). In the first step, character sequence features of the given URL are extracted and used for quick classification by deep learning. Specially, the CNN. in the second step, they combine URL statistical features, webpage code features, webpage text features and the classification result of deep learning into multidimensional features , which are then classified by XGBoost.

TOMMY CHIN [12] presented PhishLimiter, a new detection and mitigation approach, where they first propose a new technique for deep packet inspection (DPI) and then leverage it with software-defined networking (SDN) to identify phishing activities through e-mail and web-based communication. PhishLimiter utilizes an Artificial Neural Network (ANN) model to adjust to phishing attacks such that PhishLimiter can do self-training for new threat advancement and detection. The proposed Deep Packet Inspection (DPI) approach has two modes: Store and Forward (SF) and a Forward and Inspect (FI) on SDN switching devices running Open vSwitch (OVS). PhishLimiter provides better network traffic management as it has the global view of a network due to SDN.

REFERENCES

[1] Ankit Kumar Jain and B. B. Gupta “Phishing Detection: Analysis of Visual Similarity Based Approaches” National Institute of Technology, Kurukshetra, India Security and Communication Networks Volume 2017.

[2] R. Kiruthiga, D. Akila “Phishing Websites Detection Using Machine Learning” (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019.
[3] JIAN MAO, WENQIAN TIAN, PEI LI, TAO WEI, AND ZHENKAI LIANG “Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity” August 23, 2017.
[4] A. MahaLakshmi, N. Swapna Goud, Dr. G. Vishnu Murthy, “A Survey on Phishing And It’s Detection Techniques Based on Support Vector Method (SVM) and Software Defined Networking(SDN)” IJEAT ISSN: 2249 – 8958, Volume-8, Issue-2S, December 2018.
[5] Moitrayee Chatterjee Akbar Siami Namin “Detecting Phishing Websites through Deep Reinforcement Learning” 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC).
[6] S. Carolin Jeeva and Elijah Blessing Rajsingh “Intelligent phishing url detection using association rule mining” 2016.
[7] Karim Hashim Al-saedi, Mustafa Dhiaa Al-Hassani Mustansiriyah University, Baghdad, Iraq “This paper explains Mobile Phishing Websites Detection and Prevention Using Data Mining Techniques”.
[8] Routhu Srinivasa Rao and Syed Taqi Ali “PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach” Department of Computer Engineering, National Institute of Technology, Kurukshetra, Haryana, India.
[9] Hemali Sampat Manisha Saharkar, Ajay Pandey and Hezal Lopes “Detection of Phishing Website Using machine Learning” IRJET,2018.
[10] Ram B.basnet Andrew H.Sung, Quingzhong Liu Colorado “Learning to detect Phishing URLs”.
[11] Peng Yang, Guangzhen Zhao, Peng Zeng, “Phishing Website Detection based on Multidimensional Features driven by Deep Learning”.
[12] TOMMY CHIN, KAIQI XIONG and CHENGBIN HU “Phishlimiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking” August 20, 2018.
[13] Neda Abdelhamid, Fadi Thabtah and Hussein Abdel-jaber “Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features” IEEE-2017.
[14] Hiba Zuhair, Mazleena Salleh, Ali Selamat “Hybrid features based prediction for novel phish websites”.
[15] www.varonis.com