WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Safe Web Search Using Phishing Detection with User Search Data Protection

Mr. Janarthanan M[1], Mr. Pandiyarajan[2]

[1](Department of CS, PG scholar, Rathinam College of Arts and Science, Coimbatore, thamizhjanathamizhjana@gmail.com)
[2](Department of CS, senior faculty, Rathinam College of Arts and Science, Coimbatore, pandi.knp@gmail.com)

ABSTRACT

Malicious URLs and websites pose a frequent and serious risk to online safety. Search engines naturally become the foundation of information management. However, the proliferation of fraudulent websites on search engines has put our users at serious risk. Most current techniques for detecting rogue websites concentrate on a particular assault. At the same time, numerous websites are unaffected by the readily available browser extensions based on blacklists. Since the server cannot deduce any useful information from the masked data, it is crucial that any data leaving the client side be effectively masked. The initial PPSB service is here suggested. Strong security guarantees are provided, something that is lacking in the current SB services. The proposed approach detects the malicious URL access with the help of blacklist storage. The input URL (given by user) was classified with the help of SVM classification. SVM is a type of machine learning algorithm that accurately detect, whether the URL is safe or unsafe. In particular, it carries over the capacity to recognise harmful URLs while safeguarding the browsing history and proprietary information of the blacklist provider (the list of unsafe URLs) as well as the user's privacy. In order to protect user privacy from outside analysts and service providers, a model that encrypts sensitive data was presented in this work. Additionally fully supports selective aggregate functions for analysing online user behaviour and ensuring differential privacy. Data about users' online behaviour is encrypted using the AES technique.

1. INTRODUCTION

Phishing is a social engineering attack that aims at exploiting the weakness found in system processes as caused by system users. For example, a system can be technically secure enough against password theft, however unaware end users may leak their passwords if an attacker asked them to update their passwords via a given Hypertext Transfer Protocol (HTTP) link, which ultimately threatens the overall security of the system. Moreover, technical vulnerabilities (e.g. Domain Name System (DNS) cache poisoning) can be used by attackers to construct far more persuading socially-engineered messages (i.e. use of legitimate, but spoofed, domain names can be far more persuading than using different domain names). This makes phishing attacks a layered problem, and an effective mitigation would require addressing issues at the technical and human layers. 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 [1], 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

2. SCOPE OF THE PROJECT

A malicious party might leverage PPSB (Privacy Preserving Safe Browsing) to degrade the client-side user experience, like inserting a number of fake or safe URLs or increasing the server-side delay. To address this potential issue, PPSB provides a flexible mechanism for users to add or remove blacklist providers. Admin could add the fake URL and keyword to this blacklist storage. User can also allow suggesting the malicious website details regarding black list. The objectives of proposed unsafe (or) malicious URL detection and search history encryption process has been listed below,

• Objective of this project is to analyzing the URL of the website to detect suspicious patterns such as misspelled domain names or unusual characters in the URL.

• A system can be developed to detect phishing websites in search engines and provide users with warnings and recommendations to avoid these sites.

• Need to supports selective aggregate functions for online user behavior analysis and guaranteeing differential privacy.

3. LITERATURE SURVEY

According to APWG, the term phishing was coined in 1996 due to social engineering attacks against America On-line (AOL) accounts by online scammers. The term phishing comes from fishing in a sense that fishers (i.e. attackers) use a bait (i.e. socially-engineered messages) to fish (e.g. steal personal information of victims). However, it should be noted that the theft of personal information is mentioned here as an example, and that attackers are not restricted by that as previously defined in Section II. The origins of the ph replacement of the character f in fishing is due to the fact that one of the earliest forms of hacking was against telephone networks, which was named Phone Phreaking. As a result, ph became a common hacking character replacement of f. According to APWG, stolen accounts via phishing attacks were also used as a currency between hackers by 1997 to trade hacking software in exchange of the stolen accounts. Phishing attacks were historically started by stealing AOL accounts, and over the years moved into attacking more profitable targets, such as on-line banking and e-commerce services. Currently, phishing attacks do not only target system endusers, but also technical employees at service providers, and may deploy sophisticated techniques such as MITB attacks.

1.6.1 Common attacks/vulnerabilities

B. Phishing Motives According to Weider D. et. al. [6], the primary motives behind phishing attacks, from an attacker's perspective, are: • Financial gain: phishers can use stolen banking credentials to their financial benefits. • Identity hiding: instead of using stolen identities directly, phishers might sell the identities to others whom might be criminals seeking ways to hide their identities and activities (e.g. purchase of goods). • Fame and notoriety: phishers might attack victims for the sake of peer recognition. C. Importance According to APWG, phishing attacks were in a raise till August, 2009 when the all-time high of 40,621 unique3 phishing reports were submitted to APWG. The total number of submitted unique phishing websites that were associated with the 40,621 submitted reports in August, 2009 was 56,362. As justified by APWG, the drop in phishing campaign reports in the years 2010 and 2011 compared to that of the year 2009 was due to the disappearance of the Avalanche gang4 which, according to APWG's 2nd half of 2010 report, was responsible for 66.6% of world-wide phishing attacks in the 2nd half of 2009 [7]. In the 1st half of the year 2011, the total number of submitted phishing reports to APWG.

D. Challenges

Because the phishing problem takes advantage of human ignorance or naivety with regards to their interaction with electronic communication channels (e.g. E-Mail, HTTP, etc. . .),

it is not an easy problem to permanently solve. All of the proposed solutions attempt to minimize the impact of phishing attacks.

From a high-level perspective, there are generally two commonly suggested solutions to mitigate phishing attacks:

• User education; the human is educated in an attempt to enhance his/her classification accuracy to correctly identify phishing messages, and then apply proper actions on the correctly classified phishing messages, such as reporting attacks to system administrators.

• Software enhancement; the software is improved to better classify phishing messages on behalf of the human, or provide information in a more obvious way so that the human would have less chance to ignore it.

The challenges with both of the approaches are:

• Non-technical people resist learning, and if they learnthey do not retain their knowledge permanently, and thus training should be made continuous. Although some researchers agree that user education is helpful [1], [11], [12], a number of other researchers disagree [13], [14].

Stefan Gorling [13] says that: "this is not only a question of knowledge, but of utilizing this knowledge to regulate behavior.

And that the regulation of behavior is dependent on many more aspects other than simply the amount of education we have given to the user"

• Some software solutions, such as authentication and security warnings, are still dependent on user behavior. If users ignore security warnings, the solution can be rendered useless.

• Phishing is a semantic attack that uses electronic communication channels to deliver content with natural languages (e.g. Arabic, English, French, etc) to persuade victims to perform certain actions. The challenge here is that computers have extreme difficulty in accurately understanding the semantics of natural languages. A notable attempt is E-mail-Based Intrusion Detection System (EBIDS) [15], which uses Natural Language Processing (NLP) techniques to detect phishing attacks, however its performance evaluation showed a phishing detection rate 4

Fig. 3. The life-cycle of phishing campaigns from the perspective of antiphishing techniques. of only 75%. In our opinion, this justifies why most well-performing phishing classifiers do not rely on NLP techniques.

IV. MITIGATION OF PHISHING ATTACKS: AN OVERVIEW

Due to the broad nature of the phishing problem, we find important to visualize the life-cycle of the phishing attacks, and based on that categorize anti-phishing solutions.

Based on our review of the literature, we depict a flowchart describing the life-cycle of phishing campaigns from the perspective of anti-phishing techniques, which is intended to be the most comprehensive phishing solutions flowchart. See Figure 3.

When a phishing campaign is started (e.g. by sending phishing emails to users), the first protection line is detecting the campaign. The detection techniques are broad and could incorporate techniques used by service providers to detect the attacks, end-user client software classification, and user awareness programs. More details are in Section IV-A.

The ability to detect phishing campaigns can be enhanced whenever a phishing campaign is detected by learning from such experience. For example, by learning from previous phishing campaigns, it is possible to enhance the detection of future phishing campaigns. Such learning can be performed by a human observer, or software (i.e. via a machine learning algorithm).

Once the phishing attack is detected, a number of actions could be applied against the campaign. According to our review of the literature, the following categories of approaches exist:

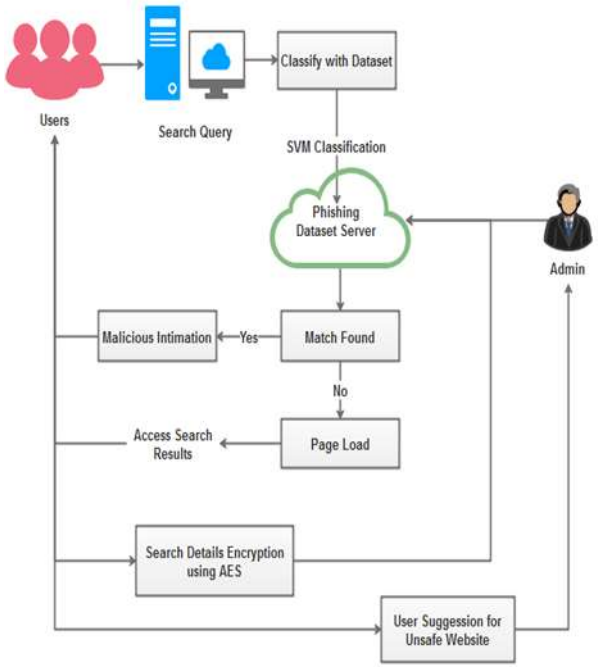
• Offensive defense — these approaches aim to attack phishing campaigns to render them less effective. This approach is particularly useful to protect users that have submitted their personal details to attackers. More details are in Section IV-B.

• Correction — correction approaches mainly focus on taking down the phishing campaign. In case of phishing websites, this is achieved by suspending the hosting account or removing phishing files.

4. Methodology

Software could be installed on the client or server side to inspect payloads of various protocols via different algorithms. Protocols could be HTTP, SMTP or any arbitrary protocol. Algorithms could be any mechanism to detect or prevent phishing attacks. Phishing heuristics are characteristics that are found to exist in phishing attacks in reality, however the characteristics are not guaranteed to always exist in such attacks. If a set of general heuristic tests are identified, it can be possible to detect zero-hour phishing attacks (i.e. attacks that were not seen previously), which is an advantage against blacklists (since blacklists require exact matches, the exact attacks need to be observed first in order to blacklist them). However, such generalized heuristics also run the risk of misclassifying legitimate content (e.g. legitimate emails or websites). Currently, major web browsers and mail clients are built with phishing protection mechanisms, such as heuristic tests that aim at detecting phishing attacks. The clients include Mozilla FireFox, Internet Explorer, Mozilla Thunderbird and MS Outlook. Also, phishing detection heuristics could be included in Anti Viruses, similar to ClamAV10 .

BLOCK DIAGRAM:



5. EXPERIMENTAL SETUP

Julie S. Downs et al. [20] surveyed 232 computer users to study what are the criteria that can predict the susceptibility of a user to fall victims for phishing emails. The survey was formed in a role play where each user was expected to analyze emails as well as answering a number of questions. The outcome of the study was that those who had a good knowledge about the definition of "phishing" were significantly less likely to fall for phishing emails, while knowledge about other areas, such as cookies, spyware and viruses did not help in reducing vulnerability to phishing emails. Interestingly, the survey showed that knowledge about negative consequences (e.g. credit card theft) did not help in reducing vulnerability to phishing emails. The study concluded that user educational messages should focus on educating users about phishing attacks rather than warning them about the dangers of negative

consequences. Another study that confirms the study in [20] was made by Huajun Huang et. al. [21], which concluded that the primary reasons that lead technology users to fall as victims for phishing attacks are: • Users ignore passive warnings (e.g. toolbar indicators). • A large number of users cannot differentiate between phishing and legitimate sites, even if they are told that their ability is being tested. A demographic study made by Steve Shen et. al. [1] shows a number of indirect characteristics that correlate between victims and their susceptibility to phishing attacks.

5.2 PROJECT OVERVIEW

This section presents our view on the various phishing detection approaches that are presented in the survey. A. User education and awareness Since phishing is a social engineering attack, an obvious solution can be educating the end-user. However, as discussed in [13] education and user awareness — alone — are not enough, and that what is needed is regulating the behavior of end-users instead. Various incidents, such as the one described in [2], have shown that even security providers themselves have fallen victims for phishing attacks. Although phishing seems to be a simple attack that exploits the naivety of end-users, it is able to persuade security-aware engineers as well. This indicates the possibility that systems' complexities are raising beyond the cognition limits of many humans, and that simply educating them is not enough. A more promising solution could be enhancing the system usability via: • Better user interfaces. It is visible to us that the software industry is moving towards this direction. For example, older versions of web browsers used passive warnings, while recent ones moved to active warnings. A security warning should be active (i.e. blocks the content) and should visually hint the user of risks even without reading its content (since most end-users do not read warning messages [23]). • Enhancing the behavior of the systems, so that the harmful messages are automatically detected and quarantined on behalf of the end-user. For example, blacklists, heuristic rules and ML techniques can be used to automatically filter harmful content from end-users. Such features are currently implemented in web browsers, email clients and server-side filters. The ideal phishing mitigation direction seems to us to be debatable so far, which could be due to the short history depth of information technology in general. However, in our opinion, systems are in reality moving towards adapting to their end-users (as opposed to having end-users adapting to their systems). B. Blacklists Blacklists are effective when minimal F P rates are required, which is achieved due to the way blacklists are constructed (e.g. many of them involve human administration, such as PhishTank). Blacklists also have the advantage of requiring low resources on the host machine. This elevates the need of extensively analyzing the content of websites and emails. However, blacklists are known to be behind the line when the objective is mitigation of zero-hour phishing attacks

5.3 MODULES DESCRIPTION

MODULES

- Framework Construction
- User Registration and Login
- URL Search
- Unsafe URL Detection
- Search URL Encryption
- Access Search History
- Feedback System

FRAMEWORK CONSTRUCTION

The detection of malicious URLs limits web-based attacks by preventing web users from visiting malicious URLs and warning web users prior to accessing content located at a malicious URL. Thus, malicious URL detection protects computing system hardware/software from computer viruses, prevents execution of malicious or unwanted software, and helps avoid accessing malicious URLs web users do not want to visit. This proposed framework uses SVM classification models to detect a malicious URL and categorize the malicious URL as one of a phishing URL. The blacklist storage models by using a set of training data (unsafe URLs and keywords) and machine learning algorithms. The training data includes a known set of unsafe URLs and a known set of malicious keywords. This framework also supports URL encryption process, to avoid the unauthorized prediction of URL details.

D. Machine Learning-based classifiers Similar to heuristic tests, ML-based techniques can mitigate zero-hour phishing attacks, which makes them advantageous when compared with blacklists. Interestingly, ML techniques are also capable of constructing their own classification models by analyzing large sets of data. This elevates the need of manually creating heuristic tests as ML algorithms are able to find their own models. In other words, ML techniques have the following advantages over heuristic tests: • Despite the complicated nature of adversarial attacks, it is possible to construct effective classification models when large data set samples are available, without the need of manually analyzing data to discover complex relationships. • As phishing campaigns evolve, ML classifiers can automatically evolve via reinforcement learning. Alternatively, it is also possible to periodically construct newer classification models by simply retraining the learner with updated sample data sets. ML-based anti-phishing classifiers in the literature, such as those presented in [4], [46], have shown that it is possible to achieve less than 1% F P, and more than 99% T P rates. According to the surveyed literature, ML-based classifiers are the only classifiers that achieved such high classification accuracy while maintaining their ability to detect zero-hour phishing attacks.

FEEDBACK SYSTEM

Feedback system helps to overcome the problems faced by user during web search process. User can send their feedback regarding, search efficiency. Also they will be allowed to provide suggestion for adding further URLs in blacklist. Admin can view URL suggestion provided by user, and add the malicious URLs in blacklist. This helps to enhance the performance of blacklist storage in phishing detection.

6. CONCLUSIONS

In this proposed work, implement a Malicious URL Detection process using machine learning techniques. This focuses on detecting unsafe website URLs and keywords with the help of encrypted blacklist storage. According to few selected features can be used to differentiate between legitimate and malicious web pages. These selected features are many such as URLs and Keywords. In proposed work a service provider that owns a high-quality blacklist, which may be more frequently updated or simply contains more items. User also allowed to directly sharing blacklists with servers in an uncontrollable way could make these dataset be obtained by every user. With the help of efficient classification approach will detect the fake websites accurately and prevent the users from accessing that websites. This also provides the secure encryption approach avoid the unknown access of search history. The security is provided to the search data which has been stored in the database.

7. FUTURE WORKS

The Future work is to fine tuning the machine learning algorithm that will produce the better result by utilizing the large URL dataset. Also implement a robust malware detection method, retaining accuracy for phishing emails.

8. REFERENCES

1.ASP.NET BOOK REFERENCES

1. ".NET: Interview Questions" by Shivprasad Koirala

2. "Dot Net Technology" by Damini Grover

3. "Functional Concurrency in .net: With Examples in C# and F#" by Riccardo Terrell

4. ".NET 4.5 Programming 6-in-1, Black Book" by cogent Learning Solutions Inc

5. "Visual Basic .NET Programming Black Book" by Steven Holzner

6. "Murach's V.B.NET database Programming with ADO.NET" by Anne Prince

2. WEBSITE REFERENCE

1.https://www.geeksforgeeks.org/introduction-to-net-framework/

2. https://www.guru99.com/net-framework.html

3.https://dotnet.microsoft.com/learn/dotnet/what-is-dotnet-framework

4.https://docs.microsoft.com/en-us/dotnet/framework/

5.https://www.tutorialspoint.com/net_framework_online_training/index.asp