ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 1, Issue 2, October 2014

Detection of Phished Websites

G. Vasanth Kumar[1], U. Seshadri[2]

M.Tech, Department of Computer Science,Vaagdevi Institute of Tech & Science, Proddatur, Kadapa, India[1]

Head of the Dept, Dept of computer science,Vaagdevi Institute of Tech & Science,Proddatur, Kadapa, India[2]

Abstract: Phishing, a criminal act of gathering personal, bank and credit card information by sending out forged e-mails with fake websites, has become the most popular recitation among the criminals of the Web. Phishing attacks are becoming more and more sophisticated and are constantly on the rise. Phishing is a major threat to information security and personal privacy. The total number of URLs used to host phishing attacks are increased to 1, 75,229 in the second quarter of 2013, up from 1, 64,023 in the first quarter of 2013 according to APWG, a Phishing Activity Trends Report. Many anti-phishing solutions, such as content analysis and HTML code analysis, rely on this property to detect fake web pages. However, these techniques failed, as phishers are now composing phishing pages with non-analyzable elements, such as images and flash objects. This paper proposes a new phishing detection scheme based on URL domain similarity, IP matching and image matching. This paper correctly estimates the phished website in three phases. At first it estimates similarity with authorized URL database, here itself we eliminate, in second phase we confirm based on IP matching and in final phase we find number of key-points matched.

Key words: Phishing; anti-phishin;,URL similarit;,ip matching; image matching.

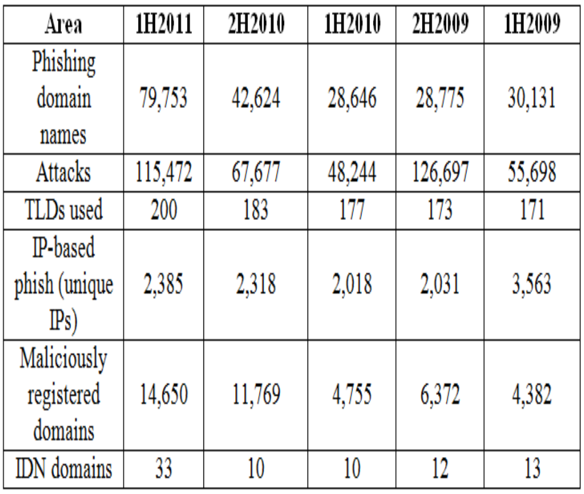
I. INTRODUCTION

Phishing refers to the process where a targeted individual is contacted by email or telephone by someone posing as a legitimate institution to lure the individual into providing sensitive information such as banking information, credit card details, and passwords. The personal information is then used to access the individual's account and can result in identity theft and financial loss. Although such scams originated sometime around the year 1995, they did not become commonly known by everyday people until nearly ten years later. That doesn't mean that phishing was not a force to be reckoned with right from the start. The fact of the matter is that phishing scams have been causing serious problems ever since day one.

A. Anti phishing Activity Trends Summary(2009-2011)

The Anti-Phishing Working Group (APWG) [1] is the global pan-industrial and law enforcement association focused on eliminating the fraud and identity theft that result from phishing, pharming and email spoofing of all types.

TABLE1. BASIC STATISTICS



Area	1H2011	2H2010	1H2010	2H2009	1H2009
Phishing domain names	79,753	42,624	28,646	28,775	30,131
Attacks	115,472	67,677	48,244	126,697	55,698
TLDs used	200	183	177	173	171
IP-based phish (unique IPs)	2,385	2,318	2,018	2,031	3,563
Maliciously registered domains	14,650	11,769	4,755	6,372	4,382
IDN domains	33	10	10	12	13

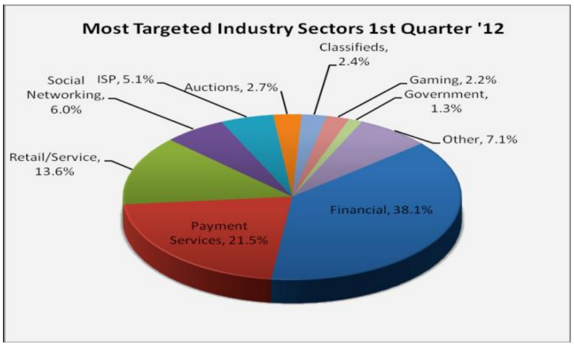


Fig.1. Most targeted industries in 1Q2012

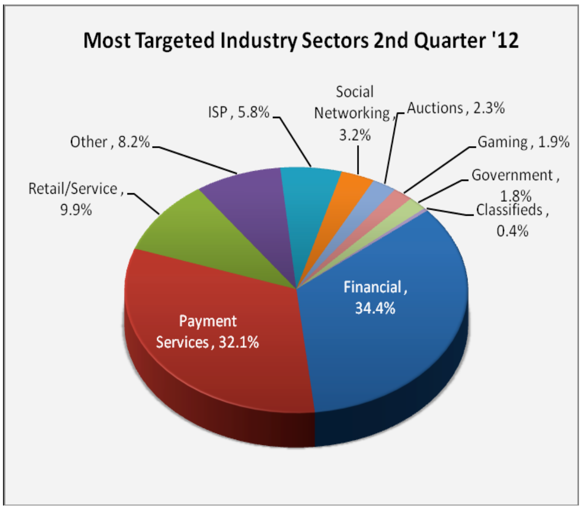


Fig.2: Most targeted industries in 2Q2012

II. PHISHING TECHNIQUES

Phishing is the method used to steal personal information through spamming or other deceptive means. There are a number of different phishing techniques used to obtain personal information from users. As technology becomes more advanced, the phishing techniques being used are also more advanced. Some of the phishing techniques are:

Email / Spam

Phishers may send the same email to millions of users, requesting them to fill in personal details. These details will be used by the phishers for their illegal activities. Phishing with email and spam is a very common phishing scam. Most of the messages have an urgent note which requires the user to enter credentials to update account information, change details, and verify accounts. Sometimes, they may be asked to fill out a form to access a new service through a link which is provided in the email.

Web Based Delivery

Web based delivery is one of the most sophisticated phishing techniques. Also known as “man-in-the-middle,” the hacker is located in between the original website and the phishing system. The phisher traces details during a transaction between the legitimate website and the user. As the user continues to pass information, it is gathered by the phishers, without the user knowing about it.

Instant Messaging

Instant messaging is the method in which the user receives a message with a link directing them to a fake phishing website which has the same look and feel as the legitimate website. If the user doesn’t look at the URL, it may be hard to tell the difference between the fake and legitimate websites. Then, the user is asked to provide personal information on the page.

Trojan Hosts

Trojan hosts are invisible hackers trying to log into your user account to collect credentials through the local machine. The acquired information is then transmitted to phishers.

Link Manipulation

Link manipulation is the technique in which the phisher sends a link to a website. When the user clicks on the deceptive link, it opens up the phisher’s website instead of the website mentioned in the link. One of the anti-phishing techniques used to prevent link manipulation is to move the mouse over the link to view the actual address.

Session Hacking

In session hacking, the phisher exploits the web session control mechanism to steal information from the user. In a simple session hacking procedure known as session sniffing, the phisher can use a sniffer to intercept relevant information so that he or she can access the Web server illegally.

System Reconfiguration

Phishers may send a message whereby the user is asked to reconFig. the settings of the computer. The message may come from a web address which resembles a reliable source.

Phishing through Search Engines

Some phishing scams involve search engines where the user is directed to products sites which may offer low cost products or services. When the user tries to buy the product by entering the credit card details, it’s collected by the phishing site. There are many fake bank websites offering credit cards or loans to users at a low rate but they are actually phishing sites.

Phone Phishing

In phone phishing, the phisher makes phone calls to the user and asks the user to dial a number. The purpose is to get personal information of the bank account through the phone. Phone phishing is mostly done with a fake caller ID.

Malware Phishing

Phishing scams involving malware require it to be run on the user’s computer. The malware is usually attached to the email sent to the user by the phishers. Once you click on the link, the malware will start functioning. Sometimes, the malware may also be attached to downloadable files.

III. LITERATURE REVIEW

Anti-phishing research is fascinating subject to many who are interested in nature and mathematics. Everyone who uses online will likely encounter some form of phishing victims. Most new internet browsers come with anti-phishing software’s that protect against phishing including legislation and technology created specifically to protect from phishing.

EARLY MODELS OF ANTI-PHISHING

Email-Level Approach [6]

Most current antiphishing strategies focus on the emails that are sent as phishing bait. Email authentication and spam filtration can help reduce phishing attacks by filtering out messages, but the risk of losing important emails is also high. Web browsers can use blacklisting to filter against known sites, but there is always latency between site reporting and blacklist updating. Indeed, as phishing-site lifetimes are reduced to hours from days, this method might prove totally ineffective.

Phishing prevention measures should be complemented with detection methods. The key strategies include

1. Monitor domain name registrations.
2. Watermark the original web pages to identify usage in phishing sites.
3. Monitor web server logs for suspicious referral entries and excessive traffic from one source IP.
4. Track double-bounce mails.
5. Setup forum for users to report phishing.

Email-Level Approach includes authentication and content filtering. The email filtering techniques, is commonly usedto prevent phishing. These are quite popular in antispam solutions because they try to stop email scams from reaching target users by analyzing email contents. Phishing messages are usually sent as spoofed emails; therefore, researchers have proposed numerous path-based verification methods.

Current mechanisms, such as Microsoft's Sender ID or Yahoo's Domain Key, are designed by looking up mail sources in DNS tables.

Browser Integrated Tool Approach

A browser-integrated tool usually relies on a blacklist containing the URLs of malicious sites to determine whether a URL corresponds to a phishing page. Attackers might not want to write their own HTML page -they might just Copy-Paste the content of the original website and make their own page. If we insert something like obfuscated javascript in the original website [which alerts us when run under any URL other than the authentic] we can get alerted against these attacks. There are many methods for watermarking your original website to track a phisher. On the original website if we are analyzing the web server logs and looking for suspicious referrers we will be able to detect an phishing attack in progress.

A popular approach to fight phishing is to maintain a list of known phishing sites and to check website against the list. Also safari3.2 and opera contain this type of anti-phishing measure. In Microsoft Internet Explorer (IE) 7, for example, the address bar turns red when a malicious page loads. A blacklist's effectiveness is strongly influenced by its coverage, credibility, and update frequency. Currently, the most well-known blacklists are those Google and Microsoft maintain for the popular browsers Mozilla Firefox and IE, respectively. However, experiments show that neither database can achieve a correct detection rate greater than 90 percent, and the worst-case scenario can be less than 60 percent.

Webpage Content Analysis

It analyzes a Web page's content, such as the HTML code, text, input fields, forms, links, and images. In the past, such content based approaches proved effective in detecting phishing pages. Phishers responded by compiling pages with non-HTML components, such as images, Flash objects, and Java applets.

Many times attackers design the phishing webpage such that the images are picked up from the original site rather than keeping a repository of images in their fake website. When the user loads the phishing webpage, the browser goes and picks the images from original website. The referrer URL as seens by the original website will be the URL of the fake website.

A phisher might design a fake page composed entirely of images, even if the original page contains only text information. In this case, content-based antiphishing tools can't analyze the suspect page because its HTML code contains nothing but HTML <img/> elements.

Visual similarity based analysis

New solution is proposed by Anthony Fu and his colleagues, detecting phishing pages based on the similarity between the phishing and authentic pages at the visual appearance level, rather than using text-based analysis.

Perhaps the most technical creativity in the phishing community today resides in the art of misdirecting users via the format of e-mailed URLs. Several tactics have been observed over 2004 and 2005, and as anti-spam researchers and developers attack one obfuscation technique, new methods emerge.

Crafted "Automatically Generated" Links

When displaying messages, most graphical mail clients take the liberty of converting plain-text HTTP links into clickable URLs, without the need for HTML processing.

An important feature of a phishing webpage is its visual similarity to its target (true) webpage. Hence, a legitimate webpage owner or its agent can detect suspicious URLs and compare the corresponding WebPages with the true one in visual aspects.

If the visual similarity of a webpage to the true webpage is high, the owner will be alerted and can then take whatever actions to immediately prevent potential phishing attacks and hence protect its brand and reputation. This module extracts the Web pages' features and measures the similarity to the true pages according to three metrics: block-level, layout, and style. If the visual similarity is higher than the corresponding threshold, the system issues a phishing report to the customer. However, this approach is susceptible to significant changes in the Web page's aspect ratio and important colors used.

This approach focuses on early detection of possible phishing Web pages without inconveniencing the end users. Within the same framework, also developed a full-image approach that converts two Web pages into pure images and then calculates their similarity using the Earth Mover's Distance(EMD), which represents the least energy required to transform one image into another. EMD offers a method for evaluating the distance (dissimilarity) between two signatures, or sets of features and their corresponding weights.

The method comes from the well-known problem of how to transport goods from one place to another with the least effort (consumption). Researchers have successfully used EMD to assess image similarity, but a phisher can defeat this approach by mimicking a small but vital block of the target Web page. Detecting phishing Web pages is similar to the problem of detecting duplicate documents and plagiarism, except that these focus on text-based features in similarity measurement, whereas phishing-page detection should focus more on visual similarities. Pure text features are insufficient for detecting phishing pages. An important feature of phishing pages is that they use similar or even exact visual effects to mimic the true pages. Hence, full-page similarity assessment is necessary in most cases.

IV. PROPOSED WORK

This system proposes a new scheme for phishing page detection in three phases. They

- URL and Domain Identity
- IP comparison
- Image Based Webpage Matching

URL and Domain Identity

Normally phishing is done via sending mails to thousands of users urging them to visit the fake website through the link or URL present in it. The input for proposed project is URLs for the detection process. These URLs are mostly similar to authorized URLs, with very minor variation which couldn't observed by normal users. Using similarity of ranking algorithm, similar authorized URLs will be searched which are stored in database (file) that is often targeted by phishers.

IP Comparison

In this phase we will calculate the IP addresses of the similar URLs. If IP addresses of the Authorized URLs do not match with the IP address of entered (input) URL then this URL could be phishing one. This URL will be considered as input for next phase which are based on the webpage's image matching.

Image Based Webpage Matching

In this phase, take a snapshot of a suspect webpage whose URL is detected as a suspected phishing URL in previous phases and treat it as an image throughout the detection process.

V. SYSTEM DESIGN

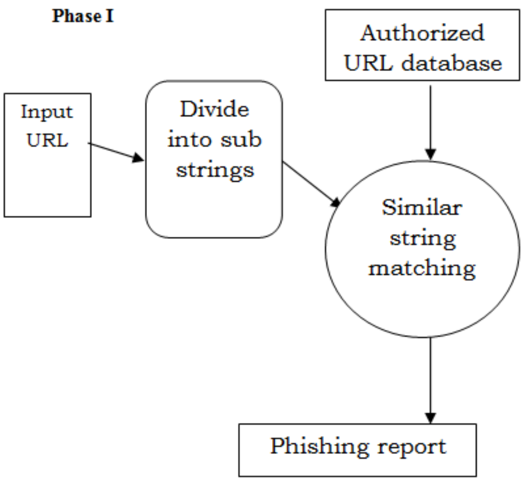


Fig.1. Data flow diagram for URL and domain identity phase

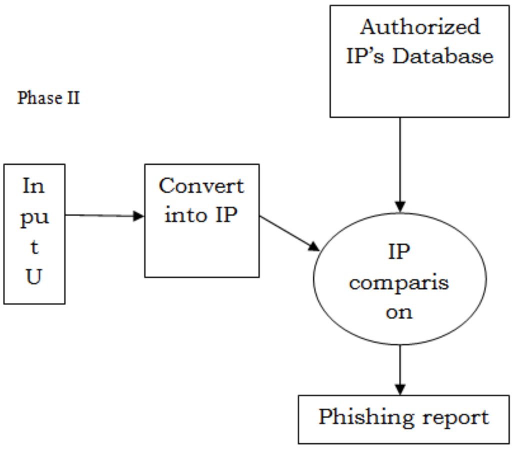


Fig.2. Data flow diagram for ip comparison phase

Phase I &I

Above data flow diagrams clearly explains that how the system is designed. Both phases are looking similar but with slight variation. Using similarity of ranking algorithm, similar authorized URLs will be searched which are stored in database (file) that is often targeted by phishers. If the similarity is greater than or equal to 60 then it is not phished otherwise it send to next phase. Here ip address taken for this URL and compared with authorized ip database. If it is not found here, then it is send to next phase.

Phase III
Image Based Webpage Matching

In this phase, take a snapshot of a suspect webpage whose URL is detected as a suspected phishing URL in previous phases and treat it as an image throughout the detection process. Here we are using a small tool, to detect and matching contrast context histogram (CCH). Object recognition can be considered as matching salient corners with similar CCH descriptors on two or more images. It shows that CCH is insensitive to image scales, rotations, viewing directions, and illumination variations. The text window shows the numbers of CCH descriptors in these two images and the number of matched descriptors.

SIMILARITY RANKING ALGORITHM

The steps of this algorithm are as follows.

Input: Input URLs are given by user.
Authorized URLs domain name stored in database.
Steps: Find out pairs of each string. Pair is formed of adjacent characters of string. E.g. Let authorized URL domain is paypal, then pairs={pa, ay, yp, pa, al}.

Then similarity between two pairs calculated by following formula Similarity (s1, s2) = | pairs (s1) Ω pairs (s2)|*100/Pairs (s2)
Where s1= Input URL String,
s2=Authorized URL,
Pairs (s1) =Pairs for each substring of URL,
Pairs (s2) = Pairs for Authorized URL
Ω=Intersection of pairs for authorized URL & input URL

Output: Similarity Value

If the similarity value is equal or greater than 60 then the input URL substring is related to authorized URLs used for pairs which are stored in database. It becomes related authorized URL. If similarity value is less than 60% then there may be possibility that no single word of input URL string related to any authorized URL in database. In this case we have to extract html source content. From these html content source we will consider only <href> content i.e. the link to other WebPages. Then treat this reference URL as input URL string and repeat above steps as like an input URL. Let take an example, pairs for each substring are as follows.
http= {ht, tt, tp}
www= {ww, ww}
paypel= {pa, ay, yp, pe, el}
com= {co, om}

Repeat the above step until all words pairs are find out. For authorized URLs, let's take two financial organizations' URLs.

Pairs for them are as follows.
paypal= {pa, ay, yp, pa, al}
ebay= {ab, ba, ay}
For each authorized URLs and input URL substring calculate similarity value.
Similarity value for paypel and paypal is
Pairs(s1)={pa,ay,yp,pe,el} Pairs(s2)={pa,ay,yp,pa,al}
Pairs (s1) ΏPairs (s2) = {pa, ay, yp}
|Pairs (s1) ΏPairs (s2)|=3
|Pairs (s2)| = 5
Similarity value= (3/5)*100=60
So, this input URL is related to paypal.

VI. CONCLUSION

Phishing has become a major threat to information security and personal privacy. Most current approaches focus on text-based analysis; however, increasingly, phishers are constructing phishing pages that look very similar to the legitimate ones, but they have totally different code compositions embedded in order to avoid detection. This paper involves new antiphishing technique based on URL domain identity and image matching mechanism. Here we first identify the related authorized URL. We used approximate string matching algorithm. In next phase image matching mechanism uses key-points detection and feature extraction methods through a software tool, i.e. CCH research tool [8]. The results of experiments also show that with high accuracy and computation time is very less.

ACKNOWLEDGEMENT

G. Vasanth kumar Author thanks to **M. NARESH BABU** Assistant Professor, Department of Computer Science & Engineering, VAAGDEVI INSTITUITE OF TECH & SCIENCE, proddatur for his valuable guidance. The real mentor and motivator of this project **U. SESHADRI** Assistant Professor, Department of Computer Science & Engineering, VAAGDEVI INSTITUITE OF TECH & SCIENCE, proddatur for giving us the opportunity to work with him and for all his efforts, patience and his encouragement, is gratefully acknowledged.

REFERENCES

[1] http://www.antiphishing.org/reports/apwg_report_dec_2007.pdf.
[2] http://www.phishing.org/
[3] http://www.phishtank.com/
[4] http://en.wikipedia.org/wiki/Phishing
[5] www.ijcst.com/vol23/3/radha.pdf
[6] www.ra.ethz.ch/cdstore/www2005/docs/p1060.pdf
[7] www.ijcst.com/vol22/2/madhuri.pdf
[8] www.infosecwriters.com/text_resources/pdf/Spam_SSiddharth.pdf

BIOGRAPHY

G. Vasanth Kumar, PG scholar. I had attended a national conference on e-commerce. This is my first publication in journal.