WC-PAD: Web Crawling based Phishing Attack Detection

Nathezhtha.T[1], Sangeetha.D[2],Vaidehi.V[3]
[1,2] Madras Institue of Technology, Anna University, Chennai, Tamil Nadu, India
[3]Mother Teresa Women's University, Kodaikanal, Tamil Nadu, India
[1] Nathezhtha31@gmail.com, [2]dsangeethabaskaran@gmail.com, [3]vaidehi@mitindia.edu.

Abstract- Phishing is a criminal offense which involves theft of user's sensitive data. The phishing websites target individuals, organizations, the cloud storage hosting sites and government websites. Currently, hardware based approaches for anti-phishing is widely used but due to the cost and operational factors software based approaches are preferred. The existing phishing detection approaches fails to provide solution to problem like zero-day phishing website attacks. To overcome these issues and precisely detect phishing occurrence a three phase attack detection named as Web Crawler based Phishing Attack Detector(WC-PAD) has been proposed. It takes the web traffics, web content and Uniform Resource Locator(URL) as input features, based on these features classification of phishing and non phishing websites are done. The experimental analysis of the proposed WC-PAD is done with datasets collected from real phishing cases. From the experimental results, it is found that the proposed WC-PAD gives 98.9% accuracy in both phishing and zero-day phishing attack detection.

keywords: Phishing, web crawler, Heuristics, Zero-day phishing, Attackers.

I. INTRODUCTION

Phishing attack is used to steal user's sensitive information such as username, password, credit card details, etc. The reports on phishing attack is increasing with the growth of economy. In phishing, emails or websites are often used for stealing information[3]. The phishing websites resemble exactly like the original websites to trap the users. Anti-phishing work group[4] reports that the phishing activates has been increasing highly. Phishers target their victim and attack by emails, messages or phone calls[3]. Figure 1 shows the number of phishing sites detected in the year of 2018. Phishing can happen in various ways. Deceptive phishing[10] is a mechanism where an attacker impersonates as organization and steal information from people. Identification of such attacks can be done by inspecting the URL and differentiating the scammer from genuine links. Spear phishing [12] is a targeted attack through emails where the attackers target an entity and collects information about the entity in social media sites like linkedin. The attacker crafts the information of the target and attack through emails. Pharming [10], a cache poisoning attack on Domain Name System(DNS).

Figure 1: Phishing sites reported on 2018

The attacker changes the IP address assigned to the DNS and redirects users to a malicious website. Dropbox Phishing[10] is a type of phishing, where the attackers wants to access the files of the dropbox users, so they create a fake dropbox sign in page, which can be hosted on dropbox itself and steal the users credentials. Google docs phishing[10] is similar to dropbox phishing, where the attackers aims to access the Google drive and its documents, such attack happened in 2015, the google page not only hosted the fake login page but also gave SSL certificate to protect the page with secure connection. Even with the anti-phishing measures, the phishing attacks still exists due to the following reasons. First, the users prefer mobile phones over desktop to use the browser and check mails[8]. They are more likely to access phishing sites which has not yet been detected by the anti-phishing sites. Secondly, the users do not use or remove the anti-phishing tools on their mobile phones because of energy consumption and memory space. Third, the available antiphising tools does not perform well in terms of detection. It has been reported that the mobile users are three times more likely to submit their information than desktop users on phishing websites[9]. Figure 2 shows the industries most targeted by the phishing attackers. The primary target of the attackers was payment but attacks on Software as a service(SaaS) has increased highly on 2018. The attackers goal is to steal the sensitive data by initiating the attacks on SaaS or webmail. The proposed WC-PAD precisely identify the attacks created for website phishing.

978-1-7281-1576-4/19/$31.00 ©2019 IEEE

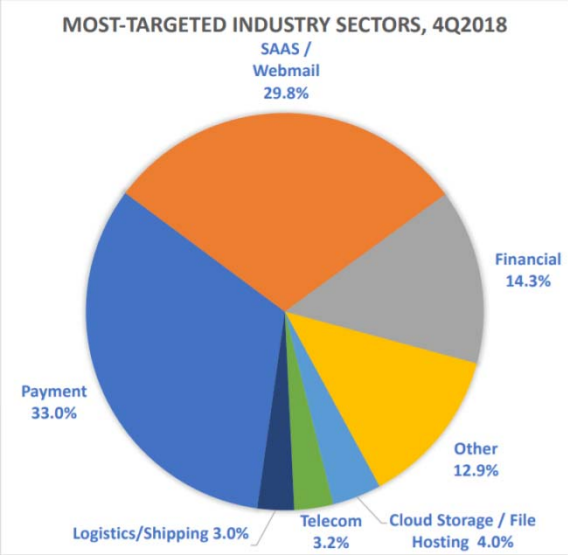

Figure 2: Most targeted industry for phishing attacks

The paper is organized as follows Section II contains the detail explanation of the related works. Section III explains the proposed WC-PAD. The experimental results have been discussed in Section IV and the paper has been concluded in Section V.

II. RELATED WORKS

Phishing[2] violates the rule of Confidentiality, Integrity and Availability. Many approaches are evolving to detect phishing attacks, yet it is viable and acts as a threat to people. The phishing websites can be detected with high accuracy with hardware devices but it is very expensive, so opting to software based approaches are preferred. Blacklist and whitelist[14]are used for phishing detection with high accuracy but these are in need of list maintenance, since it is opt to manually updating the list of URL of phishing websites. Hence, to overcome the issues in manual list update, automatic detection approaches like machine learning and heuristic approaches are currently used[16]. Many machine learning techniques use web structure or web content-based methods for identifying the phishing URL, Cantina+[5] is a well-known heuristic based approach. Sunil et al. proposed a phishing detection approach solely based on googles PageRank, by only using PageRank value it is difficult to identify phishing attack, since many new legal websites or blogs with low rank and can be misinterpreted as phishing website, but combination of multiple features can improve the performance and identification rate. Web structure approach[15] based on page ranking for phishing identification was proposed. Plug-in for browsers can be used to mitigate phishing attacks [11] but the user alert is generated only based on the blacklists again it does not resolve zero-day phishing attack. Netcraft phishing detection[14] approach is based on both blacklist and heuristics. It detects the phishing website quickly but it is not very efficient when the attackers design the attack to avoid detection. Dhamija et al proposed Security skin a dynamic approach to generate visual hash for browser window customization to indicate it's a secured site[7]. Rule based filters are the combination of keywords, IP address, syntax checkers, URLs, etc. features[6]. These features are used for generating rules for identifying the phished emails. The rules are consistently updated for better identification. Web crawler also known as web spider or internet bot is an automated scripts used to browse webpages in an automated manner. Introducing web crawlers to the research of phishing is very new, since crawlers are generally used for information extraction. It also resolves the zero-day phishing attack website which are not addressed by many other existing approaches. Most of the search engines use web crawlers[5], they are used to gather pages from web for indexing in search engines. It gathers all the useful web pages and interconnected links. Zero-day phishing websites are fresh attacks launched by the attackers less than a day or hour. It is very dangerous, since there will not be any case report with such types of attacks. The existing approaches are effective in identifying the phished website but are inefficient when it comes to zero-day phishing attack. To overcome this issue, a web crawler based phishing detection approach has been proposed.

III. PROPOSED WC-PAD

Web Crawler based Phishing Attack Detector (WC-PAD) is a three phase phishing attack detection approach. The three phases of WC-PAD include 1) DNS blacklist 2) Heuristic based approach and 3) Web crawler based approach. Here the web crawlers are used for both feature extraction and phishing attack detection. Figure 3 shows the overall architecture of the proposed WC-PAD.

A. DNS Blacklist

The Domain Name System based Black list also known as DNS Blacklist is an approach of publishing a list of IP address, which are meant to avoid and can be easily programmed on internet. It is built on top on Internet DNS. The DNS Blacklist provide the IP address which involves spam activity. The DNS Blacklist are updated frequently. The WC-PAD extracts the web address and compares it with the address in the DNS blacklist as first phase, if a match of IP address is found an immediate alert is send to the webpage user else the WC-PAD processed with its second phase.

B. Web Crawler

WC-PAD starts crawling to the websites interconnected pages and links. WC-PAD crawls from one website to other going through all the links until all the web indexed are crawled. The proposed WC-PAD uses web crawlers to crawl

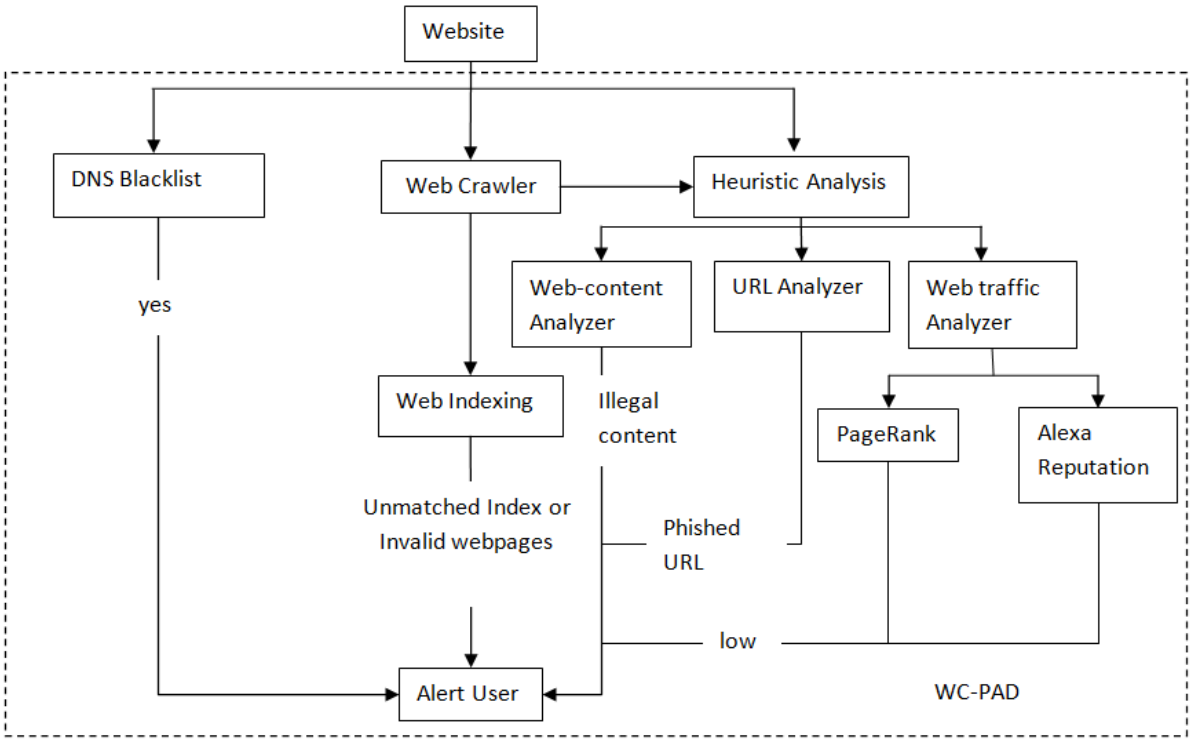

Figure 3: Architecture of Proposed WC-PAD

each web pages of a website, since attackers do not index all the web links in the phished website. Crawlers are also used to extract some features from the website. Figure 4 shows the working mechanism of web crawler.

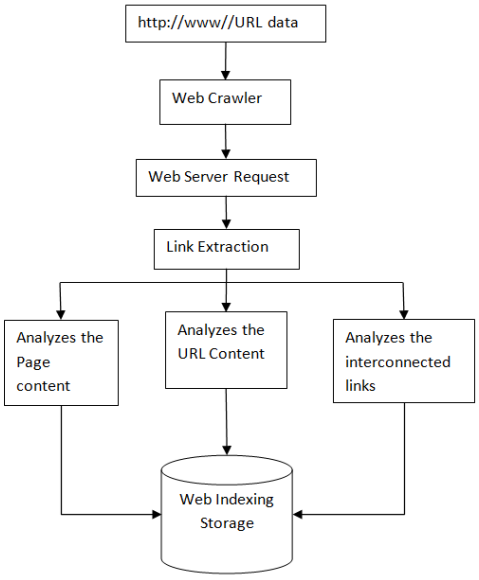

Figure 4: Web Crawler for Web Indexing

The proposed WC-PAD is used for identifying the fault in web indexing. If any of the unmatched web index is found or any of the web pages or links does not work, the WC-PAD alerts the user. Experiment analysis prove web crawlers are very effective when it comes to zero-day phishing attack detection.

C. Heuristic Analysis

Heuristic analysis takes three features Web content feature, URL feature, Web traffic feature all these features are extracted by the web crawler. Algorithm 1 explains the overview of Heuristic Analysis. The three features have separate analysis phases as follows

URL Analysis

The WC-PAD are designed in a way to extract the information form URLs all the interconnected URLs. The URL is partitioned as follows

<protocol>://<SubDomain>.<PrimaryDomain>.<TLD>/ <PathDomain>.

For example, consider the following URL:

http://paypal.abc.net/index.htm There are six elements in index.htm: http is the protocol, paypal is the SubDomain, abc is the PrimaryDomain, net is the top-level domain (TLD), abc.net is the Domain, and index.htm is the PathDomain.[1].

Algorithm 1: Heuristic Analysis

```
Input : URL features, Web content features, Web traffic features
Output: Phished website or non phished website
Copyrights→ WC-PAD (WebContent)
        if Copyrights is illegal then
                "Alert User"
        end if
WC-PAD(URLAnalyzer)←Call Algorithm 1
WC-PAD(WebtrafficAnalyzer)
        Count(total visits, pages per visit, average visit duration, Bounce rate)
        state→compute(PageRank)
        if state is low then
                "Alert User"
        end if
        state→compute(AlexaReputation)
        if state is low then
                "Alert User"
        end if
```

PrimaryDomain cannot be empty as SubDomain/ PathDomain, or a phishing Primary Domain often contains an IP address. Algorithm 2 explains the mechanism of URL Analyzer.

Algorithm 2: URL Analyzer

```
Input : URL features- d primary domain, @,- dots,ld
Output: Classification of URL as legal or phished
 if d is IP then
        State= Phishing
else if  occurrence('@','-','.')
        if '@' && '-'
                State= Phishing
        else if occurrence ('.')>5
                State= Phishing
        end if
else if ld<3
        State= Phishing
end if
if state is Phished then
        "Alert User"
end  if
```

URL Analyzer checks for the occurrence of '@' and '-' in URL, since '@' in a URL mean its left side can be discarded and only 59 characters are taken from the right side. and Legal sites does not use '-' very often. The URL of legitimate site does not contain many dots, only the phishing websites contains many dots. So the URL is also scanned for number of dots. The URL analyzer also checks the dictionary words and reports on the finding of misspelled words with levenshtein distance(ld). Levenshtein distance is used to calculate the difference in two strings sequence. If the distance is less, there is a possibility of phishing attack. URL Analyzer will also check whether the given URL contains the IP and whether that IP is the IP of its domain. Based on these Heuristics, the website will be classified as legal or phished website. WC-PAD not only extracts the URL of a website but also extracts all the interconnected URLs and crawls to the interconnected URL to find whether those URLs are valid. The URL Analyzer perform a validation on all the interconnected links.

Web content Analysis

The proposed WC-PAD has been programmed in a way to crawl through the web page contents and copyrights in the website. Based on the crawled pages, the web content analyzer classifies a web page contents to be legal or illegal. If the contents are classified as illegal, then the WC-PAD sends an alert message to the user regarding the suspicion.

Web traffic Analyzer

The web traffic analyzer takes parameters such as total visits for websites, pages per visit, average visit duration and the bounce rate. Based on these parameters, the website is classified as zero-day phishing websites or normal website. The Web traffic analyzer also take the Google PageRank and AlexaRepution to identify the bounce rate of a website. PageRank[15,16] uses a link analysis algorithm of Google search engine build to calculate PageRank values. The phishing website usually contain a low value, since these kinds of website exist only for a short time. AlexaReputation[17] value of a website is calculated based on the count of links from other webpages to itself. It is similar to PageRank. The AlexaReputation value will be low for phishng websites and higher AlexaReputation is similar to Pagerank, where AlexaReputation values of phishing websites are much lower than the values of the legitimate sites.

IV. EXPERIMENTAL RESULTS

The phishing URLs are collected from the website PhishTank and the legal URLs are scraped from the legal websites. The datasets are randomly divided in the ratio of 70:30 as legal and phished websites. The proposed WC-PAD is programmed in Java with Selenium web driver for crawling through the pages and weblinks. Initially, the WC-PAD checks the availability of the IP in DNS Blacklist, if it doesn't occur, it goes to Web Crawler phase, where both web

indexing and feature extraction is done. If WC-PAD finds any fault in web indexing the user is alerted, else WC-PAD proceeds with the third phase namely Heuristic Analysis, where the web contents are analyzed for copyrights, URL features and web traffic features are analyzed and alert is generated accordingly. Table 1 shows the allocation of weights for URL Features. The highest weight is given to the interconnected links.

TABLE 1: Heuristics weight allocation

Heuristics	weights
IP Address	0.210
Domain age	0.145
Occurrence of('@','-')	0.045
Occurrences of('.')	0.125
Interconnected links	0.475

Based on the above weights, the websites state is computed and classified as a phished or non-phished website. Table 2 shows the Accuracy, Sensitivity and specificity calculation of the proposed WC-PAD which is as follows.

- Accuracy of a website is calculated based on correctly detected websites

$$Accuracy = \frac{\#\ True\ phishing + \#\ True\ legitimate}{\#\ Total\ sites} \quad (1)$$

- Sensitivity is calculated based on correctly detected legitimate website

$$Sensitivity = \frac{\#\ True\ legitimate}{\#\ True\ legitimate + \#\ False\ alarm\ websites} \quad (2)$$

- Specificity is the rate of correctly detected phishing websites

$$Specificity = \frac{\#\ True\ phishing}{\#\ True\ phishng + \#\ Missed\ detection} \quad (3)$$

TABLE 2: Performance of WC-PAD

Evaluation in (%) percentage	Phishing Website(%)	Zero day Phishing website(%)
Accuracy	98.8	99
Sensitivity	99.2	99.5
Specificity	99.1	99.2

Figure 5 shows the classification of phishing website and legitimate website with the random picked website address of 70: 30 ratio of non-phishing and phishing website.

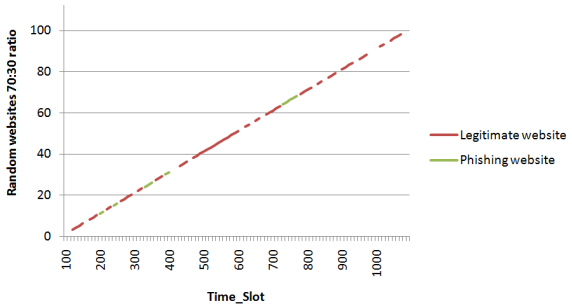

Figure 5: Phishing and legitimate website identification with 70:30 ratio

The proposed WC-PAD detects the phishing websites accurately and classifies the legitimate websites and phishing website in a précised manner.

V. CONCLUSION

WC-PAD for detecting phishing attacks has been proposed. WC-PAD performs a three phase identification mechanism, Firstly, the DNS Blacklist based detection is done. Secondly a Web Crawler based detection is accomplished followed by Heuristics based detection. The frequently used phishing IP are easily detected in DNS blacklist testing. The zero-day phishing website attacks are identified in Web crawler and heuristic analysis phase. The experimental analysis has been done for the proposed WC-PAD and it precisely detects the phishing websites. WC-PAD produces a detection accuracy of 98.9% including both phishing attacks and zero-day phishing attacks.

Reference

[1] C. Pham, L. A. T. Nguyen, N. H. Tran, E. Huh and C. S. Hong, "Phishing-Aware: A Neuro-Fuzzy Approach for Anti-Phishing on Fog Networks," in *IEEE Transactions on Network and Service Management*, vol. 15, no. 3, pp. 1076-1089, Sept. 2018.

[2] S. Marchal, J. François, R. State and T. Engel, "PhishStorm: Detecting Phishing With Streaming Analytics," in *IEEE Transactions on Network and Service Management*, vol. 11, no. 4, pp. 458-471, Dec. 2014.

[3] L. Wenyin, G. Huang, L. Xiaoyue, X. Deng, and Z. Min, "Phishing Web page detection," in Proc. IEEE 8th Int. Conf. Document Anal. Recognit., Seoul, South Korea, 2005, pp. 560–564.

[4] Anti-Phishing Working Group. Accessed: Sep. 2016. [Online]. Available http://www.antiphishing.org

[5] A. Naga Venkata Sunil and A. Sardana, "A PageRank based detection technique for phishing web sites," *2012 IEEE Symposium on Computers & Informatics (ISCI)*, Penang, 2012, pp. 58-63. doi: 10.1109/ISCI.2012.6222667

[6] C. N. Gutierrez *et al.*, "Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks," in *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 6, pp. 988-1001, 1 Nov.-Dec. 2018. doi: 10.1109/TDSC.2018.2864993

[7] R. Dhamija and J.D. Tygar, "The Battle against Phishing: Dynamic SecuritySkins", Proc. Symp. Usable Privacy and Security, 2005,pp 77-88. Mobile Marketing Statistics. Accessed: Mar. 2017.

[8] [Online]. Available:

http://www.smartinsights.com/mobile-marketing/mobilemarketing-analytics/mobile-marketing-statistics/
[9] Phishing Attacks. Accessed: Sep. 2015. [Online]. Available: https://securityintelligence.com/
[10] [Online]. Available: https://www.tripwire.com/state-of-security/security-awareness/6-common-phishing-attacks-and-how-to-protect-against-them/
[11] Blake Ross, Collin Jackson, Nick Miyake, Dan Boneh and John C. Mitchell. A Browser Plug- In Solution to the Unique Password Problem. http://crypto.stanford.edu/PwdHash/, 2005.
[12] J. Wang, T. Herath, R. Chen, A. Vishwanath and H. R. Rao, "Research Article Phishing Susceptibility: An Investigation Into the Processing of a Targeted Spear Phishing Email," in *IEEE Transactions on Professional Communication*, vol. 55, no. 4, pp. 345-362, Dec. 2012.
[13] Chou, N., R. Ledesma, Y. Teraguchi, D. Boneh, and J.C.Mitchell. "Client- Side Defense against Web-Based Identity Theft", in Proceedings of The 11th Annual Network and Distributed System Security Symposium.
[14] Netcraft, Netcraft Anti-Phishing Toolbar. Visited: Nov 20, 2006. http://tool bar.netcraft.com/
[15] A. N. V. Sunil and A. Sardana, "A PageRank based detection technique for phishing Web sites," in Proc. IEEE Symp. Comput. Informat. (ISCI), Penang, Malaysia, 2012, pp. 58–63.
[16] Checking Page Rank. Accessed: Sep. 2016. [Online]. Available: https://www.prchecker.info/check_page_rank.php
[17] [Online]. Available: http://tutology.net/category/how-php/get-alexa-rank-php-and-alexa-api.