IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-9 E-ISSN: 2347-2693

RSIPS: A Robust System to Identify Phishing Websites using Unique Addressing features of Web

V. Karamchand Gandhi[1*], M. Suriakala[2]

[1*]Dept. of Computer Science, Dr Ambedkar Government Arts College (Autonomous), Chennai, India
[2]Dept. of Computer Science, Dr Ambedkar Government Arts College (Autonomous), Chennai, India

[*]*Corresponding Author: vedhagandhi@gmail.com*

Available online at: www.ijcseonline.org

Received: 07/Aug/2017, Revised: 18/Aug/2017, Accepted: 14/Sep/2017, Published: 30/Sep/2017

***Abstract*—** Phishing is a form of internet fraud in which an attacker, also known as a phisher, attempts to fraudulently retrieve legitimate users' confidential or sensitive credentials by imitating electronic communications from a trustworthy or from the public organization in an automated fashion. There is an need of identify the phishing websites in this emerging digital era. Based on the URL and content based features of websites like length of URL, domain's age, WHOIS properties, etc, we can draw an algorithm to identify the phishing websites. Furthermore, our approach checks the legitimacy of a webpage using hyperlink features. Hyperlinks are extracted from the source code of the given website and apply that into the proposed algorithm to detect phishing site. Our experiment shows that our proposed algorithm is very effective to detect the phishing websites and it have 89.16% True Positive Rate while greater than 82% of accuracy.

***Keywords*—** Phishing URL, Phishing URL/Hyperlink

I. INTRODUCTION

In Phishing, phisher tries to acquire personal data, for example, usernames, passwords, and credit card details generally for malignant purpose. It is a web-based attack that utilizes social engineering methods to exploit users who use internet and gain their delicate information. Most phishing attacks work by creating fake version of the original or legitimate website's web interface to get user's trust and then send forged e-mails with a URL link. This link when clicked, leads to a fake webpage. In most of the cases these emails looks like professional and authorized ones, requesting individual for sensitive data[1].

In the Phishing attack, the attacker creates a fake webpage by copying or making a little change in the legitimate page. So that an internet user can not differentiate between phishing and legitimate web pages. One of the most effective solutions to detect a phishing attack is to integrate security features with the web browser which can generate alerts whenever a phishing site is accessed by an internet user. E-commerce, banks, and money transfer companies are the most targeted industries by these attacks. Seventy-five percent of phishing websites used five top level domains namely .com, .tk, .pw, .cf, and .net.

Generally, all the web browsers provide maximum security against phishing attacks with the help of list-based techniques as solutions. The list-based solutions contain either black-list or white-list. These list-based solutions match the given domain with the domains present in the black-list or white-list to take the appropriate decision. The combination of technical experts and security software verify when a new domain needs to be added in this list. Security software checks the various features of a webpage to verify identity.

Presenting evaluation metrics that are commonly used in the phishing domain to evaluate the performance of phishing detection techniques. This facilitates the comparison between the various phishing detection techniques. Presenting a literature survey of anti-phishing detection techniques, which incorporates software detection techniques as well as user-awareness techniques that enhance the detection process of phishing attacks. Since phishing attacks aim at exploiting weaknesses found in humans (i.e. system end-users), it is difficult to mitigate them. For example, as evaluated in [2], end-users failed to detect 29% of phishing attacks even when trained with the best performing user awareness program. On the other hand, software phishing detection techniques are evaluated against bulk phishing attacks which makes their performance practically unknown with regards to targeted forms of phishing attacks. These limitations in phishing mitigation techniques have practically resulted in security breaches against several organizations including leading information security providers [3], [4].

Google provides a service for safe browsing [5] that allows the applications to verify the URLs using a list of suspicious domains which is regularly updated by Google. It is an experimental API but is used with Google Chrome and Mozilla Firefox, and it is very easy
to use. The Safe Browsing Lookup API [5] allows the clients to send the suspicious URLs to Safe Browsing service which tells whether the URL is legitimate or malicious. The client API sends the URLs with GET or POST requests, which are checked using the malware and phishing lists provided by Google. Some of the shortcomings of Safe Browsing Lookup API are as follows: (i) no hashing is performed before sending URL and (ii) there is no limit on the response time by the lookup server.

Reddy et al. [7] present an anti-phishing technique which protects user at client side against phishing attacks. The proposed technique provides facility for the user to select specific image corresponding to every website he/she visits. Next time, when a user visits the same website and if the images do not match, then the system will alert the user. However, maintaining the image database required a lot of memory, and matching the images of suspicious sites with the stored images required a lot of time.

In a real-time environment, the detection of a phishing attack should be effective and very fast. Black-list-based approaches are very fast, but they cannot detect the zero-hour phishing attack. Visual similarity-based approaches are time consuming, require a lot of memory, and fail to detect the zero-hour attack. Heuristic-based approaches can detect zero-hour attack but their performance depends on the feature set, training data, and classifier. Therefore, in this paper, there is a need to propose an approach based on client side verification to protect against phishing attacks effectively.

II. PROPOSED MODEL

The proposed algorithm focuses on identifying the phishing web pages based on checking phishing websites features. After reviewing all the works on phishing, we propose an algorithm to detect the phishing websites by examines several features like number of Text fields, age of domain, sub levels of domains, length of URL, etc. According to [8], 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, etc. These features are inspected using a set of rules in order to distinguish URLs of phishing web pages from the URLs of legitimate websites. Below is a description for these rules. All these features are examined one by one to achieve the highest accuracy to find the phishing websites.

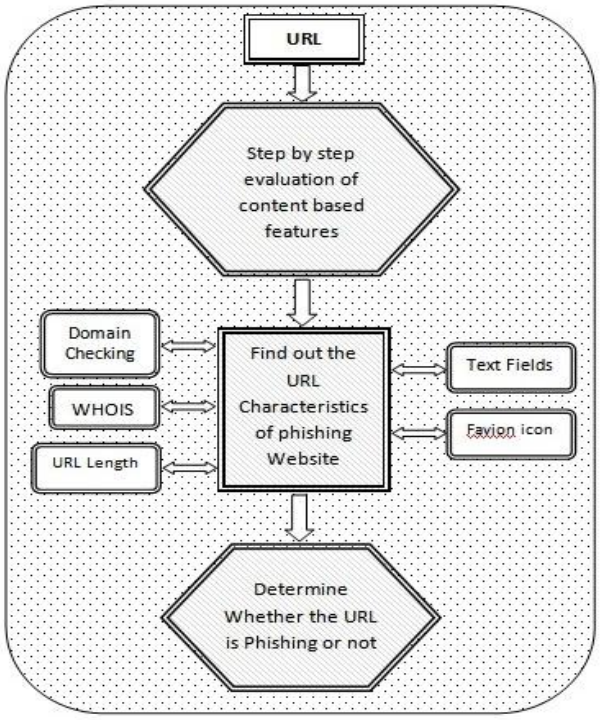

Fig.1 Proposed Phishing Detection System Architecture

III. PROPOSED ALGORITHM

Input: URL/hyper-link
Output:
Display Alert message: Alert ->Possible Phishing
Display Safe message: safe -> No Phishing

```
I.  Begin
2.      For each URL do
3.      TestI:  If (The given URL already exists) then
4.                  Alert
5.              Else
6.      Test 2: If (the given site has more than one text fields) then
7.                  Alert
8.              Else
9.      Test 3: If (The domain age is less than 6 months) then
I0.                 Alert
I1.             Else
I2.     Test 4: if (The URL has more than three dots) then
I3.                 Alert
I4.             Else
I5.     Test 5: If (WHOIS is not found) then
I6.                 Alert
I7.             Else
I8.     Test 6: If (URL length is > 54) then
I9.                 Alert
20.             Else
21.     Test 7: If (Favicon is not loaded from the respective Domain) then
22.                 Alert
23.             Else
24.                 Safe
25. End
```


IV. DESCRIPTION OF THE ALGORITHM

1. Feature of given IP address is evaluated to verify if the IP address already exists in the URL List. For instance, a URL as "http://192.100.3.114//fake.html" indicates that it is malicious site which is trying to steal some information from the user. The rule is

If (The given URL already exists)
Phishing URL
Else
Legitimate URL

2. Most of the phishing sites aimed to collect user's confidential information through text fields by asking their user name, pass word, credit/debit cards number, etc. One way of identify the phishing sites is to examine the text fields by its counts. If the site have more than one text field, the site should suspected to be a phishing.

If (the given site has more than one text fields)
Phishing URL
Else
Legitimate URL

3. The legitimate websites may away from the Blacklist by its long life and reputation. Roughly we can decide that the websites have more age than 6 months, this may be legitimate websites. And moreover, the phishing websites and theirs URL and domains are often created and target the user. So we examine the age of the domain for the website. If the age is less than 6 months, it is suspected as phishing site. The rule is

If (The domain age is less than 6 months)
Phishing URL
Else
Legitimate URL

4. The domain name have maximum one sub domain like www.facebook.com. In some cases there should be two sub domains like www.bdu.ac.in. Here bdu is the actual name of the domain and ac.in is sub level of domain in this url. Like that the entire URL has only three dots including the dot followed by www. We can examine the number of dots in the URL. If there are two dots in the URL, the site is considered 'Legitimate'. If there is three dots in the URL, the site is considered Suspicious. If there is more than three dots in the URL, the site is considered phishing.

If (The URL has more than three dots)
Phishing URL
Else
Legitimate URL

5. When we register a domain name, the Internet Corporation for Assigned Names and Numbers (ICANN) requires our domain name registrar to submit our personal contact information to the WHOIS database. Once listing appears in this online domain WHOIS directory, it is publicly available to anyone who chooses to check domain names using the WHOIS search tool. The protocol stores and delivers database content in a human readable format. The legitimate owner of the domain is identified by WHOIS database. If the DNS record for WHOIS is empty or not found, that domain considered as 'phishing', otherwise, it is considered 'legitimate'. For phishing sites, the WHOIS database is not found.

If (WHOIS is not found)
Phishing URL
Else
Legitimate URL

6. Long URL names are used to hide the doubtful part form the address bar visible. Technically, there is no standard to define the URL length for legitimate website accurately. In our study, the proposed length of legitimate URLs is 75. However, the authors did not justify the reason behind their value. To ensure accuracy of our study, we calculated the length of URLs of the legitimate and phishing websites in our dataset and produced an average URL length. The length of the URL is less than 54 characters, the URL is categories as 'legitimate'. Whereas If the length is more than 74 characters, then that URL categories as 'phishing'.

If (URL length is > 54)
Phishing URL
Else
Legitimate URL

7. Favicon is a graphical image which is used to remember the specific website by its icon. Favicon stands for Favorites Icon. It's the little icon beside your site's name in the favorites list, before the URL in the address bar and bookmarks folder and as a bookmarked website on the desktop in some operating systems. If the Favicon is loaded form the domain which is not from the respective URL shown in the address bar, then the URL is considered as Phishing.

If (Favicon is not loaded from the respective Domain)
Phishing URL
Else
Legitimate URL

The following diagram shows the flow of the algorithm

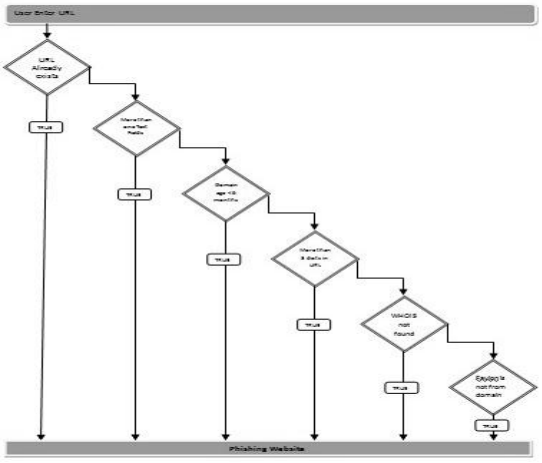

Fig.2

V. EXPERIMENTAL VALIDATION

Our proposed phishing detection algorithm is implemented in Java platform standard edition 7 (JDK 1.7). It takes the URL address of the webpage as an input to checks its legitimacy. The hyperlinks are in the webpage are extracted using Jsoup by parsing the HTML file of the webpage, and a pattern matching scheme is used to obtain the links from the web pages which are not well formed. We have used Guava libraries to find out the parent domains of the hyperlinks. The IP addresses of the parent domains of the suspicious web pages are found using Google Public DNS. Then, the legitimacy of the suspicious webpage is verified by comparing both stored and extracted URL addresses. If Google Public does not find any IP address corresponding to the domain, then we can declare the webpage as phishing. If the suspicious webpage is a phishing one, the system gives warning to the user by alert message. To evaluate the performance of the proposed approach, we have taken the dataset of 525 (480 phishing and 45 legitimate) web pages. Our dataset consists of both phishing and legitimate web pages. The phishing web pages are collected from the PhishTank which is a well known bank of verified phishing URLs. We have collected the phishing URLs during the period of 6 months (June 2016 to November 2016). Legitimate web pages are taken from three different sources. Legitimate datasets consist of the variety of web pages like payment gateway, banking sites, e-commerce, blogs, forum, and social network websites. Various experiments are performed to evaluate the performance of our proposed phishing detection system.

VI. EVOLUTION METRICS

We have calculated the true positive rate, false positive rate, true negative rate, false negative rate, and accuracy of our phishing detection system. These are the standard metrics to judge any phishing detection system. Let NL denote the total number of legitimate websites and NP denote the total number of phishing websites. Performance of RSIPS can be evaluated in the following manner

True Positive (*TP*) rate — measures the rate of correctly detected phishing sites in relation to all the existing phishing sites.

$$TP = \frac{N_{P\to P}}{N_{P\to P} + N_{P\to L}}$$

False Positive (*FP*) rate — measures the rate of legitimate sites which are incorrectly identified as phishing sites in related to all existing legitimate sites.

$$FP = \frac{N_{L\to P}}{N_{L\to L} + N_{L\to P}}$$

True Negative (*TN*) rate—measures the rate of correctly detected legitimate sites in relation to all existing legitimate sites.

$$TN = \frac{N_{L\to L}}{N_{L\to L} + N_{L\to P}}$$

False Negative (*FN*) rate — measures the rate of phishing web sites are incorrectly identified as legitimate in related to all existing phishing websites.

$$FN = \frac{N_{P\to L}}{N_{P\to P} + N_{P\to L}}$$

Accuracy (A) measures the rate of phishing and legitimate websites which are identified correctly with respect to all the websites.

$$\text{Accuracy} = \frac{N_{\text{L}\to\text{L}} + N_{\text{P}\to\text{P}}}{N_{\text{L}} + N_{\text{P}}} \times 100$$

Our system can detect the phishing webpage based on hyperlinks information. The overall true positive rate of the system is 82.28 % and false negative rate is 4.4%.

Total Phishing	Total legitimate	Phishing website classified as phishing	Phishing website classified as legitimate	legitimate website classified as legitimate	legitimate website classified as phishing	True positive rate	False negative rate	Accuracy
480	45	428	97	42	3	89.16%	4.4%	82.28%

VII. CONCLUSIONS

This paper proposes an algorithm to detect the phishing sits by verifying the URL features. Our Algorithm is able to check the legitimacy of a webpage using hyperlink features. Our experimental results showed that the proposed approach is very effective in detecting phishing attacks as it has 89.16 % true positive rate with a less false positive rate of 4.44 %. Moreover, our algorithm is suitable for a real-time environment. In the future, the performance of the proposed phishing detecting algorithm can be improved by taking the other features along with the hyperlinks. However, extracting other features will increase the running time complexity of the system. The accuracy of this proposed algorithm 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 website based on a few characteristics of websites and hyperlinks for detecting phishing attack.

REFERENCES

[1] S. Abu-Nimeh and S. Nair, "*Bypassing security toolbars and phishing filters via dns poisoning"* In IEEE GLOBECOM: Global Telecommunications Conference, 2008.

[2] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, *"Who falls for phish? : a demographic analysis of phishing susceptibility and effectiveness of interventions"* in Proceedings of the 28th international conference on Human factors in computing systems, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 373–382.

[3]B. Krebs, *"HBGary Federal hacked by Anonymous"* http://krebsonsecurity.com/2011/02/hbgary-federal-hacked-by-anonymous/, 2011, accessed December 2011.

[4] V Karamchand Gandhi, "*An Overview Study on Cyber crimes in Internet*" Journal of Information Engineering and Applications from IISTE, Volume 2, Number 1, pages 1-5, February 2012. ISSN 2224-5782 (Paper) ISSN 2225-0506 (Online).

[5] Google safe browsing API Available at: https://developers.google.com/safe-browsing/. Accessed 30 Nov 2015.

[6] W Liu, X Deng, G Huang, AY Fu, "*An antiphishing strategy based on visual similarity assessment"*, IEEE Internet Comput. 10(2), 58–65 (2006)

[7] VP Reddy, V Radha, M Jindal, "*Client side protection from phishing attack*", International Journal of advanced Engineering Science Technology 3(1), 039–045 (2011)

[8] A. Z. Broder, *"Identifying and filtering near-duplicate documents"*, In COM '00: Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching, pages 1–10, London, UK, 2000. Springer-Verlag.

Authors Profile

V.*Karamchand Gandhi* is currently pursuing Ph.D in the Department of Computer Science, Dr Ambedkar Government Arts College (Autonomous), Vyasarpadi, affiliated to University of Madras, Chennai, Tamil Nadu, India. He holds his Master's degree in Computer Science from St.Joseph's College, Tiruchirppalli. He obtained his M.Phil and MBA degrees to the feather of his educational career. He had around ten years of experience in teaching and three years of experience in research. He has published more than 11 research papers in reputed international journals and they are also available online. He written three books in Computer Networks which are available at Amazon.com His research interest includes Information Security and Computer Networks. He is a life member of various technical societies such as IAENG, IACSIT, etc.

Dr M Suriakala is currently working as an Assistant Professor in PG and Research Department of Computer Science, Dr Ambedkar Government Arts College (Autonomous), Vyasarpadi, Chennai affiliated to University of Madras, Tamil Nadu, India. She has completed her M.Phil., Computer Science from Bharathidasan University in 2004 and Ph.D., Computer Science from University of Madras in 2009. She has around 18 years of rich experience in teaching and 10 years of vast experience in research. She has published more than 56 research articles in the reputed International / National Journals and Conferences. She has chaired many technical sessions and delivered invited talks in National and International Conferences. Two books in the area of E-Commerce Security were authored by her. Her research interest includes Data Mining, Web Mining and Data Science. She is a member of various technical bodies like IAENG, IACSIT, etc.

