Phishing Page Detection Tool: Shield from Cyber Frauds

Varun Garg, Gaurav Saraswat

***Abstract*—Phishing happens to be one of the greatest frauds which are occurring inside the area with high economic growth such as Indian .Though high-level companies are aware of the risk and preventive measures, but the small and medium companies are not able to adopt security practices Even the security tools are expensive and the losses due to frauds over the Internet is increasing .Thus, we have come with phishing detection tool which could minimize these risk of phishing by the use of some simple yet effective filters we deployed but aggressive research.**

I. Keywords

Phishing, Digital frauds, Breaches, loopholes, Google search API , Page Rank API, URL, frauds detection, Forensics.

II. Introduction

India is one of the fastest growing countries in its It Sector. The future growth of information technology sector in the Indian subcontinent is expected to grow at even more rapid speed than ever. With the growing of its sector in India, cyber frauds are also increasing. Cyber frauds have many forms of implementation but phishing is a most common term in those forms. Phishing is one of the fastest growing online crime which mainly used to steal critical user information or money. Digital frauds and online thefts are becoming a global threat growing with penetration of digital technology in our daily lives .We use highly confidential personal information online like ATM credit card numbers, Bank Account Numbers, Passwords, Social Security numbers which can lead to trouble if this information is known to other individuals. It can be carried out with phishing emails, websites etc. This kind of Indian Corporations lost around USD fifty-three million(about Rs.328 crores) as a result of phishing scams.

As the cyber frauds are affecting Indian economy at a very large scale, Indian organization should give cyber security a vital importance. Using the commonly used security standards in working culture can decline the rate of these criminal activities. There are many solutions for detecting phishing websites which are using the analysis of websites URL to detect the fraud. There are various solutions are introduced in organizations to prevent and avoid these cyber frauds. Using span filter in email accounts, installing a regularly updated spyware protection system, encrypting communications , filtering the websites to be used inside the internal network and education the employees about security and cyber scams. These are some of the methods which can be used to prevent security breaches and security fraud. In our research paper, we have created a program which can detect the phishing page by various proposed filtering mechanisms. The program can also help out in guessing the original web page belonging to a particular URL.

III. Cause of Phishing

Lack of knowledge among the masses : Globally and significantly in India, there is a lack of basic knowledge concerning the cyber attacks among the public[1]. The people are not aware that their information is regularly being accessed by criminals and that they don't take correct precautions while performing on-line activities[12]. Unawareness of policy – The cyber thieves typically take the advantage of Bank/financial establishment policies and operations for contacting customers, mainly for contacting the customers for the problems like account maintenance and fraud investigation. The average user is unaware of the operations of the organization becomes a vulnerable to the phishing scam, simply because they are unable to understand the technical complexity. Technical sophistication – Hackers and cyber criminals checks for current advanced technology that can be used for operations similar to spam, distributed denial of service (DDOS), email attacks and surveillance. While the common man is being educated to the common online fraud methods hackers develop complex techniques to perform these crimes to counter this awareness. Using these techniques they can perform fake email address obfuscation and create phishing emails and websites that similar to the organization email and web page and is not noticeable to the victim. Going into a further deep level of complex technique hacks explores vulnerabilities in web browsers which enable data access and allows the attacker to send a malicious payload to the victim's system.

IV. Techniques used in phishing

1) Man-in-the-middle attacks: In this attack, the hacker or the attacker listens to the network between the client and application. Thus, the attacker proxies all communications between the systems. This manner of attack is same for each HTTP and HTTPS communications. The client connects [4] to the attacker's server thinking it as the real website and then send the interaction of the user to the real application. The advantage of this particular attack is that hacker can easily access the user login credentials and other post information entered by the client during the operation. The following attack can also be accomplished by installing fake SSL certificate in the targets system.

A. URL Obfuscation Attacks:

Attacking the universal resource locator takes in minor changes to the universal resource locator, the hacker accomplishes the task by sending a fake and lucrative request to the user to follow a link (URL) to the attacker's server, while users do not realize that he has been tricked [5]. Universal resource locator's Obfuscation uses vulnerable methods of the TCP/IP protocol to trick users into viewing an Internet site that they didn't will visit.

B. XSS (Cross-site Scripting)

Cross-site scripting attacks (XSS) is a vulnerability which exists due to unvalidated user input in the web application. XSS are categorized into two types stores XSS and reflected XSS. Phishing attacks can be performed by finding XSS in URL of the application and extracting the cookie information. The crafted XSS universal resource locator is then sent to the target victim and if the victim falls into a trap the browser cookie for the web service gets compromised and the attacker can access the victim's account using the stolen cookie information.

V. common Detection methods used

A. Blacklisting

A universal resource locator is compared against a "blacklist" [11] of URL's that are reported as phishing sites.

B. Heuristics

identifies patterns[9] within the universal resource locator love suspicious sites. Performs higher than blacklisting.

C. Machine Learning

Extremely effective at distinguishing phishing sites, however, needs massive white-lists to forestall false positives[14].

D. Trusty communication

The methodology that involves a method to spot legitimate sites and build a white-list. However, this methodology will solely determine a tiny low fraction of the whole range of legitimate sites.

E. Hybrid

This methodology combines many of the options listed to spot phishing sites[13].

F. Hashing calculation algorithm

[4]In order for the computation of image of the websites like facebook.com, gmail.com, worldbank.com, or any banking website and comparing them with the given universal resource locator [10] however once more than it's an honest approach of computing the tactic.

VI. Techniques used in proposed phishing detector

A. Google Search API Verifications

The technological implementation used in order to develop such a method for verification of genuine url for this we extract the source code and extract various parameters such as <title>, <body> or meta tags. Further on we use google search API [2] in order extract the search results and further on we analyzing the URLs extracted from the API results to the domain name of the url we need to check. This method also allows the user to browse and get the results for the original url of the page, this happens to be valid for every company every university which is indexed on google and not like other tools which are strict towards their reach and does not tell you what the original URL is.

B. URL Checking :

Getting the source code of the input url, Searching the input url in the page source itself, If the url is matched, then it could be a real page, Else it could be a phishing page. This method may not work in case of website which uses different technologies for hyperlinking different web pages such technologies are jQuery and JavaScript libraries Ember.js

C. Page Rank Checking:

The input url's domain is searched on the web, [1] Its google page rank is computed, If the page is frequently used, it would have a page rank in the range of 0-9, Else it would have a negative value.

VII. WORKING PROCESS OF DEVELOPED TOOL

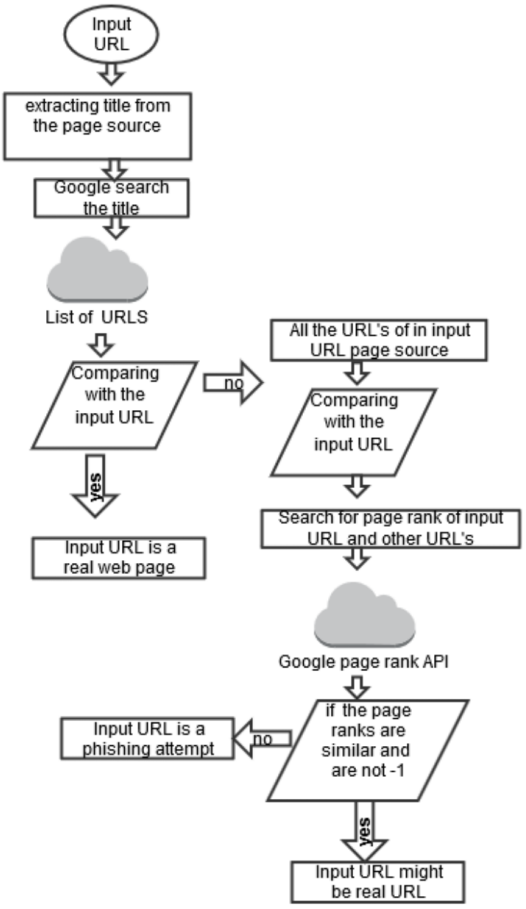

Figure 1:Flowchart to explain the working

A. working process

- Test the link page source get the Title of the page
- Explore title using google search API results
- Comparison of URL with google search API results
- Extracting all URL's from the source code of Input URL
- Link and page source URL's comparison
- Finding page rank of the given URL domain
- Evaluation of results of all tests
- Asserting the page is Real or Phishing webpage

B. False negatives

For those websites which are recently launched does not produce genuine page rank in google page rank API such that for these websites the tool returns may or may not be a phishing page] The websites whose <title> <body> tags does not specify information about them do produce false results and produces the output as a phishing page.

VIII. IMPLIMENTATION

Following images are the results obtained from the implementation of developed algorithm .The test input URL was taken from an online phishing archive Phishtank.com containing active phishing website reported .One of URL was taken and phishing detection mechanism successfully detected the phishing page.

A. IDENTIFYING THE PHISHING PAGE

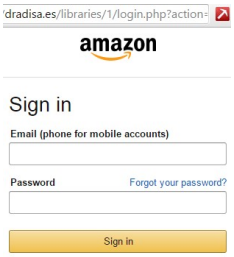

Figure 2:Testing Amazon login phishing page

PHISHING DETECTION TOOL

The Input URL :
http://dradisa.es/libraries/1/login.php?
action=billing_login=true&_session;5688c6ae40488d4

TEST 1:

The title tag data extracted from the page source : Amazon.com Sign In
The Google Search API query
http://www.google.com/search?
q=Amazon.com Sign In&start=0

The URL's obtain from Google Search API are :
https://www.amazon.com/gp/sign-in.html
https://www.amazon.com/your-account
www.amazon.in
https://www.amazon.in/your-account
https://sellercentral.amazon.in/

Matching the Input URL with the google search Api results

Test 1 Result: Match Not Found

Figure 3:Screenshot of Test 1 on the phishing URL

In the first test as shown in figure 4 from the input URL's page source the HTML tags like title tags the data is extracted .The data obtained is web search over the internet by Google web search API .If the match is not found the further test are carried on.

TEST TWO

The URL's extracted from page source of input URL
fls-na.amazon.com
//fls-na.amazon.com/1/batch/1/OP/ATVPDKIKX0DER:192-0547942-9872968:4H4DTJH31VF6SKXYN6RJuedata=s:
images-na.ssl-images-amazon.com/images/G/01/AUIClients/AmazonUI
https://www.amazon.com/ap/signin
https://www.amazon.com/ap/register?openid.pape.max_auth_age=0

MATCHING THE INPUT URL FROM THE ABOVE EXTRACTED URL'S :
Match Result =FALSE

Figure 4:Screenshot of Test 2 on the phishing URL
As the figure 5 screenshot shows that URL's from the input URL's page source are being extracted and being compared .After looking at various phishing webpages it has been observed that the page source URL's domain often is same as the domain of the webpage itself .Thus this test can be used to identify a phishing page.

TEST THREE

FINDING THE PAGE RANK OF INPUT URL DOMAIN

Page Rank Of Google Search URL DOMAIN: 8 Page Rank Of Input URL's Domain =-1

RESULT : This is a Phishing Page

Figure 5:Screenshot of Test 3 on the phishing URL
As shown in the figure above the input URL's page ranks is taken from Google page rank API .It has been observed that a phishing page often results in a Google page rank of -1 .Further the final result is declared on the basis of all test results

IX. Conclusion

The following program can be utilized not only for the diagnostics but also for forensics as it is able to provide the original web page associated with the URL since from the Google search API results top URL's are being extracted.The top search results are often the URL's of the real domain name . The program can perform diagnosis for a phishing page not only in common use web services like facebook, Gmail but also for the fraud authentication pages for small and medium companies web pages.

With this, we are also able to harvest the original page in case it is The millions of URL Obfuscation Attacks can be mitigated with this tool and further development using these methodologies as a precaution measure can save a person from the fraud.

References

[1] Joshi, V., Rajamani, N., Takayuki, K., Prathapaneni, N., Subramaniam, L. V., (2013). Information Fusion Based Learning for Frugal Traffic State Sensing
[2] wikipedia.org/wiki/Sensor$_{f}usion$
[3] Google Json API, *https://developers.google.com/custom-search/json-api/v1/overview*

Varun Garg graduated in Electronics and Electrical Department from Maharaja Agrasen Institute of Technology in 2016.

Gaurav Saraswat graduated in Computer Science Engineering Department from Maharaja Agrasen Institute of Technology in 2016.