Volume 4, Issue 3, March 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Protection from Phishing Attacks by Exploiting Page Rank, Reputation and Source Code of the Webpage

Radheshyam Panda, Rajesh Tiwari
Dept.of CSE, CSVT University,
Bhilai, India

Abstract— Phishing is a type of Internet fraud that tries to find out user's confidential credentials by deception using fraud websites, such as credit card numbers, passwords, banking account details and other confidential information. There are some techniques that distinguish phishing websites that violate the w3c standards, from legitimate websites. In this paper, we propose a phishing detection approach based on analysis of page rank, reputation and the source code of the webpage. We extract page rank and reputation of the web page using some search engines and some phishing characteristics out of the W3C standards to evaluate the security of the websites. If the page rank and reputation of the website is very high, the webpage can be concluded as a legitimate web page. During the process of checking the security level of a webpage we also check each character of the webpage's source code if required. In this process if we find a phishing character, we will decrease from the initial secure weight. Finally we calculate the security percentage based on the final weight, the high percentage indicates secure website and others indicates the website is most likely to be a phishing website. Our approach can detect the phishing website based on analysis of page rank, reputation and phishing characteristics of the webpage's source code. In this paper we present an approach to overcome the intricacy and complexity in detecting and predicting phishing websites.

Keywords— PageRank, Reputation, SFH, spoof index, SVM, W3C,

I. INTRODUCTION

Online services have become a great convenience nowadays but at the same time it has brought us some new threats. Phishing is one of them executed by spoofed email or by instant messages impersonating from legitimate sources and are used to entice recipients to click the contained link that leads to forged website to trap them into disclose their confidential information. Phishing is a technique to acquire sensitive information such as usernames, passwords and credit card details by impersonating as a truthful entity in an electronic communication. This is analogous to Fishing, where the fisherman puts bait at the hook, thus, pretending to be a genuine food for fish. But the hook inside it takes the complete fish out of the water. Communications claiming to be from popular social web sites, online payment processors, auction sites, or IT administrators are commonly used to entice the unsuspecting public [1]. Phishing is usually carried out by e-mail spoofing or direct messaging and it often directs users to enter details at a fake website whose look and feel are almost like the legitimate website. It is an example of social engineering techniques used to mislead the users [2] and takes advantage of the poor usability of current web security technologies [3]. Once a user visits the phishing website the ruse is not over. In some phishing scams JavaScript commands are used to alter the address bar. This is done by placing a picture of a legitimate URL on the address bar. Flaws in a trusted website's own scripts can also be exploited against the victim by the attacker. These kinds of attacks are known as cross site scripting and are very problematic. In these attacks the user is directed to log in at his/her bank or service's own page where each and everything from web address to security credential appears legitimate.

Nowadays social networking sites have become the primary target of phishing, because the personal details in such site can be used in identity theft. A success rate of over 70% for phishing attacks on social site is shown by the experiments. There exists some anti-phishing websites that publish exact messages that have been recently circulated in the internet, for example Millersmiles and FraudWatch International. These sites provide specific details about the message. [1]

To prevent phishing attacks, some effort has been directed towards phishing detection. Various approaches have been suggested. Detection of phishing can be broadly classified into two categories: Heuristic-Based and List-Based. Approaches based on heuristic check the characteristics of a website for phishing detection. These characteristics can be the hypertext markup language (HTML) code, page content or uniform resource locator. Generally heuristics are targeted at the HTML sourced code.

Anti-phishing approaches based on lists are widely used today. Classification of a website as a trusted or phishing is a simple database look up. List based approaches can again be classified: Blacklist [4] and Whitelist [5].

The websites, considered as a phishing website are held in blacklist. Whenever a page is loaded by the browser, a query goes to the blacklist to determine if the currently visited URL is in the list. If it exists, proper countermeasure can be taken. Otherwise it will be considered as a legitimate webpage. The blacklist can be hosted at a central server or it can

be stored locally at the client. On the other hand a list of trusted website is known as whitelist. The basic concept behind this approach is creating a list of websites that the user accesses on a daily basis. Navigation to a website that is not in the trusted list will either be failed or prompt a warning message.

In this paper we discuss about mixed approach i.e. combination of list based and heuristics approach along with the exploitation of PageRank and Reputation of the webpage to detect phishing website.

II. RELATED WORK

Phishing website has a huge adverse effect on the online commerce and finance, detecting and preventing these attacks is an important step towards saving the online business. There are several approaches have been developed to detect these attacks. In this section, we analyse some existing anti-phishing mechanisms.

A. Intelligent Phishing Website Detection System e

It is one of the anti -phishing approach based on Fuzzy Techniques [6] and turn out six criteria of website phishing attack. There are several characteristics and factors that can distinguish the forged faked phishing website from original legitimate website like long URL address and abnormal DNS record, spelling errors etc. Phishing detection rate of a website is obtained based on six criteria and there are different components for each criterion, the criteria are as follows:

1. *Encryption & Security.*
   - Certification authority.
   - Using SSL certificate.
   - Distinguished Names Certificate (DNC).
   - Abnormal cookie.
2. *Identity of URL & Domain*
   - Using the IP address.
   - Abnormal URL.
   - Abnormal request URL.
   - Abnormal URL of anchor.
   - Abnormal DNS record.
3. *Contents &Page Style.*
   - Copying website.
   - Spelling errors.
   - Using forms with "Submit" button.
   - Disabling right click.
   - Using Popup windows.
4. *Java script &Source Code.*
   - Redirect pages.
   - Straddling attack.
   - Server Form Handler (SFH).
   - Using onMouseOver to hide the Link.
5. *Social Human Factor.*
   - Public generic salutation.
   - Much emphasis on security and response.
   - Buying Time to Access Accounts.
6. *Web Address Bar.*
   - Adding a prefix or suffix.
   - Long URL address.
   - Using hexadecimal character codes.
   - Replacing similar characters for URL.
   - Using @ symbol to confuse.

The rule base has some input parameters (criterion) and one output that contain all the "IF-ELSE" rules of the system. For each criterion, the output is one of the following:

1. *Genuine*
2. *Doubtful*
3. *Fraud*

The final output is one of the following:

1. *Very Legitimate*
2. *Legitimate*
3. *Suspicious*
4. *Phishy*
5. *Very Phishy*

This represents final phishing website rates.

B. Client-side defence against web-based identity theft [7]

It requires a framework for client-side defence. In this approach a browser plug-in is needed to examine WebPages and warns the user for phishing attack. A spoof index (a measure of the likelihood that a specific page is part of a spoof attack) is computed by browser plug-in and user is warned about the attack if the index value exceeds a particular level. The browser plug-in is called SpoofGuard. A combination of page evaluation and examination of outgoing post data is used to measure the spoof index. When the username and password is entered by the user to a spoof website that contains misleading domain name, suspicious URL, image from a legitimate website or a username and password that has been used previously at a legitimate website, SpoofGuard will capture the post and notify the user with a popup that prevents the attack.

Some common parameters are used to determine spoof website, they are Suspicious URLs, Logos, User input, Copies, Short lived and Sloppiness or lack of familiarity with English and HTTPS. The test is applied by the browser plug-in on the basis of above parameter using a scoring mechanism to compute spoof index. On the basis of total index value the browser plug-in warns the user and decides the severity and type of warning.

C. Anomaly Based Web Phishing Page Detection [8]

In this approach the anomalies of the webpage are examined, particularly, mismatch between the identity of the webpage and its structural feature along with HTTP transaction. A standard webpage is composed of W3C (World Wide Web Consortium) DOM objects. The standard of the website is measured based on their relevance to the web identity under five measuring scales. These are: Request URL (RURL), URL of Anchor (AURL), Keyword/Description (KD), Main Body (MB) and Server Form Handler (SFH). These categories act as main sources where the identity and features are derived from and it lists the characteristics of phishing like Abnormal DNS record, Abnormal URL, Abnormal Anchors, Abnormal Request URL, and Abnormal Server Form Handler, abnormal certificate in SSL and Abnormal cookie. In this method the related web objects are extracted and converted into a feature vector based on the characteristics of phishing analysis. The phishing detector consists of two components Page classifier and Identity extractor. The feature vector is treated as the input by page classifier that determines if the page is legitimate or not. The ownership is identified by Identity Extractor.

Page classifier uses Support Vector Machine (SVM), a well-known algorithm for classification. The output provided by SVM is either 1 for phishing webpage or -1for authentic webpage.

III. PHISHING ATTACK ANTICIPATION

Following are the characteristics that help us to anticipate the phishing attack

A. *World Wide Web Consortium (W3C) objects:*

A structured webpage is composed of W3C objects, some of these objects are [8]:

1. URL of Anchor (AURL):

A high portion of anchors in a genuine webpage points to the same domain as the page itself. For example <a href="http://www.google.com/"> in http://www.google.com. The name of the webpage must be meaningful and short. It should be in lowercase, use hyphens, contains no space and contains no underscores between words.

2. Server Form Handler (SFH)

The finance/e-business web portals require usernames and password for security reason. Therefore these pages contain a server from handler. They are like <form action="/inetSearch/index.jsp" method="post" target="top"> in http://www.chase.com. The SFH usually is void or refers to a different domain for phishing websites.

3. Request URL (RURL)

In a webpage external objects (such as images, external scripts, CSS) are loaded from some other URLs. A large percent of those URLs are in its own domain for a normal corporate website. For example, <img src="https://home.peoplepc.com/i/60/common/ppco logo.gif"> in http://www .peoplepc.com.

B. *Common properties of Phishing attacks:*

1. Logos

To mimic the appearance, the phishing website uses logos used on the legitimate website. The logos are loaded from the genuine website domain to their phishing websites. Therefore, use of external domain can be treated as a doubtful behavior as for as the phishing attacks are concerned.

2. Suspicious URLs

Servers, Phishing websites are located on have no relation with the legitimate website. The URL of phishing website may contain the legitimate website's URL as a substring (http: //www.googletag.com), or may be similar to the legitimate URL (http://www.goog1e.com) in which the letter L in google is substituted with number 1. Sometime the IP addresses are used to mask the host name (http://25255255255/top.htm)

3. User input:

Phishing websites usually contain pages for the user to enter confidential information, such as account number, password etc.

4. Short lived:

Usually phishing websites are exists for only a few hours or days – enough time for the attacker to deceive a large number of users.

5. Copies:

HTML from the legitimate websites is copied by the attackers and minimal changes are made.

6. Sloppiness or lack of familiarity with English:

Many Phishing pages have grammatical errors, inconsistencies and misspellings.

C. *Page Rank and Reputation of a website:*

It is a heuristic based mechanism to anticipate the legitimacy of a website. In this approach the rank and the reputation of a website is used to decide the authenticity of the website. To obtain the rank and reputation of a website, some popular web search engine can be used as a tool. The URL of a website, a user wants to access is used as the input for the search engine, and the rank and the number of search result returned are used as the decision factor to decide the legitimacy of a website. Legitimate websites get back large number of search results and are ranked high, whereas phishing websites are not ranked at all and get back no result.

IV. PROPOSED METHODOLOGY FOR DETECTION OF PHISHING WEBSITE

In this paper we discuss about mixed approach to conclude the legitimacy of a website, in which combination of list based and heuristics approach will be used along with the exploitation of PageRank and Reputation of the webpage. The URL that the user wants to access is taken as an input to our system. After filtering the URL from its parameters, it is searched in the whitelist. If it is found in the whitelist, it is concluded as the legitimate website. If it is not found in the whitelist it is then searched in the blacklist, and if found it is concluded as the phishing website. If the URL is not found in the blacklist as well as in the whitelist, it is then sent to the web search engine as a search input and the number of search result and the page rank of the website are obtained from the search engine. The number of the search result provided by the search engine for the URL is known as the reputation of the webpage. If the page rank and the reputation of the webpage are high, the webpage can be concluded as a legitimate one. If not it may be a phishing website. Now to ensure if it is a phishing website, the source code file of the webpage is checked for phishing characteristics. These characteristics are extracted out of W3C standard to estimate the security and make a security percentage based on the final weight value to conclude whether the web page is secure or not.

A. *The phishing characteristics according to the W3C standards:*

Phishers use some tricks to deceive and entice the users, so our approach is to check for these tricks and factors in the source code of the webpage. The characteristics that decide the security factor of the website are as follows:

1. ***Https:*** This protocol is used to notify about the security of the website. It should be in the URL of the website but not in the body source of the webpage. Phishers use https inside source code to trick the user and redirect to the webpage of different domain. For example the normal page should be like this , however some phishers use the SSL certificate in the source code like this .
2. ***Images:*** Images of the website should be loaded from the same URL, not from another website, Therefore the entire link should be internal not external. For example, source code like <img src=https: //www.something.com/logo.jpg>" will be considered as a phishing character.
3. ***Suspicious URLs:*** Instead of actual domain name, IP addresses are used by the phishers
4. ***Domain:*** Use of external domain is also considered as a phishing characteristic. For example if user logs in to a website called www.paypal.com and he/she finds some other URL link in the source code like www.pay-pal.com which is not the source URL, it simply means that this website may try to hack our information. Phishers use domain redirection technique to make a webpage available under several URLs.
5. ***Email:*** In PHP, there is a function called mail or email that takes the information whatever is entered through the webpage and sends them to the phishers via email. This trick is used by the phishers by inserting PHP code inside HTML code.
6. ***Iframe:*** Phishers some time use iframe making it invisible i.e. without frame border and when user goes to website, he will not be able to know that some other page has also been loaded in the iframe window, more like a small website opened in current webpage. For example www.google .com can be opened in some other page like in www.mypage.com by using iframe so when somebody enters to this website he/she will see the secured website is opened. This is opened not in the page but is opened through iframe. For example http://www.mypage .com/index.php?search="'><iframe src=http://google.com ></iframe>.
7. ***Popup window:*** Popup windows are also used to steal the user’s information. There are two ways to activate popup windows, one of them is used to confirm something and are written in the HTML code like this < … onClick="window.open ('mypage.html')"> and it is by legal window and html. The other way to activate popup window is illegal because it is a JavaScript file used like: "Open Popup" onClick="javascript: Popup ('mypage.html')">.

B. *Phishing Detection Implementation Concept:*

In our system, the URL that the user wants to access will be used as an input. The first section of this approach uses the list based mechanism to detect phishing website. After filtering the URL string from its parameters it is searched in the whitelist and/or blacklist database table. If the string is matched with the whitelist data, it is concluded as the legitimate website otherwise if it is found in the blacklist data it is concluded as the phishing website. Now if the list based approach fails i.e. the URL string (without parameter) is not found in the whitelist as well as in the blacklist, the heuristic approach is used to decide the authenticity of the website. In this approach, the URL string is sent to the web search engine as a search input and the reputation and the page rank of the website is extracted. If the page rank and the reputation are high the webpage is considered as a secured webpage. If not so, the webpage may be a phishing webpage.

To ensure this the analysis of the webpage source code is needed. In order to achieve this some phishing characteristics out of W3C (Keyword) standard are extracted and the security of the webpage is evaluated. In this method each character in the source code of the webpage is checked and if a phishing character is found the initial secure weight value is decreased. The final security weight percentage is calculated based on the final weight. If the percentage is high the website is concluded as the secure website and otherwise it is concluded as a phishing website. Following are the steps to calculate the security weight percentage:

1. Initialize the counter by a secure website weight.
2. Reads each and every line in the source code individually and perform the following for every line.
   - Perform a check for each phishing characteristics in the webpage source code in the context of different domain, external link for images, suspicious URLs, domain tag, iframe, email, and popup window.
   - Phishing characteristics classification is as follows:

TABLE I: PHISHING CHARACTERISTICS

Phishing characteristics Classification	**Phishing characteristics risk**
Https	Medium
Images	Low
Suspicious URLs	High
Domain	Medium
Email	High
Iframe	Low
Popup window	Low

   - If a phishing character is found by the program the counter is decreased based on the risk of phishing character.
   - Following things are checked in the source code by the program:
     1. First of all the program will search for all images in the website to check any image has a link from some other domain. If so, it is considered as a phishing character. All the images should be in the website folder like for example <img src="Logo.PNG" border="0" width="1349" height="706">.
     2. The login or submit buttons are checked by the program. If the action of the button is linked to any email or IP like 103.838.39.0/login.php or script it is considered as a phishing character.
     3. The program also checks for the iframe, domain, script tags and popup window. If these are found, they are considered as a phishing character.
   - After performing the check for each and every line and considering all phishing characteristics, the final weight is computed to conclude if this website is legitimate and secure or not.
     1. If the final weight of the website takes 85 % of 100 % or above, the website is secured.
     2. If the final weight of the website takes between (84%- 55 %), the website is doubtful.
     3. If the final weight of the website takes under 55%, the website is phishing.

Following is the flowchart for our proposed phishing detection system.

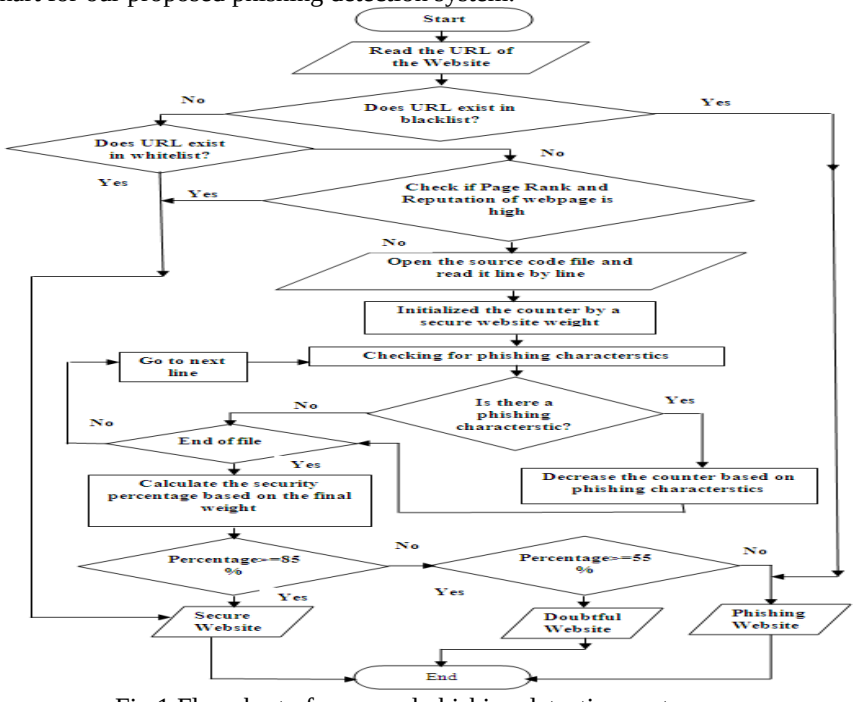

Fig.1 Flowchart of proposed phishing detection system

V. CONCLUSIONS

In this paper we proposed a phishing detection mechanism that checks the webpage security by using list based approach, page rank, reputation and the webpage source code. We extract page rank, reputation by using some popular web search engine and some phishing characteristics out of the W3C standards are extracted by analyzing the source code of the webpage to estimate the security of the websites. As a future work we can include other checks in the program and check more source codes containing many languages in it like PHP, ASP, JSP Perl, etc. Also, the browser plug-in can be developed to check the WebPages and informs the user about the phishing attack.

.

ACKNOWLEDGMENT

Working on this paper has been great learning experience for me. It will be my pleasure to acknowledge, utmost cooperation & valuable suggestion time to time given by staff members of Department. My greatest gratitude is reserved for my project guide Prof. Rajesh Tiwari. He always helped me by introducing the world of research. I am thankful for his benevolence, valuable suggestion, constructive criticism & active interest in successfully making of this paper.

REFERENCES

[1] http://commons.wikimedia.org/wiki/File:Phishing_info_ graph.svg, http:/ /www .gartner. com /it/ page
[2] STAMFORD, Conn., (April 14, 2009). "Gartner Says Number of Phishing Attacks on U.S. Consumers Increased 40 Percent in 2008". Gartner. "UK phishing fraud losses double". Finextra. March 7, 2006. http:// www. finextra. com/ fullstory asp?id=15013.
[3] http://commons.wikimedia.org/wiki/File:Phishing_info_ graph.svg, http:/ /www .gartner. com /it/ page
[4] Richardson, Tim (May 3, 2005). "Brits fall prey to phishing".The Register. http:/ /www/ theregister .co. uk/ 2005/05/03/aol_phishing/.
[5] Steve Sheng, Brad Wardman, Gary Warner, Lorrie Faith Cranor, Jason Hong, and Chengshan Zhang, "An empirical analysis of phishing blacklists," In proceedings of 6th Conference on Email and AntiSpam (CEAS 2009), Mountain View, CA, USA, July 2009
[6] N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, and J. C. Mitchell, "Client-side defense against web-based identify theft," In Proceedings of the 11th Annual Network and Distributed System Security Symposium (NDSS'04), San Diego, 2004.
[7] M. Aburrous, M.A. Hossain, F. Thabatah and K. Dahal, "Intelligent phishing website detection system using fuzzy techniques", in 3rd International Conference on Information and Communication Technologies: From Theory to Applications (ICTTA), pp. 1-6, 2008.
[8] Y. Pan and X. Ding, "Anomaly Based Web Phishing Page Detection", Proceedings of the 22nd Annual Computer Security Applications Conference (ACSAC'06), Computer Society, 2006.
[9] M. Jakobsson and A. Tsow, 2007, "Making takedown difficult", In M. Jakobsson and Steven, editors, Phishing and Countermeasures, pp461–467.
[10] M. Gupta, "Spoofing and countermeasures", 2007, In M. Jakobsson and S. Myers, editors, Phishing and Countermeasures, pp 65–104.
[11] M Tyler, C Richard and S Henry, 2009, "Temporal correlations between spam and phishing websites", In Proceedings (LEET'09) of the 2nd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more, pp5-5.
[12] R Dhamija, D Tygar, and M Hearst, 2006, "Why phishing works" In CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, pages 581–590.
[13] R.Dhamija and J. D. Tygar, 2005, "The battle against phishing: Dynamic security skins". In SOUPS '05: Proceedings of the 2005 symposium on Usable privacy and security, pages 77–888.
[14] M. Rossignol and P.sebillot, 2005 , "Combining statistical data analysis techniques to extract topical keyword classes from corpora", in Intell .Data Anal,9(10), pp- 105-127.

