Volume XXIII 2020
ISSUE no.1
MBNA Publishing House Constanta 2020

Scientific Bulletin of Naval Academy

SBNA PAPER • OPEN ACCESS

Detection of phishing attacks using the anti-phishing framework

To cite this article: Dragoș Glăvan, Ciprian Răcuciu, Radu Moinescu and Sergiu Eftimie, Scientific Bulletin of Naval Academy, Vol. XXIII 2020, pg.208-212.

Available online at www.anmb.ro

ISSN: 2392-8956; ISSN-L: 1454-864X

doi: 10.21279/1454-864X-20-I1-028

Detection of Phishing attacks using the anti-phishing framework

Dragoș GLĂVAN, Ciprian RĂCUCIU, Radu MOINESCU, Sergiu EFTIMIE

Military Technical Academy "*Ferdinand I*" – Systems Engineering for Defense and Security Doctoral School
dragos.glavan@gmail.com

Abstract. In the area of computer security, phishing is a form of criminal activity that involves obtaining sensitive data, such as access data for banking applications, e-commerce applications (such as eBay or PayPal) or credit card information, using techniques to manipulate the identity data of a person or institution. A phishing attack consists of sending the attacker an electronic message, using instant messaging software or telephone, in which the user is advised to give his confidential data to win certain prizes, or is informed that they are prizes. necessary due to technical errors that led to the loss of the original data. According to the reports of the Anti-Phishing Working Group (APWG) published in December 2018, phishing against banking and the payment processor was high. Almost all phishy URLs use HTTPS and use redirects to avoid detection. This paper is a comparative survey of the available methods for detecting phishing sites. There has been a comparative study of anti-phishing tools and their limitations have been found. An anti-phishing model, a framework to help detect these attacks, is discussed. The identification of this type of attack is a very important problem considering that, at present, people carry out many online transactions regarding money transfer, payment of bills or purchases.

1. Introduction

Phishing is an online theft of personal information and credentials of the user; it is a type of fraud through which the attacker gains access to the user's private information. An expert designer can create a fake website, identical to the original one, identifying it as fake can be a difficult procedure, so people fall into the trap of the attacker. These fake websites ask the user for credentials, claiming to be reliable (for example, by using HTTPS), this can convince a user to trust this site. Lately, the number of online transactions has increased greatly through the payment of online invoices, purchases or money transfers, so identifying phishing sites is a very important thing (in 2018 approximately 647592 type sites were reported. phishing). The figure below illustrates how a user accesses a fake site, identical to the original one, the user accesses a phishy URL that I consider legitimate and transmits the requested details, so the attacker comes in possession of the passwords they might have use them for harmful purposes.

This paper presents methods for detecting phishing sites, research shows that there are approximately 5 approaches based on:

- Rule based or Heuristics based approach;
- Blacklisting approach;
- Content based approach;
- Machine Learning based approach;
- Hybrid approach.

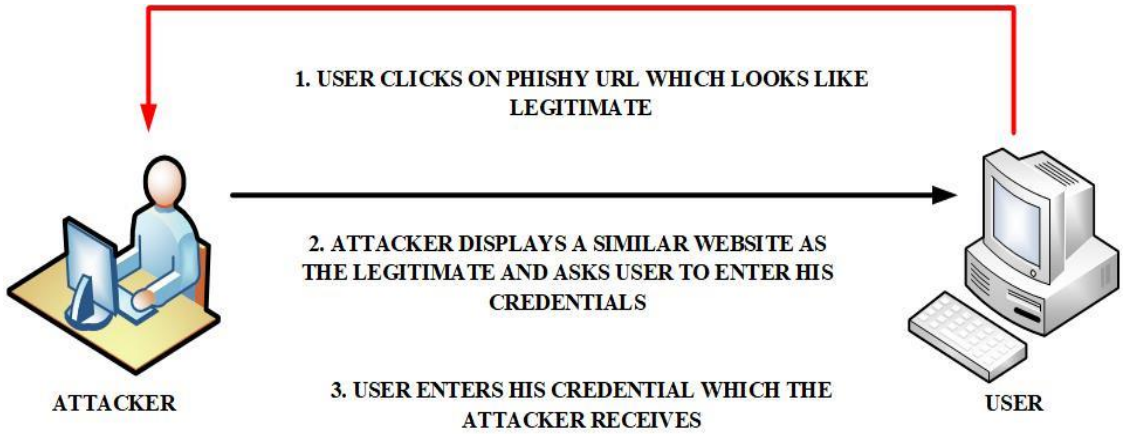

Fig.1. Web spoofing attack

2. Background

2.1. Anti-Phishing Solutions

In the last period there has been a significant increase in the number of phishing attacks, so we tried to find solutions to solve these problems. Anti-phishing solutions can be classified as follows:

- Heuristic Based Approach - This technique uses heuristics to classify URLs, these are the features that are required to verify a website;
- Content Based Approach - This technique is based on comparing the terms from the original site with those from the phishing site. Another approach is capturing images from the original website and processing them to compare them with a site similar to the original one. The information obtained from the screenshots after processing can verify the legitimacy of the site with the help of a search engine by purchasing the content, the logo can be used in this method to analyze the web page;
- Blacklist Based Approach - The blacklist includes a list of websites declared spam, which are generally administered by organizations such as Google. A major disadvantage of this procedure is the possibility that a newly created phishing address will not be blacklisted, so the URLs will be undetected, the blacklisted URLs will deny access;
- Machine Learning Approach - Within this technique the characteristics are extracted using the automatic learning techniques, the accuracy depends to a large extent on the chosen algorithm;
- Hybrid Approach - This technique involves combining several techniques to detect whether a website is fake or real (for example, blacklisting and heuristics can be combined).

Numerous anti-phishing tools are currently available to help protect against phishing sites, for example Google Chrome and Mozilla Firefox use Google Safe Browsing to block phishing sites. Google Safe Browsing uses the blacklist approach to parse a URL.

3. Features

In order to detect phishing attacks, web sites are examined using their functions, they play an important role in designing a system that is capable of detecting phishing sites (there are different types of features such as: based on JavaScript or HTML, address bar, domain). The table below includes the characteristics and their description. According to the Request for Comments (RFC) community of engineers, we affirm that special characters '@', '~', '#' are safe and can be used within URLs. Another feature says that if a URL is longer than 54, it is suspected of not being legitimate, but some phishing URLs have a length of less than 54, so it's not a criterion for us guide to discovering phishing sites. Thus, it was concluded that an address greater than 1750 is almost certainly a phishing address.

FEATURE	DESCRIPTION
Using IP address in DN (domain name)	the attacker tries to hide the name with numbers
LongURL	is used to hide keywords that are considered suspicious
'-'symbol	is used to describe it as a legitimate URL
Sub-domain(s) inURL	several sub-domains are considered phishing
Use of HTTPS	secure connection using HTTPS
Request URL	all text / images must be uploaded from the URL domain
Using pop-up Window	using the pop-up window for password entry is not an ethical way
DNS record	the URL is considered phishy if DNS record is not available
Web site traffic	the site is phishy if there are no records of website traffic
Age of domain	the site is considered phishing if the site registered for less than 1 year

Table.1. Features

4. Architecture

In order to design a model capable of detecting phishing, an appropriate way of designing a system should be adopted that will give correct and precise results, in the figure below are presented the stages and the functioning of each one.

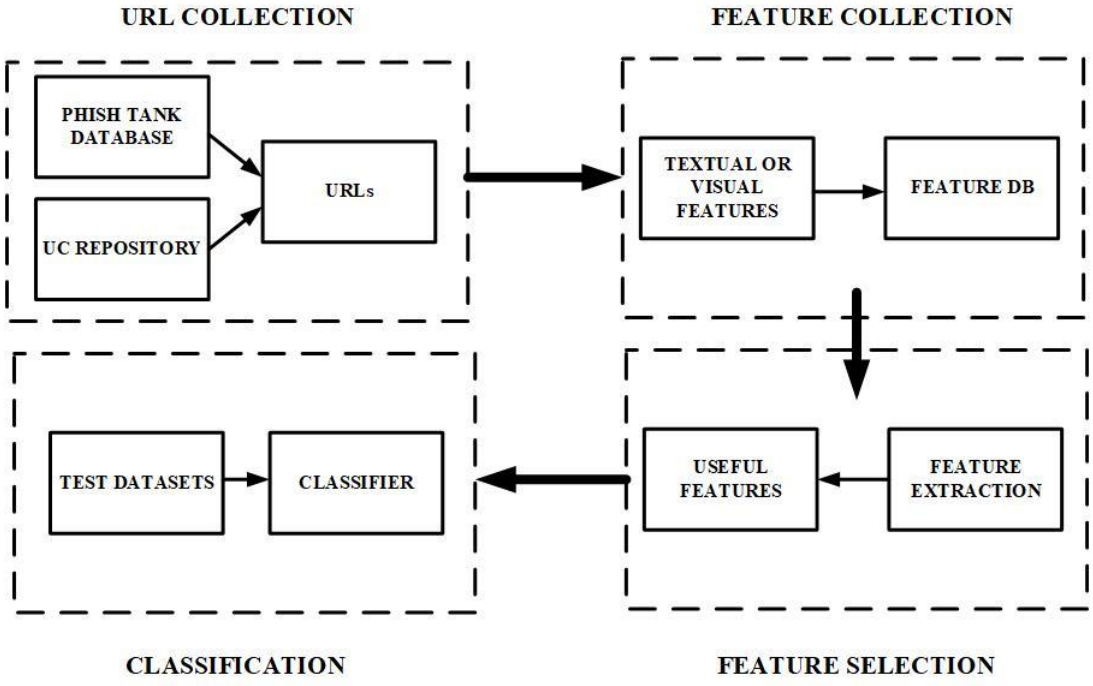

Fig.2. System flow of detection

URL collection – in this case, the entry into the system is the URL of the website to be verified. URLs are collected from sources such as PhishTank, this is a site where suspicious URLs are found and are polled whether or not they are spam. Here you can find even the sites declared as finding phishing sites. There is also a database called Alexa, this is where most authors collect legitimate URLs. A correct set of data will result in a correct model, if a correct set of data is not used, the trained model will not be efficient enough and thus the wrong results will be obtained.

Feature Collection – while creating your own data set, elements such as URL length, IP address, domain, DNS, etc. are collected, features that are saved and processed. Mohamed used some rules to generate a dataset with binary or ternary values, where 0.1 or -1 are values for each URL where: 1 is phishing, 0 is suspicious and -1 is legitimate. This step is performed and in the next stage in frames the URL is transmitted to the model for validation, so this step becomes necessary only in the testing phase.

Feature Selection – characteristic selection is a very important step to filter the characteristics of the institutions that contribute very little to the classification process, thus increasing the falsification rate and thus reaching incorrect results. Methods such as Consistency subset or Correlation Feature Selection are used for feature selection.

Classification – in this final stage, several classification algorithms are used such as Random Forest or Naïve Bayes, automatic learning algorithms and JRIP or PRISM which are associative classification techniques. To design an efficient system from all points of view, choosing the classification algorithm is a very important step.

5. Results

They studied about all the different types of approaches available for phishing detection as well as the currently used Anti-Phishing tools. It turned out that the Blacklisting spam detection technique is less useful with regard to recently registered domains. False websites created can be generated every day, if their URLs are not blacklisted, these websites cannot be detected. A heuristic approach can fail when there is no rule for a particular attribute. Due to this fact, that attribute remains undetected and, therefore, we must ensure that all rules or heuristics are added to the system. Research shows that heuristics and hybrid approaches are the most effective. Although the accuracy of the Heuristics approach is high, the possibility of this method of not being able to analyze a spam feature recently added to the website can lead to the classification of the site as legitimate. The model that uses the Machine Learning approach offers good accuracy and uses a total of 30 features, several machine learning algorithms have been used, and Random Forest has proven to be the best.

At the same time, in the case of hybrid approaches, an important number of functions that play an important role must be considered while designing an effective model for detecting a phishing website. The other important thing is the training set or the data set that is taken into account during the experimentation. A dataset should consist of some spam URLs as well as legitimate URLs. A system must use a data set with different types of instances that covers all the features of the verified website. No feature should be left unchecked to ensure complete verification of the URL. Feature selection must be done to minimize processing and improve the training process. A classification algorithm should be chosen to provide the best accuracy. With each new solution, attackers find new ways to create a spam website.

6. Conclusions

Phishing attacks become a permanent threat to this rapidly developing technology world. Nowadays, everyone is targeting online transactions without cash, online business etc. Any non-specialist individual who does not fully know how to identify a Security threat will never take the risk of conducting online monetary transactions. This paper analyzed various anti-phishing methods that were discussed to give a clear idea of the existing techniques. At the same time, the idea was stressed that the general public should be aware of phishing attacks and the use of anti-phishing tools during web browsing. We have presented the most important steps to be taken in building an efficient anti-phishing model that is responsive to today's technology.

References:

[1] A. Ahmed, "*Real Time Detection of Phishing Websites*", Electronics and Mobile Communication Conference ,2016

[2] Naga Venkata Sunil, "*A PageRank Based Detection Technique for Phishing Web Sites*", 2012

[3] H. Shahriar, "*Information Source-based Classification of Automatic Phishing Website Detectors*", 2011

[4] E. H. Chang and K. L. Chiew, "*Phishing Detection via Identication of Website Identity*", 2013

[5] M. Aydin and N. Baykal, "*Feature Extraction and Classification Phishing Websites Based on URL*", 2015

[6] S. Asghar and S. Gillani "*A Hybrid Model to Detect Phishing-Sites using Supervised Learning Algorithms*", 2016

[7] R. M. Mohammad, "*Intelligent Rule Based Phishing Website Classification*", 2014