International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 4, Number 17 (2014), pp. 1761-1770
© International Research Publications House
http://www. irphouse.com

A Hybrid Approach to Detect Zero Day Phishing Websites

Namrata Singh[1], Nihar Ranjan Roy [2]

[1,2]Galgotias University, Greater Noida, U.P.

Abstract

Phishing is a significant problem that tricks unsuspecting users into revealing private information involving fraudulent email and websites. This causes tremendous economic loss every year. In this paper, we proposed a novel hybrid phish detection method based on phishing blacklists and phishing properties. We used some fresh phish from PhishTank that were recently added to test that it can be detected by blacklist or not. We found that 70 % of the phishing websites in our dataset lasted less than two hours. Blacklists were ineffective when protecting users initially, as most of them caught less than 20% of phish at zero hour.

Another check used in this approach is phishing characteristics. Phishing characteristics are properties occur in phishing websites. It caught significantly more phish at zero hour than those using only blacklists. Finally we tested this approach on a set of legitimate URLs for false positives, and did not find any mislabeling.

Keywords: Anti-Phishing, Blacklist, Evaluation, Phishing, Phishing Properties, Zero-day Attack.

Introduction

Cybercriminals require virtual real estate on the Internet to execute various aspects of their phishing attacks. And the use of legitimate websites enables them to more easily evade common web security measures. There has been a dramatic increase in phishing, a kind of attack in which victims are tricked by spoofed emails and fraudulent websites into giving up sensitive information such as bank account number, password, social security number, and so on [2-4]. Phishing is a rapidly growing problem, with 42,890 unique phishing sites detected in December 2013 alone and USA is the most phishing sites hosting country [1].

74.67% phishing websites contain some form of target name in URL. 1.60% phishing websites just used IP address, no hostname. 0.78% phishing websites not using port number 80, according to the report. These are the results of December alone. Phishing patterns evolve constantly, to achieve a high true positive (TP) rate while maintaining a low false positive (FP) rate it is hard for a detection method. In this paper, we propose a novel hybrid approach based on blacklist and phishing properties with URL detection and page content check. This is an attempt to achieve a good balance between TP and FP. The blacklist based detection component of this algorithm proposed in this paper, examines the URL and domain names in the blacklist database. As the blacklist contains two strings one of corrupted URL and another of corrupted domain. This approach makes use of the best way to detect zero day phishing as it uses phishing properties to examine those websites that are not in the blacklist. According to the above mentioned report, most of the phishing websites contain some form of manipulated URL from legitimate sites this approach uses URL detection method to handle with URL manipulation. We present a summary evaluation, comparing our proposed approach to two popular anti-phishing toolbars that the most effective tools for detecting phishing sites. Our experiments show that this approach has comparable or better performance to SpoofGuard with fewer false positives, and does well with Netcraft too. SpoofGuard is a heuristic based anti phishing toolbar; Netcraft is a blacklist and heuristic based anti phishing toolbar. Finally, we show that proposed approach is effective at detecting phishing websites.

This paper is organized as follows. Section 2 introduces relevant anti phishing detection methods. In section 3, we introduce the proposed detection methodology, followed by experiment and results in section 4, and finally, conclusion in section 5.

Related Work

Efforts to detect phish can be implemented at the phishing website level. To prevent phishing websites from reaching potential victims, traditional detection techniques such as Bayesian filter, blacklist, and rule based rankings can be applied. Generally speaking, research to detect phish at the website level falls into two categories: heuristic approaches, which use HTML or content signature to identify phish, and blacklist based methods.

The method web browsers use to identify phish is to check URLs against a blacklist of known phish. This method leverage human verified phishing URLs in an effort to control the FP rate. Blacklists of known spammers have been one of the predominant spam filtering techniques. It may contain IP addresses or domains used by known spammers, IP address of open proxies, country and ISP netblocks that send spam, RFC violators, and virus. Though, there is a problem to block a phishing website because some phishing websites are hosted on hacked domains. It is not possible to block the whole domain because of a single phish on that domain. So a blacklist of Phishing URLs is a better solution in this scenario [6].

The heuristics based approaches utilize HTML or content signature to discriminate cases and controls. For this, machine learning algorithms are usually applied. A variety of heuristics have been proposed for phish detection. TF-IDF is a well-known algorithm often used for information retrieval and text mining [11]. The algorithm is used for comparing and classifying documents. The importance of a word in a document is measured by this algorithm and a proper weight is assigned to it according to its importance level. The importance of the word increases proportionally to the number of times a word appears in the document.

Password hashing is another method to detect phishing in which outgoing passwords are checked. This method provides unique passwords to a domain [7].

Researchers at Carnegie Mellon have created an anti-phishing solution, named as Cantina. The base of this solution is TF-IDF algorithm. TF-IDF yields a weight to measures how important a word is to a document in a corpus and that importance is proportional to the number of times a word appears in the document. But it is offset by the frequency of the word in the corpus. It first calculates the TF-IDF scores of each term and then creates a lexical signature of that webpage by taking the five terms with highest TF-IDF weights.

There is another approach Bayesian Toolbar. This is an anti phishing toolbar that uses a Bayesian filter. Bayesian filter is used as spam filter for email filtering. This filter has a property to detect never before seen items. Bayesian toolbar uses white list that is a collection of US financial institutions and ecommerce sites and other phishing prone sites.

SpoofGuard is a web browser plug-in the monitors a user's Internet activity and computes a spoof index. The user is warned if the index exceeds the level fixed by the user. SpoofGuard uses domain name, URL, link, and image checks.

Methodology

This approach makes use of blacklist as well as phishing properties for detecting zero day phishing sites. Blacklist is a well known method to detect phishing sites it contains a list of Phishing URLs and domain names and if a user click on a website the URL of that website is checked against the blacklist database. In this section first we describe the phishing properties and then how this algorithm works.

Phishing Properties

There are some properties that are very common in phishing attacks.

- Suspicious URLs: Phishing websites are located on servers that have no relation with the legitimate website.

The phishing websites URL may contain the legitimate websites URL as a substring (http: //www.ebaymode.com), or may be similar to the legitimate URL (http://www.paypa1.com) in which the letter „L‟ in PayPal is substituted with number „1‟. IP addresses are sometimes used to mask the host name (http://25255255255/top.htm). Others use @ marks to make host names difficult to understand (http://ebay.com:top@255255255255/top.html) or contain suspicious usernames in their URLs (http://middleman/http://www.ebay.com).

- User input: Phishing websites typically contain pages for the user to enter sensitive information, such as account number, password and so on.
- Logos: The Phishing website uses logos found on the legitimate website to mimic its appearance. So phishers can load it from the legitimate website domain to their phishing websites (external domain).

- Short lived: Most phishing websites are available for only a few hours or days – just enough time for the attacker to defraud a high enough number of users.
- Sloppiness or lack of familiarity with English: Many Phishing pages have misspellings, grammatical errors, and inconsistencies.
- Copies: Attackers copy HTML from the legitimate websites and make minimal changes.

Detection Approach

Detect the phishing websites by checking the webpage source code, we extract some phishing properties out of the W3C standards to evaluate the security of the websites, and check each character in the webpage source code, if we find a phishing property, and we will increase the risk factor according to that property. Finally we calculate the total risk factor based on the phishing properties weights, the high percentage indicates phishing website and others indicates the website is most likely to be legitimate website.

Blacklist

When a user want to open a webpage, the URL of that page is compared to the blacklist if the URL is found in the list then that site is said to be a phishing site. Blacklists contain URLs and domain names. This is a very good approach for detecting pre-declared phishing sites.

URL Detection

The Uniform Resource Locator (URL) detection algorithm is based entirely on the URL of the website. This is a standalone test completed on the client machine that does not rely on information from outside sources such as an anti-phishing database or content from the webpage itself. Because many phishing websites try to impersonate the URL of the target or include suspicious items, this is a good start for detecting phishing sites that are not in a blacklist.

URL detection works best when it is combined with other detection techniques. Many of these techniques are common phishing detection techniques borrowed from other research while some new techniques such as keywords, port no. and URL redirection have been added. Every alphanumeric URL has a numeric only address associated with it. This numeric value is called an IP (Internet Protocol) address and is a sequence of numbers that uniquely identifies a computer on a network. Phishing URLs often contain IP addresses to hide the actual URL of the website.

Much like the URL can be represented with a numeric only value, each character on the keyboard can be represented with a numerical value that the computer understands. This numeric decimal value can easily be converted into hexadecimal base 16. Web browsers can understand hexadecimal values and they can be used in URLs by preceding the hexadecimal value with a '%' symbol. For example the value %20 is the hexadecimal equivalent of the space character on the keyboard. Typically phishing sites use hexadecimal values to disguise the actual letters and numbers in the URL.

When a computer connects to another computer the service uses a specific IP address and a port number. There are several standard port numbers shown in the table 1. These correspond to common services used in web browsers such as FTP, Gopher, web, secure web, and SOCKS. If a suspicious unknown port number is used the phishing score is increased because attackers often use different port numbers to bypass security detection programs that may monitor a specific port number. SpoofGuard identified two characters common in phishing URLs the '@' and '- 'characters. By far the most common and dangerous character used in phishing URLs is the '@' character. This character is used by web browsers to automatically pass a username to a secure site.

Table 1: List of Standard Port No. used in URLs

Name	Port No.
FTP	21
Gopher	70
Web	80
Secure Web	443
SOCKS	1080

The username proceeds the '@' symbol and the destination URL follows the '@' symbol. The problem is that if the website isn't setup to handle a secure connection, the web browser will navigate to the destination URL without any error message. For example the URL "http://www.bankofamerica.com@phishingsite.com" will navigate to the destination URL which is "phishingsite.com" and will attempt to login using "www.bankofamerica.com" as the username. Obviously this is a great way to disguise the actual URL of the website and combined with an IP address one can really hide the phishing site while the URL appears to be legitimate.

Phishing websites use keywords in the URL to lure victims to the phishing sites and provide a false sense of security. Researchers at Google and Johns Hopkins University have identified several common keywords found in phishing URLs. They are: secure, account, webscr, login, signin, banking, and confirm [19]. The phishing score increases if one of the keywords in this list is found in the URL. Phishing websites want to appear as legitimate as possible so they often contain the name of the company they are targeting. Most of these targeted sites are included in the company check.

The list includes eBay, Paypal, Volksbank, Wells Fargo, Bank of America, Private Banking, HSBC, Chase, Amazon, Banamex, and Barclays [19]. The total phishing score increases if one of the keywords in this list is found in the URL.

Anonymizers can conceal the actual URL and provide a way for users to surf a website anonymously. Phishers can use an anonymizer or web proxy to act as a middleman between the user and the phishing site. Generally this is used to conceal the actual phishing site's URL as the anonymizer's URL will include a specific parameter that translates to the actual URL that is displayed. The list of common anonymizers and proxies found in phishing websites is found in Table 2. If an anonymizer or proxy is found the phishing score increases.

Table 2: Web Proxies used by Phishing Sites

No.	Web Proxies
1	Web.archieve.org
2	www.schoolproxy7
3	Supahproxy.com
4	www.hidemyass.com
5	www.nijacloak.com
6	www.the-cloak.com
7	Freesurf11.info
8	www.proxco.info
9	www.behidden.com

Page – Content Check

Page content detection relies on detecting certain properties of the actual content found in the webpage source code. Many phishing detection solutions use page content detection in some form. The goal of a phishing website is to obtain sensitive confidential information from the user such as a username and password, social security number, credit card number, bank account number, or pin number. A user must enter some information into the phishing website in order for the information to leak. The most common way of entering information into a website is through an HTML input field. It begins with the phrase "<input" in HTML so any webpage with this tag should be further examined. . If the input tag is not present in the website then it is most likely not a phishing threat so one can examine the page for the input tag and decide whether to proceed with other detection methods.

Because phishing sites attempt to steal personal information, certain keywords are often used in the webpage. Keywords such as username, password, login, pin, account number, name, address, etc. are used to prompt the user to enter sensitive information. A phishing detection program could search for these keywords and add to a phishing score if they are present.

One of the easiest ways to copy a website and store it locally is to open the site in a web browser and choose the "File->Save As" option. This allows one to store a local copy of the current webpage in view. Often additional information is added to the top of the webpage in the form of a comment. A comment is a message inside the HTML code that is not visible when the webpage is viewed, but is visible if one examines the HTML source code. Comments are usually reserved for web programmers to make notes about the HTML code, but in some cases web browsers automatically add a comment to the webpage if it is stored locally. The comment will often begin with the phrase "saved from url" such as the following comment from a Google search engine page that was saved locally. For example: <!—saved from

url=(0061)https://www.google.co.in/?gws_rd=cr&ei=Mwy6Us-RN4qWrAf63YCQDg -->

System Design and Implementation

- A window opens demanding URL as input, where in the input box the URL is entered.
- Check the webpage security through URL comparison with blacklist.
- If found in the blacklist then a message displayed that “This is a Phishing Website”, otherwise, go to the next step.
- The source code of the URL entered is loaded automatically and streaming in read mode.
- Reads every line in the source code individually.
- Checks for each phishing property in the webpage source code.
- If there any phishing property found in the source code the risk factor will increase by the value based on the phishing property risk.
- After all the source code in read, the total security percentage will be evaluated based on the final risk to tell us if the website is legitimate or phish.

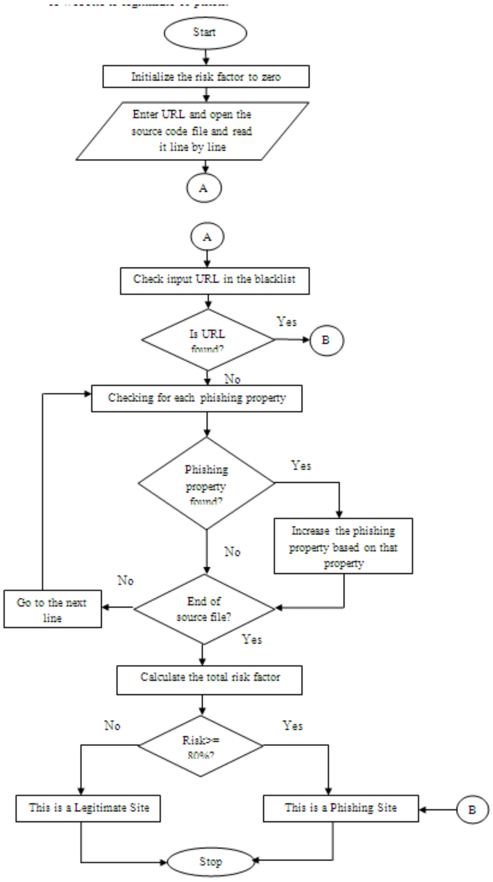
Start
Initialize the risk factor to zero
Enter URL and open the source code file and read it line by line
A
A
Check input URL in the blacklist
Is URL found?
Yes
B
No
Checking for each phishing property
Phishing property found?
Yes
Increase the phishing property based on that property
No
No
Go to the next line
End of source file?
Yes
Calculate the total risk factor
No
Risk>= 80%?
Yes
This is a Legitimate Site
This is a Phishing Site
B
Stop

Fig. 1: Flowchart of Proposed Approach

Experiment and Result

We conducted four experiments to assess the performance of proposed method.

Experiment 1 – Evaluation of Blacklist method

We chose the phishing sites from the phishtank [13] and test the blacklists. We ran the experiment for 3-4 hours a day for 5 weeks. During this time, batches of new unique phish were sent to the algorithm. It takes some time to examine if the site is phishing or not. For each website, the performance of the algorithm at hour 0, 1, 2, 3, 4, 5, 8, 12, 24, 48 was tested. The websites whose lifetime is almost 24 hours or more are detected by the blacklists others are not tested by this approach. We found that blacklist was ineffective when protecting users initially, as it caught less than 20% of phish at hour zero.

Experiment 2 – Evaluation of URL Method

Hundreds of phish were used to test this approach. The experiment ran 3-4 hours daily for 10 weeks. During the experiment whenever suspicious properties of the URL were detected the algorithm labelled them with appropriate weight. The result show that 80% of the good URLs were not detected as phishing which translates to a high false positive rate of 20%. The phishing URLs had 23% escape detection which translates to a true positive rate of 77%. These results show that detection by URL alone leaves much to be desired. After examining the result, one quickly realizes that the two major categories where the good URLs are being falsely detected are suspicious characters and keywords in the URL. The keyword detection is expected as keywords are commonly found in both legitimate and phishing URLs. The decision was made to remove the dash '–'from the suspicious character detection as legitimate sites also use this character. After removing the dash there was a drastic decrease in suspicious character detection from both sets of URLs.

Experiment 3 – Evaluation of Page content Check

This experiment focuses on *eBay* because it is one of the most targeted phishing sites. Each match found to contain "ebay" was manually reviewed by a human to verify that it was indeed a phishing site imitating eBay. Many sites were discarded from this experiment because they contained the word "ebay" and were sorted by the program, but after manual review they were found not to be eBay phishing sites. A total of 20 confirmed eBay phishing sites were used for this experiment. The result is out of 20, 19 phishing sites of eBay are detected as phishing and 1 is not detected. Overall the phishing detection algorithms performed excellent on the eBay phishing sites with a true positive rate of 99% and a false negative rate of 1%. *PayPal* is consistently listed in the top targets of phishing sites along with eBay so an experiment similar to the eBay experiment was conducted with PayPal phishing sites. A total of 20 confirmed PayPal phishing sites were used for this experiment. The result is out of 20 phishing sites, 20 sites are detected as phishing. The results are very similar to the eBay phishing sites. Here all the phishing sites are detected. Saved from comment and input field properties are also checked through the source code of 20 URLs. We saved all 20 websites locally to our computer and then check for <input tag and <!—saved tag

in the source code of locally stored websites. From saved from comments all 20 websites are detected as phishing.

Experiment 4 – Evaluation of complete Algorithm

To see the effectiveness of phishing properties have in impacting true positives and false positives, as well as comparing our overall performance with two popular anti phishing toolbars SpoofGuard and Netcraft. We select these because we found that SpoofGuard had the highest true positive rate and Netcraft was one of the best toolbars overall. All 100 URLs are English language sites. All the approaches in the algorithm finally did the best with 97% true positives in figure 2(a). SpoofGuard had a fairly high false positive rate of 48% and Netcraft is not much different from the final approach in the algorithm. The algorithm and Netcraft had 1% and 0%false positive rates respectively in figure 2(b).

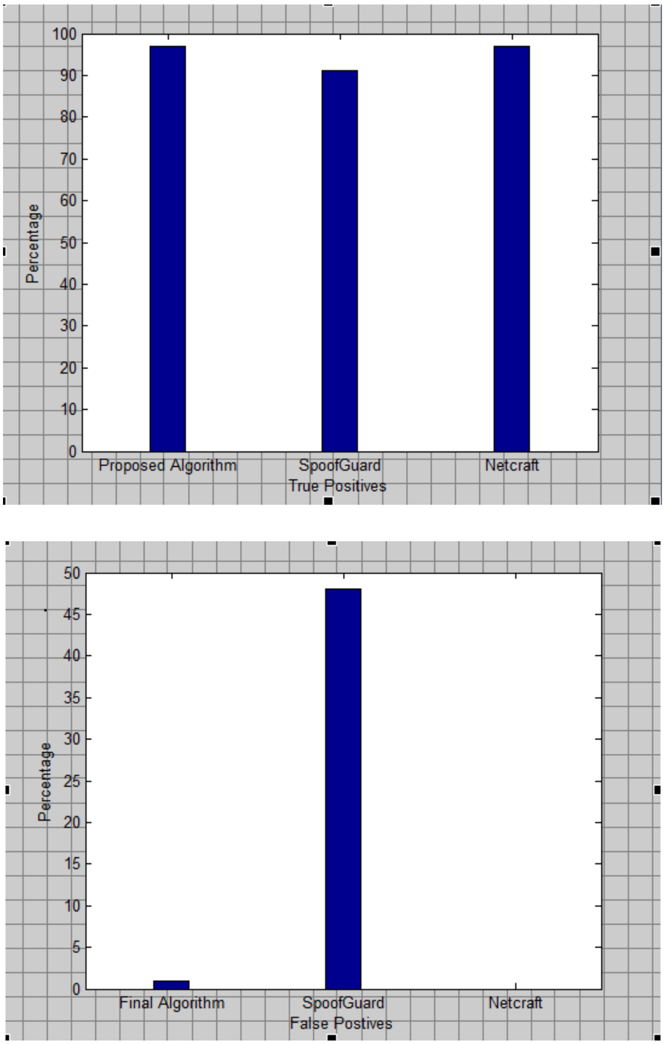

Figure 2(a): Comparison of TP; 2(b) Comparison of FP.

Conclusion

In this paper, we presented the design and evaluation of proposed approach. This method takes URL as input and examine that URL based on blacklist and phishing properties. Blacklist detects the previously detected phishing site through its database but with the help of phishing properties this approach can detect zero hour phish also. This is an idea to overcome the problem of zero day attacks. We described our implementation and discussed results also. We also presented the evaluation process, showing that complete approach can catch about 97% phishing sites with about 1% false positives.

In future work we plan on refining our approach in preparation for wider-scale deployment and evaluation. As previous research has shown, even if an anti-phishing toolbar is highly accurate, user's might still fall victim to fraud.

References

[1] APWG. *Phishing Activity Trends Report* Available:2013,available:http://docs.apwg.org/reports/apwg_trends_report_q4_2013.pdf.

[2] M. Aburrous, et al., "Predicting Phishing Websites Using Classification Mining Techniques with Experimental Case Studies", in *Seventh International Conference on Information Technology,* IEEE Conf, Las Vegas, Nevada, USA, 2010, pp. 176-181.

[3] J. Chen and C. Guo, "Online detection and prevention of phishing attacks", in *Proceedings of the fifth Mexican International Conference in Computer Science,* IEEE Conf, 2006, pp. 1-7.

[4] Net Applications. Inc. Browser market share q4, 2008. http://marketshare.hitslink.com/report.aspx?qprid=0&qpmr=15&qpdt=1&qpct=3&qpcal=1&qptimeframe=Q&qpsp=39.

[5] The 2012 internet crime report. 2012. The Internet Crime Complaint Centre (IC3). http://www.ic3.gov/media/annualreport/2012_IC3Report.pdf

[6] B. Ross, C. Jackson, N. Miyake, D. Boneh, and Mitchell, 2005. Stronger password authentication using browser extensions. In *Proceedings of the 14th Conference on USENIX Security Symposium – Volume 14* (Baltimore, MD, July 31 – August 05, 2005).

[7] P. Likarish, E. Jung, D. Dunbar, T. Hansen, and J. P. Hourcade B-APT: Bayesian Anti-Phishing Toolbar. In Communications, 2008. ICC '08. IEEE International Conference on Communications, 2008.

[8] PhishTank: http://www.phishtank.com/