IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 7, PP 38-40

Study of Various Phishing Detection Methods

Diksha Todankar[1], Sakshi Seth[2], Vidhika Jain[3] and Prof.Reena Mahe[4]
(Information Technology, Atharva College of Engineering, India)
Corresponding Author: Diksha Todankar

Abstract: *Phishing websites are essential to determine as the world is moving towards digitalisation and all the major communication and transactions are carried out through various websites. However, many reported methods include Blacklist, content based, non-content based which use various algorithms to differentiate between legitimate and phishing sites. These methods consider various features. To name some are page rank, URL features, using blacklisted database, positioning of content in a website, domain name, etc. Given a URL the goal of differentiating it as legitimate or phishing is quite challenging based on the reported methods as the attacker will try their best to deceive the user. Numerous techniques have been developed to detect it efficiently and the purpose of this paper is to discuss these methods and their limitations.*
Keywords: *phishing, legitimate, features, URL, random forest, blacklist*

I. Introduction

Phishing is a fraudulent activity attempted to gain access to a user's personal information and financial details. An attacker will create a legitimate looking website and force the user to provide private data such as pins and passwords. Another way would be to pose as a genuine organization and send an email which points towards the attacker's site. This fools the user into clicking the link and filling their personal details. Since the availability of free web hosting services and website builder attackers have exploited these facilities and made their attacks more sophisticated. According to APWG various companies are being targeted every week, some small numbers of companies were targeted more frequently. In 2017, 50,720 unique phishing sites were detected and 452 different brands were targeted. Attackers were targeting the financial and webmail sectors[1].

Due to the boom of social media in the past few years attackers are starting to see the potential in using social media to spread their attacks and reel in as many people as possible into their trap. A recent example would be a phishing attempt made on popular photo sharing social network Instagram. The attackers made a fake account of Paytm and wrote in the bio section that they would offer 4000 to 0.1 million users if they completed a certain list of activities. They had uploaded chat screenshots to trick users into believing the scheme was real.

Following sections describe the various phishing detection methods that exist. Sections 2, 3, 4, 5 describe Enhanced Blacklist, Non content based method using machine learning algorithms like Random Forest, Google's phishing detection system and Visual feature detection respectively.

II. Enhanced Blacklist

It is similar to that of traditional method with an advantage that it focuses more on finding similarities between two websites based on features like attribute values of tag, css used, dynamic nature of website, path links, script used, background images, filenames of url, where the content is placed can be obtained from the source code of a website. These features are then applied to simhash algorithm which gives fingerprints of bits as an output by using many hash functions. Later these fingerprints of website are compared with that of the fingerprints present in blacklist database using hamming distance, if the value comes out to be less than the threshold it is classified as phishing else legitimate. This approach focuses on finding near similarities of a website. Limitations of this approach is it is unable to classify websites if it is an exact replica of the legitimate website or the attacker does not use some or all attribute tags[2].

III. Non content based method

In this method features of URL are used to detect phishing website namely primary domain, sub domain and path domain. Number of times a website is accessed by the user or linked by other websites is used as a heuristic to classify it as phishing. To get access frequencies and number of linkage Alexa rank, page rank, and Alexa reputation are used. The system can be proposed in a format as follows. Firstly, URLs from browsers or datasets are received which are then divided into components to get primary domain, sub domain and path domain. Later on the heuristic values are calculated for which a whitelist database is maintained which contain primary domain of legitimate websites. If the heuristic value comes out to be negative it is considered as a suspicious website else if the value is positive, it is a legitimate website. The heuristic of the system is calculated

by using the product of the individual heuristic obtained from the previous step and weight of that heuristics. Finally, the system value is taken as a threshold to compare the URLs as phishing or legitimate. By keeping a threshold of 0.5 an accuracy of about 97% is obtained[3].

Random forest

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates a forest with a number of trees. In general, the more trees in a forest the more robust the forest will be. This logic is used in the random forest algorithm. The random forest classifier can be used for both classification and regression tasks. It can handle missing values as well. Even though there are more number of trees in the forest the classifier does not over fit the model. k features out of m features are given as an input to the trees of random forest where k<<m.

In the next stage k features are used to find root node using best split, subsequent daughter nodes are also generated using best split. The above methods are followed until forest is generated. For prediction rules defined by the randomly created trees are applied on test features. Votes are counted based on the prediction results and the result which has receive highest votes is considered as the final prediction[4].

IV. Google's phishing detection system

Google has created a phishing detection system which detects potential phishing sites every day. The system has four main processes which are extracting features, getting domain information and crawling the page, if the crawl succeeds then content features are extracted and a final score is assigned. If the score is above a threshold the site will be added to the blacklist. Blacklist has patterns which are created from similar blacklisted pages.

Each process is divided into tasks. A task management system assigns tasks to workers. If the task fails it is put back into the queue and reassigned. Since tasks are independent communication between workers is not needed. If the load cannot be handled by single worker more workers are assigned.

URLs are gathered from spam messages and reports sent by users to Google. We can understand whether a page is phishing by checking the URL. A whitelist is created manually. If the URL matches with the whitelist, then the URL is not considered for further classification. Some URLs have an IP address in the hostname. This is an important feature which can be detected easily. Another feature of phishing URLs is the number of host components. Common words used in URLs to deceive users are collected and converted as Boolean values and are checked with incoming URLs. Other features like URL metadata and reputation score are used. Feature selection is done to retain important features only.

Hosting information of web page is fetched using "Content fetcher" [5]and crawler modifies the page to see what a normal user would see when they visit the page. Crawler looks for iframe, image and JavaScript content of the page. It also makes sure that the number of fetches is less so that high traffic is not generated on the site. Since the attacker cannot use false DNS details, this information can be used for detection.

The collected features are converted into values. Value of a feature is 0 when it is not present. These values are used to compute final score. The final score is computed by combining these scores using logistic regression. The final score is between 0 and 1 where 0 means the site is legitimate and 1 means the site is phishing. In the case where the score is more than 0.5 the classifier marks it as a phishing site. Before a page is added to the blacklist the page's PageRank is checked to make sure there is no misclassification. If it turns out to be a popular page it has to be manually verified before blacklisting.

A function called "blacklist aggregator"[5] will create the blacklist to be displayed. the common broader patterns are combined to avoid repetition of similar patterns. It displays results using the safebrowsing protocol format.

To detect new phishing trends, the classifier is trained offline once every day. URLs from three months ago along with their features are used as dataset. For balancing dataset, not more than 150 URLs per domain are selected per week. Published blacklist is used as noise. URLs not in blacklist considered as legitimate. If a mistake is found it is corrected for future runs and new phishing pages added to blacklist [5].

V. Visual detection method

Various visual detection techniques are: DOM structure, visual feature, CSS, pixel based features, visual perception, hybrid features. To detect phishing sites, the DOM structure of the site can be checked. Phishing sites have similar DOM structures to the original sites. Information of the site from the past is checked with current information of the site. If information matches, domain name is different and layout similarity is beyond the threshold the website is considered phishing. Limitation of this technique is that only previously visited information is checked.

In visual feature detection features like page content, font size, color, logo, etc. are checked. Three techniques viz. visual signature, site signature, Phishzoo (checks SSL and URL of the site) are used here. CSS

features and images of phishing sites have to be kept similar to the original site. This similarity can be used to detect phishing sites.

In pixel based technique image processing is used to detect phishing sites. Screenshots are taken and compared. If the screenshots are similar but the domains are different then one of the sites is phishing. In hybrid detection any of the above techniques are combined. A limitation of visual detection is that this method is used offline and data collection takes too much time [6].

VI. Conclusion

In this paper, we discussed various methods for phishing site detection. Each method has its own advantages and disadvantages. Current methods of phishing detection take too much time which makes it difficult to stop these attacks dynamically. Phishing ventures are most of the time temporary thus the detection techniques have to be fast enough. Detection services need to be made available to the users easily. Lack of awareness is another reason that phishing attacks become successful. Users should be educated about the attacks and how to avoid them.

References

[1]. APWG. (n.d.). APWG Reports. Retrieved from Unifying the Global Response to Cybercrime APWG: http://docs.apwg.org/reports/apwg_trends_report_h1_2017.pdf

[2]. Circulation. 2002;106(25, article 3143).Colin Whittaker, B. R. (n.d.). Research at Google. Retrieved from Research at Google: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35580.pdf

[3]. Gupta, A. K. (n.d.). Phishing Detection: Analysis of Visual Similarity Based Approaches. Retrieved from Hindawi: https://www.hindawi.com/journals/scn/2017/5421046/

[4]. How the random forest algorithm works in machine learning. (n.d.). Retrieved from Dataaspirant: http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/

[5]. Luong Anh Tuan Nguyen, B. L. (2013). Detecting Phishing Web sites: A Heuristic URL-Based Approach. *The 2013 International Conference on Advanced Technologies for Communications (ATC'13).*

[6]. Pais, R. S. (2017). An Enhanced Blacklist Method to Detect. Springer International Publishing AG.