Detection of PhishingWebsites using Machine Learning

Uday singh shergill,
B.tech, computer science
Vellore Institute of technology, Chennai

Abstract— **Phishing attacks continue to be a major security threat for organizations and individuals alike. Phishing attacks are designed to steal sensitive information, such as passwords, credit card numbers, and personal identification numbers (PINs). In this paper, we propose a machine learning-based approach to detect data phishing attacks. Our proposed approach uses a combination of supervised and unsupervised learning algorithms to identify phishing emails and URLs.**

Keywords- Phishing, Extreme Learning Machine.

I. INTRODUCTION

Internet has become an important part of our life to obtain, spread information in social media. While Mobile Social Networks enrich people's lives, it also creates some security issues. In one of the previous studies the author defined phishing as a type of semantic attack in an online environment, where the victims are sent spoofed emails whichessentially deceive them into providing confidential data suchas account numbers, passwords and other personal information to the attacker. To understand what phishing does, we must know the different types of phishing.

Types of Phishing Attacks:

Numerous different types of phishing attacks have now been identified. Some of the more prevalent are listed below.

1. Deceptive Phishing

Deceptive phishing is the most common type of phishing. In this case, an attacker attempts to obtain confidential information from the victims. Attackers use the information to steal money or to launch other attacks.

2. Spear Phishing

Spear Phishing targets specific individuals instead of a wide group of people. Attackers often research their victims on social media and other sites. That way, they can customize their communications and appear more authentic.

3. Whaling

When attackers go after a "big fish" like CEO, it's called Whaling. These attackers often spend considerable time profiling the target to find the opportune moment and means of stealing login credentials.

4. Pharming

Similar to phishing, Pharming sends users to a fraudulent website that appears to be legitimate. However, in this case, victims do not even have to click a malicious link to be taken to the bogus site.

II. RELATED WORKS

The point of this section is to highlight work done by others that uses different techniques to achieve the maximum accuracy result and improve the whole system. Fadi Thabtah *et al.* Experimentally compared large numbers of ML techniques on real phishing datasets and with respect to different metrics. The purpose of the comparison is to reveal the advantages and disadvantages of ML predictive models and to show their actual performance when it comes to phishing attacks. The experimental results show that Covering approach models are more appropriate as anti- phishing solutions. Muhemmet Baykara *et al.* Proposed anapplication which is known as "Anti Phishing Simulator", it gives information about the detection problem of phishing and how to detect phishing emails. Spam emails are added to the database by Bayesian algorithm. Phishing attackers use JavaScript to place a legitimate URL of the URL onto the browser's address bar. The recommended approach in the study is to use the text of the e-mail as a keyword only to perform complex word processing. "Anti Phishing

Simulator" was developed to check the content and determine whether the related message contained phishing elements. Tianrui Peng *et al*. Proposed and named a system as SEAHound processes a document, one sentence at a time and returns tree if the document contains a social engineering attack. It focuses on the natural language text contained in the attack, performing semantic analysis of the text to detect malicious intent. This approach performs a semantic analysis of the text transmitted by the attacker to verify the appropriateness of each sentence. Jhen-Hao Li *et al.* Proposed an approach, called PhishBox to effectively collect phishing data and generates models for phishing validation and detection. It integrates phishing websites collection, detection and validation into an online tool which monitors the blacklisted phishing sites, validates and detects them in real-time. Naghmeh Moradpoor *et al.* Proposes a neural network-based model for detection and classification of phishing emails. It uses real benign emails from "SpamAssassin" dataset and real phishing emails from "Phishcorpus" dataset. Python and MATLAB is used to measure the accuracy, true-positive rate, false positive-rate, network performance, and error histogram. R.Aravindhan *et al.* Proposed a list based anti phishing approach, which has two types 1.Black list 2.White list. In black list some online databases such as phish tank provides list of phishing websites. In white list the user manually builds a white list byadding the trusted website to the white list. In heuristics based anti phishing approach the characteristics are determined such that it reflects the nature of the websiteaccurately, machine learning techniques is used to find the phishing. Mustafa Aydin *et al.* Proposed a classification algorithm for phishing website detection by extractingwebsites' URL features and analyzing subset based feature selection methods. It implements feature extraction and selection methods for the detection of phishing websites. The extracted features about the URL of the pages and composed feature matrix are categorized into five different analyses as Alpha-numeric Character Analysis, Keyword Analysis, Security Analysis, Domain Identity Analysis and Rank Based Analysis. Most of these features are the textual properties of the URL itself and others based on third parties services. Samuel Marchal *et al.* presents PhishStorm, an automated phishing detection system that can analyze in real time any URL in order to identify potential phishing sites. Phish storm is proposed as an automated real-time URL phishingness rating system to protect users against phishing content. PhishStorm provides phishingness score for URLand can act as a Website reputation rating system.

III. PROPOSED METHODOLOGY

I) Classification of the hyperlinks in the phishing e-mails

In order to (illegally) collect useful information from potential victims, phishers generally tries to convince the users to click the hyperlink embedded in the phishing e-mail. A hyperlink has a structure as follows.

<a href="URI "> Anchor text <\a>

where 'URI' (universal resource identifiers) provides the necessary information needed for the user to access the networked resource and 'Anchor text' is the text that will be displayed in user's Web browser.

Examples of URIs are:

- http://www.google.com
- https://www.icbc.com.cn/login.html
- ftp://61.112.1.90:2345.

'Anchor text' in general is used to display information related to the URI to help the user to better understand the resources provided by the hyperlink. In the following hyperlink, the URI links to the phishing archives provided by the APWG group, and its anchor text "Phishing Archive" informs the user what's the hyperlink is about. <a href http://www.antiphishing.org/phishing-archive.html">Phishing Archive </a>

Note that the content of the URI will not be displayed in user's Web browser. Phishers therefore can utilize this fact to play trick in their 'bait' e-mails. In the rest of the paper, we call the URI in the hyperlink the actual link and the anchor text the visual link. After analyzing the 203 (there are altogether 210 phishing e-mails, with 7 of them with incomplete information or with malware attachment and do not have hyperlinks).Phishing email archives from Sep. 21st 2003 to July 4th 2005 provided by APWG.

Different algorithms that can be used to detect the phishing websites are:

A. *Artificial Neural Networks (ANN)*

An artificial neural network (ANN), inspired from biological neural networks, is a set of interconnected nodes (neurons). Each connection between nodes is typically assigned weights. The network learns by adjusting the weights, in the learning phase for correct prediction process. ANNs were considered less suitable for data mining due to their poor interpretability and long training times. However, their advantages include ability to classify patterns on which they have not been trained and high tolerance for noisy data.

B. *K-Nearest Neighbour (k-NN)*

Learning for k-NN classifiers occurs by analogy, that is, by comparing the test tuple to similar training tuples. These are distance-based comparisons that intrinsically assign equal weights to each attribute; therefore, accuracy could be poor when noisy or irrelevant data is presented. However, methods of editing and pruning have been introduced to solve the problem of useless and noisy data tuples respectively. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. The good value for the number of neighbors can be determined experimentally.

C. *Support Vector Machine (SVM)*

Support vector machines (SVMs) are used for the classification of both linear and nonlinear data. In short, when given an original training data, the algorithm uses a nonlinear mapping to transform it into a higher dimension. In this dimension, a linear optimal hyper plane is searched, to keep the data of any two classes separate. SVMs can be used for classification and numeric prediction as well. The simplest form of SVM is a two-class problem, where the classes are linearly separable. For a 2-D problem, a straight line can be drawn to separate the classes, in fact, multiple lines could be drawn.

D. *Random Forests (RF)*

Random Forests can be built in tandem with random attribute selection using bagging. Random Forests follow an ensemble approach to learning, that is a divide and conquer approach for improving performance. In a simple decision tree, the input or test in added at the top and it traverses down the tree, ending up in smaller subsets. In a random forest, the ensemble mechanism combines various random subsets of trees. The input/test traverses through all the trees. The result is calculated based on average or weighted average of the individual results, or the voting majority in case of categorical data. The accuracy of a random forest depends on a measure of the dependence between the classifier and the strength of the individual classifiers and they improve the problem of over fitting of the decision trees.

The components for detection and classification of phishing websites include the discussion on thirty distinct attributes ofwebsites. They are as follows:

A. *Address Bar based Features*

1. Using the IP address

If IP address is used instead of domain name in the URL e.g. 125.98.3.123 the user can almost be sure someone is trying to steal his personal information.

2. Long URL to hide the Suspicious Part

Phishers can use long URL to hide the doubtful part in the address bar.

3. Using URL shortening services "TinyURL"

URL shortening is a method on the "World Wide Web" in which a URL may be made considerably smaller in length and still lead to the required webpage.

4. URL's having "@" symbol

Using "@" symbol in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol.

5. Redirecting using "//"

The existence of "//" within the URL path means that the user will be redirected to another website.

6. Adding Prefix or Suffix Separated by (-) to the Domain The dash symbol is rarely used in legitimate URLs. Phisherstend to add prefixes or suffixes separated by (-) to the domainname so that users feel that they are dealing with a legitimate webpage.

7. Sub Domain and Multi Sub Domains

Let us assume we have the following link: http://www.hud.ac.uk/students/. A domain name might include the country-code top-level domains (ccTLD).

8. HTTPs (Hyper Text Transfer Protocol with Secure Sockets Layer)

The existence of HTTPS is very important in giving the impression of website legitimacy, but this is clearly not enough.

9. Domain Registration Length

Based on the fact that a phishing website lives for a short period of time, we believe that trustworthy domains are regularly paid for several years in advance. In our dataset, we find that the longest fraudulent domains have been used for one year only.

10. Favicon

A favicon is a graphic image (icon) associated with a specific webpage.

11. Using Non-Standard Port

This feature is useful in validating if a particular service is upor down on a specific server.

12. The existence of "HTTPS" Token in the Domain Part of the URL

The phishers may add the "HTTPS" token to the domain part of a URL in order to trick users.

B. *Abnormal Based Features*

1. Request URL

Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain.

2. URL of Anchor

An anchor is an element defined by the <a> tag. This feature is treated exactly as "Request URL".

3. Links in <meta>, <Script> and <Link> tags

Given that our investigation covers all angles likely to be used in the webpage source code, we find that it is common for legitimate websites to use <Meta> tags to offer metadata about the HTML document; <Script> tags to create a client side script; and <Link> tags to retrieve other web resources.

It is expected that these tags are linked to the same domain of the webpage.

4. Server From Handler(SFH)

SFHs that contain an empty string or “about:blank” are considered doubtful because an action should be taken upon the submitted information.

5. Submitting Information to Email

Web form allows a user to submit his personal information that is directed to a server for processing. A phisher might redirect the user’s information to his personal email.

6. Abnormal URL

This feature can be extracted from WHOIS database. For a legitimate website, identity is typically part of its URL.

C. HTML and JavaScript Based Features

1. Website Forwarding

The fine line that distinguishes phishing websites from legitimate ones is how many times a website has been redirected. *Status Bar Customization*

2. Disabling Right Click

Phishers use JavaScript to disable the right-click function, so that users cannot view and save the webpage source code. This feature is treated exactly as “Using onMouseOver to hide the Link”.

3. Using Pop-Up Window

It is unusual to find a legitimate website asking users tosubmit their personal information through a pop-up window.

4. IFrame Redirection

IFrame is an HTML tag used to display an additional webpage into one that is currently shown.

D. Domain Based Features

1. Age of Domain

This feature can be extracted from WHOIS database. Most phishing websites live for a short period of time. By reviewing our dataset, we find that the minimum age of the legitimate domain is 6 months.

2. DNS Record

For phishing websites, either the claimed identity is not recognized by the WHOIS database or no records founded for the hostname. If the DNS record is empty or not found then the website is classified as “Phishing”, otherwise it is classified as “Legitimate”.

3. Website Traffic

This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit.

4. Page Rank

PageRank is a value ranging from “0” to “1”. PageRank aims to measure how important a webpage is on the Internet.

5. Google Index

This feature examines whether a website is in Google’s index or not. When a site is indexed by Google, it is displayed on search results.

6. Number of Links Pointing to Page

The number of links pointing to the webpage indicates its legitimacy level, even if some links are of the same domain.

7. Statistical-Reports Based Feature

Several parties such as PhishTank formulate numerous statistical reports on phishing websites at every given period of time; some are monthly and others are quarterly.

I. INVESTIGATION ON RESEARCH GAPS

So far we have understood that phishing is a specialized social engineering attack whereby the attacker very intelligently uses spoofed emails or websites to trick the victims into sharing their confidential and sensitive information. There is a need to understand the psychology of online consumers that whether they are concerned about the security issues when they are having the authority to change the security features. There are many academic literatures about security against phishing. However, there are a number of issues that concern the gap between academic literature andpractical evidence.

A major research gap exists between research and the industry “in terms of true positives”. While academic and literary research essentially focuses on machine-learning and heuristics, assuming very good true positives, these true positives are sometimes high false positives. Hence, these heuristics are only reasonable enough to identify phishing sites that have not been encountered before. However, the industry primarily relies on blacklists for classification of phishing websites. But, the blacklists fail to generalize to the future unseen cases and are also potentially slow in responding to zero-hour attacks.

II. CONCLUSION AND FUTURE WORK

Phishing has becoming a serious network security problem, causing finical lose of billions of dollars to both consumers and e-commerce companies. And perhaps more fundamentally, phishing has made e-commerce distrusted and less attractive to normal consumers. In this paper, we have studied the characteristics of the hyperlinks that were embedded in phishing e-mails. We then designed an anti-phishing algorithm, Link Guard, based on the derived characteristics. Since Phishing Guard is characteristic based, it can not only detect known attacks, but also is effective to the unknown ones.

We have implemented Link Guard for Windows XP. Our experiment showed that Link Guard is light-weighted and can detect up to 96% unknown phishing attacks in real-time. We believe that Link Guard is not only useful for detecting phishing attacks, but also can shield users from malicious or unsolicited links in Web pages and Instant messages. Our future work includes further extending the Link Guard algorithm, so that it can handle CSS (cross site scripting) attacks.

FUTURE ENHANCEMENT

Our future work includes further extending the Link Guard algorithm, so that it can handle Cross Site Scripting (CSS) attacks. **Cross-site scripting** (**CSS**) is a type of computer security vulnerability typically found in web applications which allow code injection by malicious web users into the web pages viewed by other users. Examples of such code include HTML code and client-side scripts. An exploited cross-site scripting vulnerability can be used by attackers to bypass access controls such as the same origin policy. Vulnerabilities of this kind have been exploited to craft powerful phishing attacks and browser exploits. Cross-site scripting was originally referred to as **CSS**, although this usage has been largely discontinued due to the confusion with cascading style sheets.

REFERENCES

[1] Yasin Sönmez, Türker Tuncer, Hüseyin Gökal and Engin Avci, "Phishing Web Sites Features Classification Based on Extreme Learning Machine", International Symposium on Digital Forensic and Security (ISDFS), May 2018.

[2] Anjum N. Shaikh, Antesar M. Shabut and M. A. Hossain, " A literature review on Phishing Crime, Prevention Review and Investigation of gaps", 2016 10th International Conference on Software, Knowledge, Information Management & Applications.

[3] Neda Abdelhamid, Fadi Thabtah and Hussein Abdel-jaber, "Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features", IEEE Int. Conf. on Intelligence and Security Informatics (ISI), pages 72–77, 2017.

[4] Muhemmet Baykara, Zahit Ziya Gürel, Detection of phishing attacks, 2018.

[5] Tianrui Peng, Detecting Phishing Attacks Using Natural Language Processing and Machine Learning , 2018.

[6] Jhen-Hao Li, PhishBox: An Approach for Phishing Validation and Detection, 2017.

[7] Naghmeh Moradpoor, Employing Machine Learning Techniques for Detection and Classification of Phishing Emails, July 2017.

[8] R.Aravindhan, Dr.R.Shanmugalakshmi, Certain Investigation on Web Application Security: Phishing Detection and Phishing Target Discovery, January 2016.

[9] Mustafa Aydin, Nazife Baykal, Feature Extraction and Classification Phishing Websites Based on URL, 2015.

[10] Samuel Marchal, Radu State, Jerome Francois, and Thomas Engel, PhishStorm: Detecting Phishing with Streaming Analytics, December 2014.