WEB PHISHING DETECTION – A LITERATURE SURVEY

This article surveys the literature on the detection of phishing attacks. 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, 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

Due to the broad nature of the phishing problem, this phishing detection survey begins by:

Defining the phishing problem. It is important to note that the phishing definition in the literature is not consistent, and thus a comparison of a number of definitions is presented.

Categorizing anti-phishing solutions from the perspective of phishing campaign life-cycle. This presents the various anti-phishing solution categories such as *detection*. It is important to view the overall anti-phishing picture from a high-level perspective before diving into a particular technique, namely: phishing detection techniques (which is the scope of this survey).

Presenting evaluation metrics that are commonly used in the phishing domain to evaluate the performance of phishing detection techniques. This facilitates the comparison between the various phishing detection techniques.

Presenting a literature survey of anti-phishing detection techniques, which incorporates software detection techniques as well as user-awareness techniques that enhance the detection process of phishing attacks.

Presenting a comparison of the various proposed phishing detection techniques in the literature.

This survey begins by defining the phishing problem, presenting background and related works. Next in this survey will then focus on phishing detection techniques, which include detection techniques through user awareness. The evaluations of the surveyed detection techniques are presented. The conclusion is finally drawn.

DEFINITION

The definition of phishing attacks is not consistent in the literature, which is due to the fact that the phishing problem is broad and incorporates varying scenarios. For example, according to PhishTank:

"Phishing is a fraudulent attempt, usually made through email, to steal your personal information"

PhishTank's definition holds true in a number of scenarios which, roughly, cover the majority of phishing attacks (although no accurate studies have been made to reliably quantify this). However, the definition limits phishing attacks to stealing personal information, which is not always the case.

Another definition is provided by Colin Whittaker et. al.

"We define a phishing page as any web page that, without permission, alleges to act on behalf of a third party with the intention of confusing viewers into performing an action with which the viewer would only trust a true agent of the third party"

The definition by Colin Whittaker et. al. aims to be broader than PhishTank's definition in a sense that attackers goals are no longer restricted to stealing personal information from victims. On the other hand, the definition still restricts phishing attacks to ones that act on behalf of third parties, which is not always true.

For example, phishing attacks may communicate socially engineered messages to lure victims into installing MITB malware by attracting the victims to websites that are supposed to deliver safe content (e.g. video streaming). Once the malware (or crimeware as often named by Anti-Phishing Working Group (APWG)[2]) is installed, it may log the victim's keystrokes to steal their passwords.

In order to address the limitations of the previous definitions above, we consider phishing attacks as semantic attacks which use electronic communication channels (such as E-Mails, HTTP, SMS, VoIP, etc…) to communicate socially engineered messages to persuade victims to perform certain actions (without restricting the actions) for an attacker's benefit (without restricting the benefits).

Definition 1

Phishing is a type of computer attack that communicates socially engineered messages to humans via electronic communication channels in order to persuade them to perform certain actions for the attacker's benefit.

HISTORY AND BACKGROUND WORKS

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.

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 end-users, but also technical employees at service providers, and may deploy sophisticated techniques such as MITB attacks.

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.

According to APWG, phishing attacks were in a raise till August, 2009 when the all-time high of 40,621 unique 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 gang 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. In the 1st half of the year 2011, the total number of submitted phishing reports to APWG was 26,402, which is 35% lower than that of the peak in the year 2009.

On the other hand, the 2nd half of 2011 saw a raise in phishing reports and websites, which seems to be correlated with holidays season. The year 2011 saw a number of notable spear phishing attacks against well-known security firms such as RSA and HB Gary, which resulted in further hacks against their clients such as RSA's client Lockheed Martin. This shows that the dangers of phishing attacks, or security vulnerabilities due to the human factor, are not limited to the naivety of end-users since technical engineers can also be victims.

Minimizing the impact of phishing attacks is extremely important and adds great value to the overall security of an organization.

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 learn they do not retain their knowledge permanently, and thus training should be made continuous. Although some researchers agree that user education is helpful, a number of other researchers disagree. Stefan Gorling says that:

"This is not only a question of knowledge, but of utilizing this knowledge to regulate behaviour. And that the regulation of behaviour 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 behaviour. 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), which uses Natural Language Processing (NLP) techniques to detect phishing attacks, however its performance evaluation showed a phishing detection rate of only 75%. In our opinion, this justifies why most well-performing phishing classifiers do not rely on NLP techniques.

DETECTION APPROACHES

In this survey, we consider any anti-phishing solution that aims to identify or classify phishing attacks as detection solutions. This includes:

User training approaches — end-users can be educated to better understand the nature of phishing attacks, which ultimately leads them into correctly identifying phishing and non-phishing messages. This is contrary to the categorization in where user training was considered a preventative approach. However, user training approaches aim at enhancing the ability of end-users to detect phishing attacks, and thus we categorize them under "detection".

Software classification approaches — these mitigation approaches aim at classifying phishing and legitimate messages on behalf of the user in an attempt to bridge the gap that is left due to the human error or ignorance. This is an important gap to bridge as user-training is more expensive than automated software classifiers, and user-training may not be feasible in some scenarios (such as when the user base is huge, e.g. PayPal, eBay, etc…).

The performance of detection approaches can be enhanced during the learning phase of a classifier (whether the classifier is human or software). In the case of end-users, their classification ability can be enhanced by improving their knowledge of phishing attacks by learning individually through their online experience, or by external training programs. In the case of software classifiers, this can be achieved during the learning phase of a Machine Learning-based classifier, or the enhancement of detection rules in a rule-based system.

Detection techniques not only help in *directly* protecting end-users from falling victims to phishing campaigns, but can also help in enhancing phishing honeypots to isolate phishing spam from non-phishing spam.

It is also important to note that the detection of phishing attacks is the starting point of the mitigation of phishing attacks. If a phishing campaign is not detected, none of the other mitigation approaches can be applicable. For example, all of the mitigation techniques, such as *correction, prevention* and *offensive defence* depend on a functional and accurate *detection* phase.

Offensive defence solutions aim to render phishing campaigns useless for the attackers by disrupting the phishing campaigns. This is often achieved by flooding phishing websites with fake credentials so that the attacker would have a difficult time to find the real credentials.

Two notable examples are:

- **BogusBiter** — A browser toolbar that submits fake information in HTML forms whenever a phishing website is encountered. Instead of simply showing a warning message to the end-user whenever a phishing website is visited, BogusBiter also submits fake data into HTML forms of the visited phishing website. Submitting fake data into the HTML forms is intended to disrupt the corresponding phishing campaigns, with the hope that such fake data may make the attackers task of finding correct data (among the fake data) more difficult.
- **Humboldt** — Similar to BogusBiter, except that BogusBiter relies on submissions from end-user clients, while Humboldt relies on distributed and dedicated clients over the Internet instead of end-user toolbars that may visit phishing sites, in addition to a mechanism to avoid causing DOS floods against servers. This can make Humboldt more effective against phishing websites due to the more frequent submission of data to phishing pages.

Once a phishing campaign is detected, the correction process can begin. In the case of phishing attacks, correction is the act of taking the phishing resources down. This is often achieved by reporting attacks to Service Providers.

Phishing campaigns often rely on resources, such as:

- Websites — could be a shared web host owned by the phisher, a legitimate website with phishing content uploaded to it, or a number of infected end-user workstations in a botnet[6].
- E-mail messages — could be sent from a variety of sources, such as: free E-mail Service Provider (ESP) (e.g. Gmail, Hotmail, etc…), open Simple Mail Transfer Protocol (SMTP) relays or infected end-user machines that are part of a botnet.
- Social Networking services — web 2.0 services, such as Facebook and Twitter, can be used to deliver socially engineered messages to persuade victims to reveal their passwords.
- Public Switched Telephone Network (PSTN) and Voice over IP (VoIP) — similar to other forms of phishing attacks, attackers attempt to persuade victims to perform actions. However, the difference is that attackers attempt to exploit spoken dialogues in order to collect data (as opposed to clicking on links). Moreover, due to the way VoIP protocols (e.g. Session Initiation Protocol (SIP)) function, and the way many VoIP provider systems are configured, spoofing Caller IDs are used by attackers as tools to increase their persuasion.

In order to correct such behavior, responsible parties (e.g. service providers) attempt to take the resources down. For example:

- Removal of phishing content from websites, or suspension of hosting services.
- Suspension of email accounts, SMTP relays, VoIP services
- Trace back and shutdown of botnets.

This also extends to the shutdown of firms that frequently provide services to phishing attackers.

The shutdown process can be initiated by organizations that provide brand protection services to their clients, which may include banking and financial companies that are possible victims of phishing attacks. When phishing campaigns are identified, they can be reported to their hosting

Internet and web hosting service providers for immediate shutdown. Depending on the country where phishers and phishing campaigns exist, the penalties and procedures can differ.

The “prevention” of phishing attacks can be confusing, as it can mean different things depending on its context:

- Prevention of users from falling victim — in this case, phishing detection techniques will also be considered prevention techniques. However, this is not the context we refer to when “prevention” is mentioned in this survey.
- Prevention of attackers from starting phishing campaigns — in this case, law suits and penalties against attackers by Law Enforcement Agencies (LEAs) are considered as prevention techniques.

In this survey, whenever the keyword “prevention” is mentioned, it refers to the second previous item which is minimizing the possibility of attackers starting phishing campaigns via LEA.

Usually, LEA may take a number of weeks to complete their investigation and response procedures. Thus, it is common to apply prevention techniques after all other mitigation techniques, which is due to the expensive nature of LEA investigations that makes them consume a relatively large period of time.

Once the sources of the phishing attacks are traced, LEA can then file law suits which in turn may issue penalties such as: imprisonment, fines and forfeiture of equipments used to convey the attacks.

PHISHING DETECTION

Phishing detection are done by the following approaches:

- **BACKLISTS**: Blacklists are frequently updated lists of previously detected phishing URLs, Internet Protocol (IP) addresses or keywords. Whitelists, on the other hand, are the opposite, and could be used to reduce *FP* rates. Blacklists do not provide protection against zero-hour phishing attacks as a site needs to be previously detected first in order to be blacklisted. However, blacklists generally have lower *FP* rates than heuristics. Blacklists are found to be ineffective against zero-hour phishing attacks, and were able to detect only 20% of them. The study also shows that 47% to 83% of phishing URL were blacklisted after 12 hours. This delay is a significant issue as 63% of phishing campaigns end within the first 2 hours.
- **HEURISTICS**: 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 ClamAV.

- **DATA MINING**: Techniques that are described in this section consider the detection of phishing attacks as a document classification or clustering problem, where models are constructed by taking advantage of Machine Learning and clustering algorithms, such as k-Nearest Neighbors (k-NN), C4.5, Support Vector Machines (SVM), k-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). For example, k-NN stores training instances in memory which are represented as multi-dimensional vectors, where each vector component represents the extracted value from a particular feature (e.g. number of URLs in an email message). The classification task is then performed by similarly processing testing instances, and calculating the distance (e.g. euclidean distance) between the testing instance and the other training instances. When K=3, the classes of the 3 nearest neighbors (as obtained during the training phase) are considered. When the task is classification, majority voting can be used to determine the class of the testing instance. Algorithms such as *C4.5* and SVM follow a different approach where they generalize a classification model (as opposed to k-NN, which does not generalize a model). For example, *C4.5* constructs a decision tree that should be generic enough to correctly classify unseen instances. The decision tree is composed of nodes with splitting branches. The splitting is generally performed to maximize the conditional Information Gain after the split. On the other hand, SVM aims at finding an effective separation plane in a vector space by analyzing the training instances. The separation plane should be generic enough so that it should still be able to separate unseen instances. However, clustering algorithms such as k-means and DB-SCAN partition instances in an unsupervised manner (i.e. knowing the class label is not required to construct the clusters). k-means algorithm aims at constructing k partitions by randomly setting k initial partition centers, followed by iteratively assigning instances to a partition with the smallest distance (e.g. euclidean distance) towards its center, and then updating the partition center to be the mean of the instances in the same partition. This iterative process is repeated until the clusters converge. On the other hand, DBSCAN is able to partition the data based on the density (i.e. using a distance function measure, such as euclidean distance) of the instances. Contrary to k-means, DBSCAN does not need to know beforehand the number of partitions that should be found, which is achieved by the concept of *density reachability*

EVALUATION METRICS

Based on our review of the literature, the following are the most commonly used evaluation metrics:

- True Positive (*TP*)rate — measures the rate of correctly detected phishing attacks in relation to all existing phishing attacks.
- False Positive (*FP*)rate — measures the rate of legitimate instances that are incorrectly detected as phishing attacks in relation to all existing legitimate instances.
- True Negative (*TN*)rate — measures the rate of correctly detected legitimate instances in relation to all existing legitimate instances.
- False Negative (*FN*)rate — measures the rate of phishing attacks that are incorrectly detected as legitimate in relation to all existing phishing attacks.
- Precision (P) — measures the rate of correctly detected phishing attacks in relation to all instances that were detected as phishing.

- Recall (R) — equivalent to TP.
- f1 score — Is the harmonic mean between P and R.
- Accuracy (*ACC*)— measures the overall rate of correctly detected phishing and legitimate instances in relation to all instances. See Equation (8) for details.
- Weighted Error (WErr) — measures the overall weighted rate of incorrectly detected phishing and legitimate instances in relation to all instances.

User education or training is an attempt to increase the technical awareness level of users to reduce their susceptibility to phishing attacks.

It is generally assumed that the addition of user education materials compliments technical solutions (e.g. classifiers). However, the human factor is broad and education alone may not guarantee a positive behavioural response.

As shown in the previous sections, most of the educational materials were also associated with a decrease in the *TN* rate, with an exception of only one educational material, namely: *Anti-Phish Phil*. This shows that the addition of user training approaches is not *always* the right answer.

CONCLUSION

User education materials can complement software solutions. However, it should also be noted that none of the existing studies empirically show enough evidence that user education can practically complement software solutions. This is due to the fact that all of the publicly available user education studies have evaluated educational materials independently from software solutions.

The study concludes that *Anti-Phish Phil* training material reduced *FN* rate from 46% to 29%, which is not enough evidence to assume that it would also complement software solutions that, for example, achieve a *FN* rate of less than 1%. The un-answered question is: what is the percentage of overlap between the classification performed by end-users following a user training phase, and the classification performed by a software classifier? If the overlap is 100%, then the addition of user training can be redundant and will not be worth the added cost and complexity. However, if the overlap is less than 100%, then they can be complementary to each other — however, such a study is not available in the public literature.

This survey reviewed a number of anti-phishing software techniques. Some of the important aspects in measuring phishing solutions are:

Detection accuracy with regards to zero-hour phishing attacks. This is due to the fact that phishing websites are mostly short-lived and detection at hour zero is critical.

Low false positives. A system with high false positives might cause more harm than good. Moreover, end-users will get into the habit of ignoring security warnings if the classifier is often mistaken.

Generally, software detection solutions are:

- Blacklists.
- Rule-based heuristics.
- Visual similarity.
- Machine Learning-based classifiers.

The findings in Section XII show that the use of Machine Learning techniques is promising as they have led to the most effective phishing classifiers in the publicly known literature. The Machine Learning-based detection techniques achieved high classification accuracy for analysing similar data parts to those of rule-based heuristic techniques.

As a future work in this field, it would be beneficial to conduct a study that:

- measures the effect of the addition of user training from the perspective of correcting software classification mistakes.
- analyses the phishing detection techniques from the perspective of their computational cost and energy consumption.