Real Time Detection of Phishing Websites

ABDULGHANI ALI AHMED, NURUL AMIRAH ABDULLAH

University Malaysia Pahang

abdulghani@ump.edu.my, nurulamirah.abdullah@hotmail.com

Abstract: Web Spoofing lures the user to interact with the fake websites rather than the real ones. The main objective of this attack is to steal the sensitive information from the users. The attacker creates a 'shadow' website that looks similar to the legitimate website. This fraudulent act allows the attacker to observe and modify any information from the user. This paper proposes a detection technique of phishing websites based on checking Uniform Resources Locators (URLs) of web pages. The proposed solution is able to distinguish between the legitimate web page and fake web page by checking the Uniform Resources Locators (URLs) of suspected web pages. URLs are inspected based on particular characteristics to check the phishing web pages. The detected attacks are reported for prevention. The performance of the proposed solution is evaluated using Phistank and Yahoo directory datasets. The obtained results show that the detection mechanism is deployable and capable to detect various types of phishing attacks maintaining a low rate of false alarms.

Keywords*: Phishing Attack; URL; Real Time Model; Phishing Detection*

1. INTRODUCTION

Social engineering attack is a common security threat used to reveal private and confidential information by simply tricking the users without being detected [1]. The main purpose of this attack is to gain sensitive information such as username, password and accounts numbers. According to [2], phishing or web spoofing technique is one example of social engineering attack. Phishing attack may appear in many types of communication forms such as messaging, SMS, VOIP and fraudster emails. Users commonly have many user accounts on various websites including social network, email and also accounts for banking. Therefore, the innocent web users are the most vulnerable targets towards this attack since the fact that most people are unaware of their valuable information, which helps to make this attack successful.

Based on the report prepared by the Anti-Phishing working group organization [2], there were about 163,333 phishing attacks reported in 2014. A recent study by McAfee Lab [3] showed that there were about 30,000,000 new suspected URLs in Quarter 3 for the year 2014. These reports also showed that web browser was classified as the top most network threat which was about 26% compared to the other network threats. For some crime groups, phishing attack is actually a business. Billions of dollars have been reported stolen from banks in US, Russia and Eastern Europe.

Typically phishing attack exploits the social engineering to lure the victim through sending a spoofed link by redirecting the victim to a fake web page. The spoofed link is placed on the popular webpages or sent via email to the victim. The fake webpage is created similar to the legitimate webpage. Thus, rather than directing the victim request to the real web server, it will be directed to the attacker server. Figure 1 shows the steps involved in web spoofing attack.

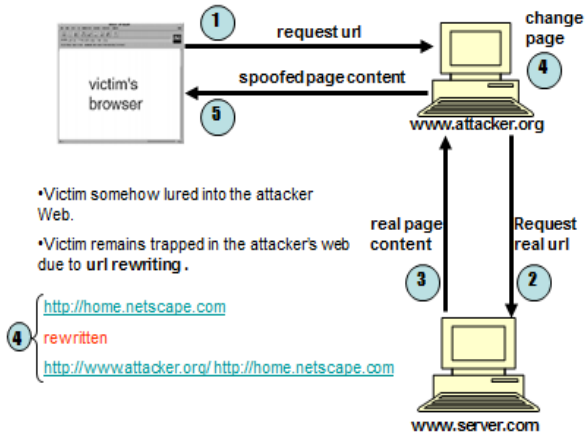

Figure.1: Steps involved in web spoofing attack

There are many researches conducted to detect web spoofing attacks. However, these researches are not effective enough to stop the sophisticated attack of web spoofing. The use of various media communication such as social network leads to the increase of the numbers of attacks. According to [4], 70% of successful phishing attacks are launched through social network. In fact, the lack of awareness and education on web spoofing attack causes the fall of the victims. Inability to distinguish between the fake and legitimate web pages is still a challenge in the existing prevention solutions of web spoofing.

Moreover, the current solutions of antivirus, firewall and designated software do not fully prevent the web spoofing attack. The implementation of Secure Socket

978-1-5090-0996-1/16/$31.00 ©2016 IEEE

Layer (SSL) and digital certificate (CA) also does not protect the web user against such attack. In web spoofing attack, the attacker diverts the request to fake web server. In fact, certain type of SSL and CA can be forged while everything appears to be legitimate. According to [5], secure browsing connection does virtually nothing to protect the users especially from the attackers that have knowledge on how the "secure" connections actually work.

This paper develops an anti-web spoofing solution based on inspecting the URLs of fake web pages. This solution developed series of steps to check characteristics of websites Uniform Resources Locators (URLs). URLs of a phishing webpage typically have some unique characteristics that make it different from the URls of a legitimate web page. Thus, URL is used in this paper to determine the location of the resource in computer networks. Figure 2 shows the architecture of the proposed solution of phishing websites detection.

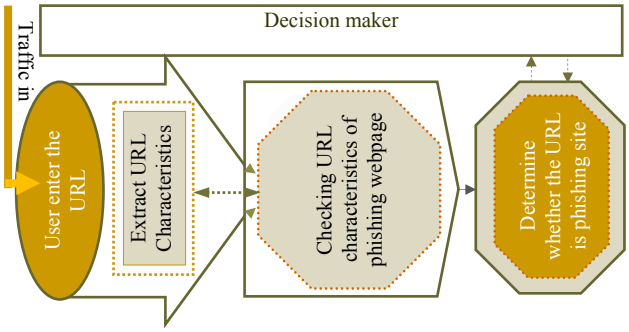

Figure.2: The Proposed Solution Architecture

The rest of the paper is organized as follows: Section 2 discusses the related work; Section 3 describes the proposed solution. Section 4 presents the performed experimental results. The obtained result is analysed and tested in Section 5. Section 6 provides the conclusion and future work; and Section 7 includes the acknowledgments.

2. RELATED WORK

This section reviews the most related works of web spoofing attack. Various researches on web spoofing attack have been done for the past few years. Various researches and methods have been done to study the details of web spoofing attack. Prevention methods of website spoofing are survived and classified into various approaches: content-based, heuristic-based and blacklist-based approaches as shown in figure 3.

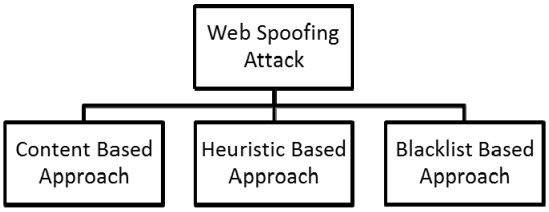

Figure. 3: Web Spoofing Attack Detection Approaches

2. 1 Content Based Approach

This approach identifies web spoofing by inspecting the similarity between the original and spoofed web pages. The similarity between the two web pages is calculated based on the similarities of the web page content. In general, this approach has sufficient accuracy and low false alarms in determining the fake web page. One research belongs to this approach is conducted by CANTINA [6]. This research detects the phishing websites by using Term Frequency/Inverse document Frequency (TF-IDF). Using TF-IDF technique to retrieve information and text mining successfully reduces the false positive rate. The results of CANTINA research show that it catches about 97% phishing site with around 6% false positive. The results also show that with combining some simple heuristics approach to TF-IDF technique, it catches around 90% of phishing sites with only 1% false positives.

CANTINA is successful in identifying the phishing website, but it disables the keyword extraction. Nowadays, some attackers use hidden text in HTML to evade the keyword extraction technique. Moreover, CANTINA suffers from a performance challenge as it needs a considerable time for querying Google.

GoldPhish is another content-based solution [7]. This solution uses Google as a search engine. The philosophy of this solution is that fake website usually active for short period of time. GoldPhish algorithm depends on capturing an image for the current website in the user's web browser. Later, the captured image is converted into computer readable text using an optical character recognition technique. The converted text in this solution is used as an input into a search engine for analyzing the page rank and identifying the possible phishing attack. The finding of GoldPhish shows that it is effective to reduce the false positive and detect zero day phishing. However, this solution is limited with delays in the webpage exploring. Also, GoldPhish solution is vulnerable to attacks on Google's PageRank algorithm and Google's search service.

2. 2 Heuristic Based Approach

Heuristic based approach uses HTML or URL signature identify the spoofed webpages. There are several researches conducted based on this approach. SpoofGurad is one of the solutions that uses heuristics approach [8]. It is an anti-

phishing browser plug-ins. This approach uses a combination of stateless page evaluation, state full page evaluation and examination of outgoing post data to compute spoof index value. If the computed spoof index is greater than a pre-defined threshold value, the page is classified as phishing page and the user is notified about this page. If the spoof index is less than threshold value, the page is classified as legitimate page. SpoofGuard solution has a limitation of generating high rate of false positive in case a sophisticated phishing attack.

Structure of a page [1] and Analyzing the phishing URLs [10] are another studies to distinguish between a legitimate and phishing web pages. Both depend on identifying the features of the characteristics for detecting the phishing web page. By using this solution, phishing attack may be identified and reported as soon as it is launched without the need to maintain a blacklist. However, this solution also generates high false negative rate since there are too many phishing websites classified as legitimate.

2.3 Blacklist Based Approach

This approach has been used for a long time and has been adopted as anti-phishing solutions. This approach has an updated blacklist for the known phishing websites. The phishing blacklist contains all entries that are denied access [11]. Thus user is prevented from accessing web pages that appear in the blacklist. The most important part in blacklist based approach is retrieving the URLs from phishing pages in order to maintain and create the blacklist. The URLs can be retrieved from the users phishing emails, spam, or from the organization that serve the anti-phishing such as Anti-Phishing Working Group (APWG) and Phish Tank [12]. Once a URL is reported, it will be verified first before it is added into the blacklist.

Net Craft Toolbar [14] is one anti-phishing solution that uses blacklist method. It detects the security risk of the web page based on a few criteria such as time of sitting the Net craft web server survey, times of visiting the web page, country that hosted the website, name of organization that hosting the current site and risk rating.

Net Craft Toolbar approach assists to decrease the chances of phishing attack towards users. It also protects the users from downloadable malicious files that may be used by the phishers to collect users' sensitive information. Moreover, Net craft can protect the user from the DNS poisoning and protect the user from the pop-up windows that hides the address bar. Despite the advantages of this approach in protecting the user, the user might encounter new types of phishing attack. The blacklist database requires a continuous updating in order to add the URLs of the new detected phishing websites.

3. PROPOSED MODEL

Web spoofing attacks occur when the user is directed to the fake web page by using fake URLs. This section describes the proposed model of phishing attack detection. The proposed model focuses on identifying the phishing attack based on checking phishing websites features. Further detail about the features of phishing websites is provided in the following subsection.

3.1 Phishing Features Checking

One of the challenges faced in this research is the unavailability of complete dataset to be used as a standard for phishing websites features. According to [14], few selected features can be used to differentiate between legitimate and spoofed web pages. These selected features are many such as URLs, domain identity, security & encryption, source code, page style & contents, web address bar and social human factor. This study focuses only on URLs and domain name features. Features of URLs and domain names are checked using several criteria such as IP Address, long URL address, adding a prefix or suffix, redirecting using the symbol "//", and URLs having the symbol "@". These features are inspected using a set of rules in order to distinguish URLs of phishing webpages from the URLs of legitimate websites. Below is a description for these rules.

a) Feature of IP address is checked to verify if the IP address exists in the URLs. For instance, a URL as "http://192.100.3.124//fake.html" indicates that someone is trying to steal some information from the user. In this study, this URL is checked using the following rule:

$$\text{If } \begin{cases} \text{IP address exist} \rightarrow \text{Phishing Web Page} \\ \text{else} \rightarrow \text{Legitimate Web Page} \end{cases} \quad (1)$$

b) Long URLs usually uses by the phisher to hide the suspicious part. There is no exact length to indicate the phishing site; however, authors in [15] reported that normal length of URL does not exceed 54 characters. Thus, in this study URL with length greater than 54 characters is suspicious link for phishing web pages. This study checks such URLs using the following rule.

$$\text{If } \begin{cases} \text{URLs length is} > 54 \rightarrow \text{Phishing Web Page} \\ \text{else} \rightarrow \text{Legitimate Web Page} \end{cases} \quad (2)$$

c) Phisher tend to add prefixes or suffixes separated by the mark (-) so that the user will trust the URLs as a legitimate web page URL. Below is the rule which can be used to check this feature.

$$\text{If } \begin{cases} \text{URLs include (-) symbol} \rightarrow \text{Phishing Web Page} \\ \text{else} \rightarrow \text{Legitimate Web Page} \end{cases} \quad (3)$$

d) Some URLs of phishing web page have an addition at the front of the real URLs. An example of this addition is http://www.legitimate.com//http://www.phishing.com. This feature checks the location of the symbol "//" in the URL. If the URL starts with "HTTP", this means that symbol "//" should appear in the sixth position. However, if the URL employs "HTTPS" then the symbol "//" should appear in the seventh position. This study checks this feature using the following rule.

$$\text{If } \begin{cases} \text{Position of "//" symbol in the URLs} > 7 \\ \quad \rightarrow \text{Phishing Web Page} \\ \text{else} \rightarrow \text{Legitimate Web Page} \end{cases} \quad (4)$$

e) The use of "@" symbol leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol. Thus, this study classifies any URL includes @ symbol as phishing URL using the following rule.

$$\text{If } \begin{cases} \text{Position of "//" symbol in the URLs} > 7 \\ \quad \rightarrow \text{Phishing Web Page} \\ \text{else} \rightarrow \text{Legitimate Web Page} \end{cases} \quad (5)$$

Figure 4 shows the algorithm of the proposed model which inspects web pages URL features and decides whether the web page is spoofed or normal.

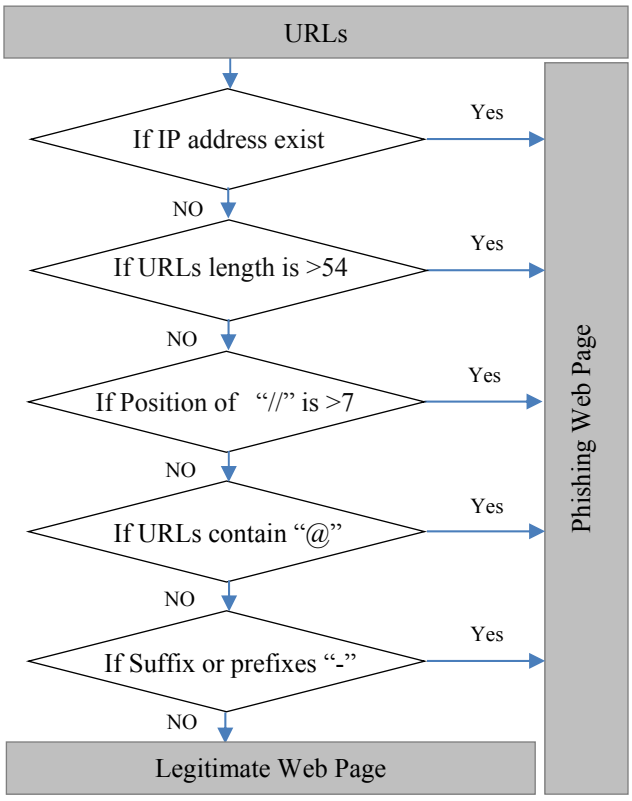

Figure. 4: Phishing Attack Checking Algorithm

4. EXPERMENT RESULTS

In this study, Uniform Resource Locator (URLs) is used as an indicator to distinguish the phishing web page from the legitimate ones. By using the URLs, it can be determined whether the URL comes from a phishing site or legitimate site. In this experiment, Microsoft Visual Studio Express 2013 and C# language were used to create the application that can differentiate the difference between the legitimate and phishing web pages. The designed application is named PhishChecker for short. PhishChecker contains a blank box for entering the URLs which need checking.

Figure 5 shows the main interface of the PhishChecker. When the user enters the URLs into the blank space, the checking rules will inspect the characteristics of the web page URL. If the URLs contain any phishing characteristic, an alert pops up to indicate that the web page is a phishing web page. Figure 6 illustrates the result which the PhishChecker produces when the URL redirect to a phishing web page.

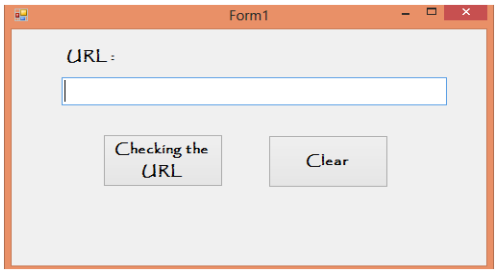

Figure. 5: PhishChecker Interface

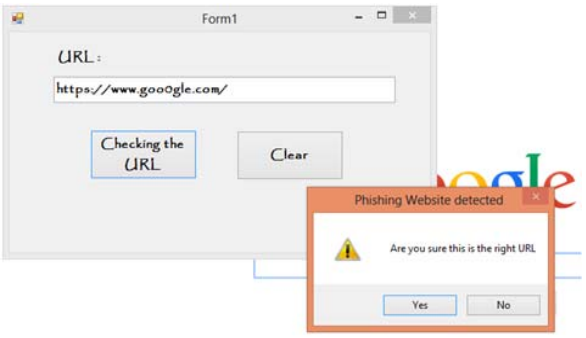

Figure. 6: Phishing Web Page Detected

If the URL does not contain any phishing characteristics, an alert pops up to indicate that the web page is a legitimate web page. Figure 7 illustrates the result which PhishChecker produces when the URL redirect to a legitimate web page.

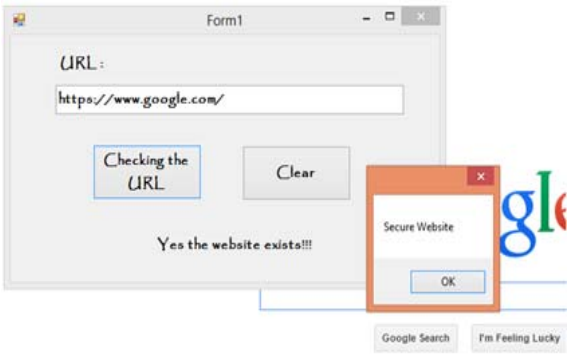

Figure. 7: Legitimate Web Page Verified

5. RESULTS TESTING AND EVALUATION

In this section, the performance of the PhishChecker is tested to verify its efficiency in detecting the phishing web pages. For this purpose, a list of 100 URLs is used (59 legitimate web pages and 41 fake web pages). The used URLs are randomly chosen from the Phistank [12] and Yahoo directory [13] database. For every URL, PhishChecker checks whether the URL has the characteristics of the phishing web page or not. Phistank [16] and Yahoo directory datasets are provided in Appendix A.

The obtained result shows that from the 100 URLs that have been tested, PhishChecker classifies 68 of the URLs as legitimate web page, while the other 32 URLs are classified as fake web pages. Figure 8 shows that 68% from the tested URLs are classified as legitimate web pages and 32% are classified as phishing web pages.

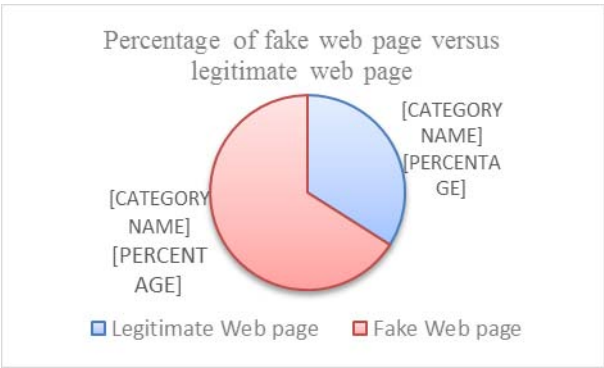

Figure 8: Phishing Detection Accuracy

This result is evaluated through computing the accuracy and false alarm rates. According to [14], the accuracy of attack detection is computed using the following equation.

$$Accuracy = \left(\frac{TP+TN}{TP+TN+FP+FN}\right) \times 100 \quad (6)$$

False negative alarm rate refers to the percentage of phishing web pages URLs falsely filtered as legitimate web pages URLs with respect to the percentage of all phishing URLs. False alarm rate was measured using the following formula, as described in [14]:

$$FalseNegative = \left(\frac{FN}{FN+TP}\right) \times 100\% \quad (7)$$

The obtained results show that PhishChecker detect the phishing web pages with accuracy of 0.96. Moreover, the false negative rates in PhishChecker does not exceed 0.105.

6. CONCLUSION AND FUTURE WORK

Lack of awareness on phishing education makes the attack successful. Even with the help of few indicators used by the browser such as pad lock identification, lock icon, and site identity button, the user still cannot identify the attack. Web spoofing attack is not easy to detect. Even with the newest security prevention method, these attacks still occur. The main aim of this study is to help the users especially to differentiate between the legitimate and phishing web pages by using URL as an indicator. Finding of this research demonstrates its ability to identify the fake webpages based on their URLs.

As a conclusion, the most important way to protect the user from phishing attack is the education awareness. Internet users must be aware of all security tips which are given by experts. Every user should also be trained not to blindly follow the links to websites where they have to enter their sensitive information. It is essential to check the URL before entering the website.

There are a few limitations in this work. The accuracy of this heuristic-based depends on the discriminative features that may help in distinguishing the type of website whether it is a legitimate or phishing site. This study only checks the validity of Universal Resource Locator (URLs) based on a few characteristics for detecting phishing attack.

Future works of this study will include the automatic detection of the web page and the compatibility of the application with the web browser. Additional work also can be done by adding some other characteristics to distinguishing the fake web pages from the legitimate web pages. PhishChecker application also can be upgraded into the web phone application in detecting phishing on the mobile platform.

ACKNOWLEDGMENTS

RDU grant number RDU1403162, Faculty of Computer System & Software Engineering, Universiti Malaysia Pahang supported this work.

REFERENCES

1. Ludl, C., McAllister, S., Kirda, E., & Kruegel, C. (2007). On the effectiveness of techniques to detect phishing sites. In Detection of Intrusions and Malware, and Vulnerability Assessment (pp. 20-39). Springer Berlin Heidelberg.
2. Anti-Phishing Working Group Phishing, (2014). Anti-Phishing Working Group Phishing Trends Report. [Online] Available at: https://apwg.org/ [Accessed 30 Mar. 2015].
3. McAfee Labs Threats Report: February 2015. Retrieved from http://www.mcafee.com/us/resources/reports/rp-quarterly-threat-q4-2014.pdf.
4. Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. Communications of the ACM, 50(10), 94-100.
5. Why HTTPS and SSL are not secure as you think (2014, March 12) .Retrieved from http://scottiestech.info/2014/03/12/why-https-and-ssl-are-not-as-secure-as-you-think.
6. Zhang, Y., Hong, J. I., & Cranor, L. F. (2007, May). Cantina: a content-based approach to detecting phishing web sites. In Proceedings of the 16th international conference on World Wide Web (pp. 639-648). ACM.
7. Dunlop, M., Groat, S., & Shelly, D. (2010, May). Goldphish: Using images for content-based phishing analysis. In Internet Monitoring and Protection (ICIMP), 2010 Fifth International Conference on (pp. 123-128). IEEE.
8. Chou, N., Ledesma, R., Teraguchi, Y., & Mitchell, J. C. (2004, February). Client-Side Defense Against Web-Based Identity Theft. In NDSS.
9. Garera, S., Provos, N., Chew, M., & Rubin, A. D. (2007, November). A framework for detection and measurement of phishing attacks. In Proceedings of the 2007 ACM workshop on Recurring malcode (pp. 1-8). ACM.
10. Sheng, S., Wardman, B., Warner, G., Cranor, L. F., Hong, J., & Zhang, C. (2009). An empirical analysis of phishing blacklists.
11. PhishTank | Join the fight against phishing. (n.d.). Retrieved March 3, 2015, from https://www.phishtank.com/
12. Cranor, L. F., Egelman, S., Hong, J. I., & Zhang, Y. (2007, December). Phinding Phish: An Evaluation of Anti-Phishing Toolbars. In NDSS.
13. Yahoo Business Pages. (n.d.). Retrieved April 12, 2015, from https://business.yahoo.com.
14. B. S. Osareh, "Intrusion Detection in Computer Networks based on Machine Learning Algorithms," IJCSNS International Journal of Computer Science and Network Security, vol. 8, November 2008.