URL Scrutiny to Circumvent Phishing

Ms. P. Mahalakshmi¹, Bharat Motwani², Akhil Bajpai³

¹Asst. Professor Department of CSE SRM Institute of Science and Technology, Chennai, India.
²,³Department of CSE SRM Institute of Science and Technology, Chennai, India

ABSTRACT-- Increasing the number of cyber-crimes and frauds it has become the need of hour to seriously consider the threat and make some solid security arrangement against it. The links or Uniform Resource locator (URL) can be classified as benign, spam or malicious thus informing the user beforehand about its safety. As emails are hook to attach malicious URL, luring to catch the users. Malicious URL are the most common tool used in breaking identity and attacking either the system database or some financial frauds. Instead of focusing the source of such links if the rigorous testing on various attributes is performed on the URL itself using Random forest algorithm for threat detection and Support Vector Machine (SVM) to consider even the extreme cases is expected to improve the coverage and thus can provide the system user with high level security from counterfeited URL.

Keywords: Clustering, Classification, Apriori, Tokens, Neural Network, Firefly.

1. Introduction
Phishing is an ill-social engineering method to obtain data from the victim’s system and exploit it. Most often email are attached with the links/URL which can redirect the user to a malicious site or it may automatically start downloading some software which may pertain threat to the system health and data.

Thus to counter all this becomes really very important as nowadays people are moving towards complete digitization whereby storing there important data onto the clouds or system. Thus Data mining technologies can help in classifying whether the URL is trustworthy or not. Data mining is also termed as Knowledge Discovery in Database i.e. KDD. It is the process to extracting knowledge from the database i.e. extracting valuable information from the database. Data mining involves few common classes of tasks: Anomaly Detection, Clustering, Classification and Regression. The mail is made in such a way that it attracts the user without any suspicion, onto which the link is b eing attached thus the mail is being treated as a “hook” to attach the URL, the “bait”. No sooner the user click on the link he is being re-directed to some malicious sites, thus he is a “catch” to the trap.

Random Forest Algorithm which will put the URL to test on a decision tree, based on several attributes. Once this is being completed Support Vector Machine will come up with covering the extreme cases so that none are left uncovered.

2. Literature Survey
Blacklisting: The attempt to avoid un-safe URL is by blacklisting them all. But since this is an avoidance mechanism and not a prevention one it requires the beforehand knowledge of the URL. This method simply needs to know all the counterfeited URL, make a list of them and put them under blacklist. It safeguards from the existing threats and not the newer ones [1].

JavaScript Code Analysis: This is basically a browser scripting language to develop the client-side interface for web applications. This code might be designed such that it attacks the user’s browser and its extensions. This is possible by “drive-by downloads” [2], which automatically downloads some malwares on to the user’s system and taking over its platform. The dynamic nature of JavaScript is what leads it to be the dangerous one and its amalgamation with the web browser leads to a deadliest attack as all this are so dynamic that the user’s permission is generally not required and thus can’t be interrupted once started. Thus in order to counter this, the anomalies are detected along with its
emulation over a number of features making use of machine learning algorithms, thus reproducing its behaviour which in turn is differentiated with the existing profiles. It’s also possible to generate signatures for the system which are based on signatures [3].

Extract Meta Data: With the increase in awareness about the phishing, the way to attack the user in a newer ways are also being improved with time making it difficult for the user to defend even after being aware. The same case appeared in 2011, when exponential increase in the attacks were observed as the hackers improvised the existing methods with more sophistication, upgrading the methods as per the technology and finding the loop-holes in our defensive security tools thus this breach necessary measure to be countered. In order to defend from these so-called sophisticate attacks it became pervasive to evaluate the meta-data of the site through data mining processes and algorithms [4]. Once such algorithm is Logistic Regression Classifier [5], which is basically a binary classification algorithm in order to classify a URL as false positive or false negative. The same can be done using few learning based algorithms like Naïve-baiyes, Passive-aggressive, perceptron etc. [6].

Detect Phishing Mail: Evaluating the URL on the basis of hybrid features obtained from the email header and content. Making use of 10-fold cross validation at the time of pre-processing to obtain the features plays a vital role and essential influence on the results of classification the email as safe or not. The basis of this method is to treat the mail itself with suspicion and check its legitimacy before moving on to the URL or website thus saving the significant time as if the hook in itself is malicious there remains no need to check the URL or website [7].

Anomaly Detection: Detecting the anomalies in the link is also an interesting method to determine its legitimacy. Working on the anomalies of the URL message and the sender’s behaviour, a scoring metric is generated and the scores of the same is compared with some threshold. If score is satisfactorily above the threshold it is accessed else blocked. Various anomalies are being taken into consideration, like regular delay time of system, regular response time, entropy delay time etc. [8].

Apriori & Predictive Apriori: It’s important not just to tackle the possible nodes of attack but doing it in the possible efficient way is also important. Upgrading the anomaly detection methodology by making use of associative rule based mining algorithms like apriori and predictive apriori on the keywords in the URL helps to classify the URL as being false positive or false negative in a much efficient way. The apriori is basically a bottom-up approach which generates the candidates and tests them against the data-set, thus the same way it moves across the URL creating its each keyword as its candidate and thus evaluating them against the already fed data-set thus producing the output [9].

Neural Networks and Firefly: Identifying the phished websites from the internet is the basis of this model. It performs the complete task in three steps firstly, identifies the features for the websites secondly, gives the features already found as input to the Neural Networks where respective computations are performed based on the activation functions and its weight the output is produced thirdly, the firefly provides the optimum weight using the objective function. Thus the websites can be classified as being safe or a threat [10].

3. Analysis of the existing testing model

The security researchers are developing newer advance methods of analysing the links / URL and limiting their detrimental impacts as there is always some areas left with some probability of being attacked it becomes essential to comprehend the existing system with the algorithm to defend against the newer threats. The existing model, Focuses either on the hook i.e emails or the websites but email is not the only source nowadays, Lack in focus on URL, Narrow coverage in terms of attributes/nodes of attack, outdated i.e. not updated with the newer threats, lacking the appropriate structure.

Upgrading is required in the existing model in order to cope up with the danger of data breach, financial frauds and taming of information etc. The structure need to be re-defined in order to tackle the current threats and vulnerabilities.
3

Fig. 1: General Control Flow in the existing analysis models.

The figure 1 depicts the general test flow, implementing respective algorithms in order to combat the prospective threats. It is a basic diagram representing the common flow followed by the majority of the existing model, the only difference is that the focus is either email or website and then URL at last and that too not a comprehensive one.

4. Proposed Work

Nowadays sending phished URL to trap the users by sending mails attaching such links is not the only way, counterfeited links can be sent through messages of pop-up so it becomes more significant to focus on the URL rather than the mails or websites. This requires proper structuring of the testing model along with inclusion of few newer nodes of attack. Analysing an URL and detecting it for its safety using an amalgamation of Random Forest vector and Support vector machine algorithm escalates the capacity to be more accurate covering the extreme cases as well i.e. broader coverage.

Fig. 2: Structured attributes for URL analysis.

The figure 2 represents the control flow of the testing attributes. At first testing an URL for its rank on the basis of traffic if the result is satisfactory then moving on to the next level checking whether the link comprises of the ip address or a definite link, in case of it being a non-IP address it is moved further to test the length of URL, if the length is larger than the average length of an URL it is being put under suspicious category but if it is larger than the largest URL available currently then it is simply ignored. Moving on further next attribute tokens are being checked, in order to complete this test an URL is being divided into 2 parts: Domain and Path. Domain part is being tested for http/https, (-), (.) tokens and the path part is tested for the following tokens @, //, email to. At last the URL satisfying all the above conditions is given a green signal to direct the access.
The figure 3 reflects the relationship between the various entities participating in the testing model in producing the correct analysis output, the ER diagram makes it more comprehensible to work at the possible nodes of attack. The above diagram can be used to clearly represent not just the loop-holes but also the version of analysis code being used to counter such nodes. As it is a comprehensive evaluation diagram.

5. CONCLUSION
Analysing the previous and current security vulnerabilities, the various security mechanisms, their working and limitations and how the researchers are coming up with ideas to restrict the detrimental impacts. However if the focus remains directly on the URL irrespective of the source its obtained from and then implementing random forest algorithm in supplementary to support vector machine it might come up with higher accuracy in terms of classifying the URL correctly as phishing or non-phishing since it has broader coverage. Thus the current analytic method needs to be upgraded ensuring wider coverage of vulnerabilities to meet the exponential rise of phishing threats.

6. REFERENCES
[1]. S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, C. Zhang, "An empirical analysis of phishing blacklists", 2009
[2]. “Download Attacks”, Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC) 2010.
[3]. Marco Cova, Christopher Kruegel, Giovanni Vigna, "Detection and analysis of drive-by-download attacks and malicious javascript code", 2010.
[4]. IEEE 2017 - Feature selection for machine learning based detection of phishing websites.
[5]. “Examination of data, rule generation and detection of phishing URLs using online logistic regression” by Mohammed Nazim FerozSusan Mengel Published in: 2014 IEEE International Conference on Big Data (Big Data).
[6]. R. B. Basnet and A. H. Sung, “Learning to Detect Phishing Webpages”, Journal of Internet Services and Information Security (JISIS) 2014.
[7]. 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA) 2015 , “Detection of Phishing Emails using Data Mining Algorithms”.
[8]. Mohiuddin Ahmed, Abdun Naser Mahmood, Jiankun Hu, “A survey of network anomaly detection techniques”, 2016.
[9]. Intelligent phishing url detection using association rule mining by S.Carolin Jeeva and Elijah Blessing Rajsingh Published on Dec 2016.
[10]. Phishing Detection in Websites Using Neural Networks and Firefly, 2016.
[11]. NHK K. ISMAIL*, "Estimation Of Reliability Of D Flip-Flops Using Mc Analysis”, Journal of VLSI Circuits And Systems 1 (01), 10-12.2019.
[12]. MN BORHAN, "Design Of The High Speed And Reliable Source Coupled Logic Multiplexer”, Journal of VLSI Circuits And Systems 1 (01), 18-22, 2019