International Journal of Innovative and Emerging Research in Engineering
Volume 3, Special Issue 1, ICSTSD 2016

A Review on Useful Features for Introducing Accuracy in Phishing Website Detection

Mr. Shailesh P. Thakare[1], Mr. Shrikant N. Sarda[2], Mr. Nitin M. Shivratriwar[3]

spthakare82@gmail.com[1]
shrikantnsarda@gmail.com[2]
nitinshivratriwar@rediffmail.com[3]
[1, 2, 3] Assistant Professor, Department of Information Technology
Prof. Ram Meghe Institute of Technology & Research, Badnera - Amravati

***Abstract*— phishing is the act of getting someone's personal details directly or indirectly and use for some malicious purpose. Phishers get successful because there is no concrete mechanism to detect and avoid it. Now a day's phishers use online resources to get personal details of the target persons. Due to wider use of social media and availability of personal information on internet it is very easy for phishers to get such details. For these purpose they create fake websites which exactly looks like authentic websites and send links of fake websites to target users so that they can visit it and reveal their personal information then this information is used by phishers for some unethical purpose. We carried out some survey on it and we come to know that there are some standard features associated with standard or authentic websites which are absent in fake or phishing website if we make use of these features then it is helpful for detecting and avoiding phishing websites.**

***Index Terms*—Phishing, Phishing Websites, Domain Based features, Address bar Features, Java script based features.**

I. INTRODUCTION

Now a day's almost all of us use the internet or some online resources for some purpose and reveal our personal details on many websites directly or indirectly by filling form online. But most of time peoples are not aware about some tricks and tactics used by phishers to collect personal details online and they reveal their information. Normal users can't tolerate the difference between phishing websites and authentic website due to unawareness about phishing or new tricks and tact's used by phisher. Phishers creates fake websites and convince target peoples to access and send their private information such as usernames, passwords and credit card resulting in stealing their information. Many research works have been proposed to train applications using data mining and machine learning to detect phishing sites using some basic features of websites and WebPages. These features are commonly divided into four categories as: Address Bar based Features, Abnormal Based Features, HTML and JavaScript based Features, Domain based Features. These features have some standards or normal values associated with it for authentic website if we consider these threshold values for some important features then it will gives us accurate result. This article focuses on about such features [1, 2].

II. PHISHING WEBSITES

Phishing is a new form of internet crime as compare to other like, virus and hacking. Now a day's many phishing web pages are available on internet. Its impact is the breach of Information security through the compromise of confidential data and the victims may finally suffer losses of valuable resources. A phishing website launched social engineering attack that attempts to defraud people of their personal information including credit card number, bank account information, social security number, and their personal credentials in order to use these details fraudulently against them. Phishing has a huge negative impact on organizations' revenues, customer relationships, marketing efforts, and overall corporate image.

Definition:
Since phishing is continuously evolving with new tricks and trends there are many definitions of phishing website. One definition by Anti-Phishing Working Group (APWG)'s is, "Phishing attacks use both social engineering and technical subterfuge to steal consumers' personal identity data and financial account credentials". Typically a phishing attack is a combination of fraudulent emails, spoofed websites, and identity theft. Internet users or customers of many banks and financial institutions are the targets of phishing attacks.

Phishing websites use a number of different techniques to hide the fact that they are not authentic. In practice, these tricks make it extremely difficult for the general user to distinguish a phishing site from an authentic one.

Phishing websites are fake web pages created by phishers to mimic web pages of authentic websites. Most of these types of web pages have similarities to cheat their targets. Victims of phishing web pages may reveal their bank account, password,

credit card number, or other important information to the phishing website [1, 2, 3, 4].

III. PROBLEMS AND CHALLENGES

Detection of phishing website is a very challenging and complex task, since it contains technical as well as social issues for which there is no known methodology to solve it completely. There are many applications available for phishing website detection, many of these uses soft computing methods to give the accuracy in detection. Most of these applications use some data sets to test and train their systems by considering this we are providing some efficient features which will helpful for accuracy in detection.

IV. PHISHING WEBSITE FEATURES

Due to uncertain nature, new tricks and tactics used by phishers, researchers in this area face many problems. There are many articles are available on phishing website detection but it is very complicated to gather required features from them and continue. Today's most of projects uses soft computing methods like fuzzy logic, data mining, neural network etc. these methods requires datasets for training and testing purpose. But due to rapid change in nature of attacks and available datasets (outdated) such systems will not produce accurate results. So if researchers concentrate on basic features then systems will give more accurate results. The aim of this article is to focus on some features which will helpful for detection of phishing websites and will be helpful for new researchers and students.

Useful features for detection of Phishing website:

There are many features which are useful for deciding whether a particular website is normal or phishing one. These features are classified into four groups as follows

FEATURE	CATEGORY
1. Using the IP Address 2. URL-Length 3. Shortining-Service 4. having-At-Symbol 5. double-slash-redirecting 6. Prefix-Suffix 7. having-Sub-Domain 8. SSL final-State 9. Domain-registration-length 10. Favicon 11. port 12. HTTPS-token	Address Bar based Features
1. Request-URL 2. URL-of-Anchor 3. Links-in-tags 4. SFH 5. Submitting-to-email 6. Abnormal-URL	Abnormal Based Features
1. Redirect 2. on-mouseover 3. RightClick 4. popUpWidnow 5. Iframe	HTML and JavaScript based Features
1. age-of-domain 2. DNSRecord 3. web-traffic 4. Page-Rank 5. Google-Index 6. Links-pointing-to-page 7. Statistical-report	Domain based Features

1. **Address Bar based Features:** This will contain all features associated with the address bar. Features like IP address of the domain name in the URL, URL length, URL with @ symbol, the existence of □ in the domain name part of the URL, the usage of https and issuer of the website, the expiration of the domain, open ports, Favicon (graphic image (icon) associated with a specific webpage) and the existence of HTTPS Token in the Domain Part of the URL.

2. **Abnormal Based Features:** This will contain all features related with the abnormal behaviors of the website. Features like, examining whether a webpage contains external objects such as images, if the <a> tags and the website have different domain names, if <Meta>, <Script> and <Link> tags linked to the same webpage, examining if the Server Form Handler (SFH) is an empty string or blank, if the mail()" or "mailto:" functions are used in the source code of the webpage to submit user's information to phisher's personal email and if the host name is included in the URL of the examined website.

3. **HTML and JavaScript based Features:** This will contain all features related with HTML and JavaScript source code of the WebPages included in the examined website. features like, how many times a website has been redirected, if a fake URL in the status bar is shown to the users, if the right click function is disabled to prevent the users from viewing and saving the webpage source code, if website asks user to submit her/his personal information through a pop-up window, and if the IFram HTML tag is used to display an additional webpage into the currently one shown.

4. **Domain based Features:** This will contain all features of the domain part in the URL of the website. Features like checking if the websites live for a short period of time, if no DNS record (mapping between IP address and the domain associated with) exists for the domain, popularity of the website, the page rank, if the website is indexed by

Google, number of links pointing to the webpage, and if the website is reported as a phishing site by several parties that track phishing such as phishTank and StopBadware. [5, 6, 7]

1. Address Bar based Features

a. Using the IP Address

One may use an IP address as an alternative of domain name in URL, like "http://123.92.1.123/abcd.html". Sometimes, the IP address is converted into hexadecimal code as "http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".

b. Long URL to Hide the Suspicious Part

Sometimes phishers can use long URL to hide the doubtful part in the address bar. Like, http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html Generally in most of the cases size of URL is fixed i.e up to some characters long, if this size is greater than expected then there may be a abnormality in that URL.

c. 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. This is accomplished by means of an "HTTP Redirect" on a domain name that is short, which links to the webpage that has a long URL. For example, the URL "http://portal.hud.ac.uk/" can be shortened to "bit.ly/19DXSk4".

d. 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.

e. Redirecting using "//"

The existence of "//" within the URL path means that the user will be redirected to another website. An example of such URL's is: "http://www.legitimate.com//http://www.phishing.com". We examin the location where the "//" appears. We find that if the URL starts with "HTTP", that means the "//" should appear in the sixth position. However, if the URL employs "HTTPS" then the "//" should appear in seventh position.

f. Adding Prefix or Suffix Separated by (-) to the Domain

The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example http://www.Confirme-paypal.com/.

g. Sub Domain and Multi Sub Domains

Consider http://www.hud.ac.uk/students/. A domain name might include the country-code top-level domains (ccTLD), which in our example is "uk". The "ac" part is shorthand for "academic", the combined "ac.uk" is called a second-level domain (SLD) and "hud" is the actual name of the domain. To produce a rule for extracting this feature, we firstly have to omit the (www.) from the URL which is in fact a sub domain in itself. Then, we have to remove the (ccTLD) if it exists. Finally, we count the remaining dots. If the number of dots is greater than one, then the URL is classified as "Suspicious" since it has one sub domain. However, if the dots are greater than two, it is classified as "Phishing" since it will have multiple sub domains. Otherwise, if the URL has no sub domains, we will assign "Legitimate" to the feature.

h. 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 not enough. Generally the minimum age of a reputable certificate is of two years.

i. 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.

j. Favicon

A favicon is a graphic image (icon) associated with a specific webpage. Many existing user agents such as graphical browsers and newsreaders show favicon as a visual reminder of the website identity in the address bar. If the favicon is loaded from a domain other than that shown in the address bar, then the webpage is likely to be considered a Phishing attempt.

k. Using Non-Standard Port

This feature is useful in validating if a particular service (e.g. HTTP) is up or down on a specific server. In the aim of controlling intrusions, it is much better to merely open ports that you need. Several firewalls, Proxy and Network Address Translation (NAT) servers will, by default, block all or most of the ports and only open the ones selected. If all ports are open, phishers can run almost any service they want and as a result, user information is threatened. The most important ports and their preferred status are shown in Table 2.

Table 1. Common ports to be checked

PORT	Service	Meaning	Preferred Status
21	FTP	Transfer files from one host to another	Close
22	SSH	Secure File Transfer Protocol	Close
23	Telnet	provide a bidirectional interactive text-oriented communication	Close
80	HTTP	Hyper test transfer protocol	Open
443	HTTPS	Hypertext transfer protocol	Open

		secured	
445	SMB	Providing shared access to files, printers, serial ports	Close
1433	MSSQL	Store and retrieve data as requested by other software applications	Close
1521	ORACLE	Access oracle database from web.	Close
3306	MySQL	Access MySQL database from web.	Close
3389	Remote Desktop	allow remote access and remote collaboration	Close

l. 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. For example, http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/. [6 , 7, 8, 9]

2. Abnormal Based Features

a. Request URL

Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain. In legitimate webpages, the webpage address and most of objects embedded within the webpage are sharing the same domain.

b. URL of Anchor

An anchor is an element defined by the <a> tag. This feature is treated exactly as "Request URL". However, for this feature we examine:

1. If the <a> tags and the website have different domain names. This is similar to request URL feature.

2. If the anchor does not link to any webpage, e.g.:

A. <a href="#">
B. <a href="#content">
C. <a href="#skip">
D. <a href="JavaScript ::void(0)">

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.

c. Server Form Handler (SFH)

SFHs that contain an empty string or "about:blank" are considered doubtful because an action should be taken upon the submitted information. In addition, if the domain name in SFHs is different from the domain name of the webpage, this reveals that the webpage is suspicious because the submitted information is rarely handled by external domains.

d. 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. To that end, a server-side script language might be used such as "mail()" function in PHP. One more client-side function that might be used for this purpose is the "mailto:" function.

e. Abnormal URL

This feature can be extracted from WHOIS database. For a legitimate website, identity is typically part of its URL. [6, 7, 8, 9,]

3. HTML and JavaScript based Features

a. Website Forwarding

The fine line that distinguishes phishing websites from legitimate ones is how many times a website has been redirected. In our dataset, we find that legitimate websites have been redirected one time max. On the other hand, phishing websites containing this feature have been redirected at least 4 times.

b. Status Bar Customization

Phishers may use JavaScript to show a fake URL in the status bar to users. To extract this feature, we must dig-out the webpage source code, particularly the "onMouseOver" event, and check if it makes any changes on the status bar.

c. 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". Nonetheless, for this feature, we will search for event "event.button==2" in the webpage source code and check if the right click is disabled.

d. Using Pop-up Window

It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window. On the other hand, this feature has been used in some legitimate websites and its main goal is to warn users about fraudulent activities or broadcast a welcome announcement, though no personal information was asked to be filled in through these pop-up windows.

e. IFrame Redirection

IFrame is an HTML tag used to display an additional webpage into one that is currently shown. Phishers can make use of the "iframe" tag and make it invisible i.e. without frame borders. In this regard, phishers make use of the "frameBorder"

attribute which causes the browser to render a visual delineation. [6, 7, 8, 9]

4. Domain based Features

a. Age of Domain

Most phishing websites live for a short period of time. Generally minimum age of the legitimate domain is 6 months.

b. DNS Record

If the DNS record is empty or not found then the website is classified as "Phishing", otherwise it is classified as "Legitimate".

c. Website Traffic

This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit. If the domain has no traffic or is not recognized by the Alexa database, it is classified as "Phishing". Otherwise, it is classified as "Suspicious".

d. PageRank

PageRank is a value ranging from "0" to "1". PageRank aims to measure how important a webpage is on the Internet. The greater the PageRank value the more important the webpage. Most of phishing webpages have no PageRank.

e. 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. Usually, phishing webpages are merely accessible for a short period and as a result, many phishing webpages may not be found on the Google index.

f. Number of Links Pointing to Page

The number of links pointing to the webpage indicates its legitimacy level. Most of phishing dataset items have no links pointing to them. On the other hand, legitimate websites have at least 2 external links pointing to them.

g. Statistical-Reports Based Feature

Several parties such as PhishTank and StopBadware formulate numerous statistical reports on phishing websites at every given period of time; some are monthly and others are quarterly. . [6, 7, 8, 9,]

V. CONCLUSION

Phishing websites have some abnormal values with its basic features. By considering such features if we develop system then system will detect fake website more accurately. Although phishers are getting successful because lack of knowledge about it within people so awareness and education about phishing is necessary to stop it.

REFERENCES

[1] Minal Chawla, Siddarth Singh Chouhan, "A Survey of Phishing Attack Techniques" *International Journal of Computer Applications (0975 – 8887) Volume 93 – No 3, May 2014* 32.

[2] Nirmala Suryavanshi, Anurag, " A Review of Various Techniques for Detection and Prevention for Phishing Attack", Jain *International Journal of Advanced Computer Technology (IJACT)* ISSN:2319-7900.

[3] Gaurav Kumar Chaudhary, "Development *Review on Phishing: A Computer Security Threat* "*International Journal of Advance Research in Computer Science and Management Studies Volume 2, Issue 8, August 2014 pg. 55-64,* ISSN: 2327782 (Online)

[4] Atul M. Tonge , Surbhi R. Chaudhari, "Phishing Susceptibility and Anti-Phishing Security Strategies-Literature Review ",International Journal of Scientific & Engineering Research, Volume 4, Issue 12, December-2013 67 ISSN 2229-5518

[5] Qingxiong Ma "The process and characteristics of phishing attacks: A small international trading company case study", Journal of Technology Research

[6] Doaa Hassan, "On Determining the Most Effective Subset of Features for Detecting Phishing Websites", International Journal of Computer Applications (0975 - 8887), Volume 122 - No.20, July 2015

[7] Rami M. Mohammad, Fadi Thabtah, Lee McCluskey, "Intelligent Rule based Phishing Websites Classification".

[8] A.Martin, Na.Ba.Anutthamaa, M.Sathyavathy, Marie Manjari Saint Francois, Dr.Prasanna Venkatesan, "A Framework for Predicting Phishing Websites Using Neural Networks", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011, ISSN (Online): 1694-0814, www.IJCSI.org

[9] Maher Aburrous, M.A. Hossain, Keshav Daha, Fadi Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining", Expert Systems with Applications 37 (2010) 7913–7921, journal homepage: www.elsevier.com/locate/eswa