WEB PHISHING DETECTION: LITERATURE SURVEY

INTRODUCTION:

Phishing attacks are the practice of sending fraudulent communications that appear to come from a reputable source. It is usually done through email. The goal is to steal sensitive data like credit card and login information, or to install malware on the victim's machine. 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.

Since phishing attacks aim at exploiting weaknesses found it is difficult to mitigate them. 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.

This 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.

HISTORY:

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

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.

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 learn they do not retain their knowledge permanently, and thus training should be made continuous. 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.

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 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. . .).

EVALUATION METRICS:

In any binary classification problem, where the goal is to detect phishing instances in a dataset with a mixture of phishing and legitimate instances, only four classification possibilities exist.

	Classified as phishing	Classified as legitimate
Is phishing	*NP→P*	*NP→L*
Is legitimate	*NL→P*	*NL→L*

NP→P is the number of *phishing* instances that are correctly classified as *phishing*,
NL→P is the number of *legitimate* instances that are incorrectly classified as *phishing*,
NP→L is the number of *phishing* instances that are incorrectly classified as *legitimate*, and
NL→L is the number of *legitimate* instances that are correctly classified as *legitimate*.

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.
- *f*1 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.
- Weighted Error (*WErr*) — measures the overall weighted rate of incorrectly detected phishing and legitimate instances in relation to all instances.

DETECTION OF PHISHING ATTACKS:

Inputs to the decision making process are:

- External information: could be anything learned through the User Interface (UI) (Web/mail client and their content), or expert advice. The phisher only has control over what is presented by the UI. Usually, the user does not ask for expert advice unless he is in doubt (i.e. if a user is convinced that a phishing site is legitimate, he might not ask for expert advice in the first place).
- Knowledge and context: the user's current understanding of the world, which is built over time (e.g. news, past experience).
- Expectation: users have expectations based on their understanding and the outcome of their actions. During the decision making process, two types of decisions can be made, which are:
- Planning a series of actions to be taken.
- Deciding on the next action in sequence to be taken. This is influenced by the outcome resulting from the previous action.

Each of the two types of decisions mentioned above, follow the following steps:

- Construction of perception: constructed through the context where the user reads (say) an email message. Such as, senders/recipients, conversation cause, or suggested actions by the email. In legitimate messages, there are no inconsistencies between the reality and message claims (e.g. senders are the real senders whom they claim to be, and suggested actions by email content does what it says). However, in phishing messages there are inconsistencies (e.g. if the sender's ID is spoofed, or the message's content claims to fix a problem while attempting, in reality, to obtain personal

information). If the end-user discovers inconsistencies in a given phishing message, the phishing attack would then fail to persuade the end user.

- Generation of possible solutions: users usually find solutions through available resources. However, with phishing emails, the user is not requested to generate a possible solution in the first place, as the phisher already suggests a solution to the user. For example, if the phishing email content presents a problem, such as account expiry, it will also present a solution, such as activating the account through logging in a URL from which expiry is prevented.
- Generation of assessment criteria: different users have different criteria that reflects how they view the world, their emotional state, personal preferences, etc. . . . As the paper claims, most phishing attempts do not take into account such details, but rely on generic common-sense criteria instead; for example: an attacker might place a tick box labelled "Secure login" to meet a security criterion most users require. Phishing attacks aim to match user criteria as much as possible.

PHISHING DETECTION BY BLACKLISTS:

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.

PROPOSED RULES FOR DETECTION:

The proposed rules fall under:

- Analysis performed on URL that fall within the email's body.
- Analysis performed on email headers.

The proposed rules are (where *positive* indicates phishiness):

- Rule 1: If a URL is a login page that is not a business's real login page, the result is positive. The paper specifies that this is analyzed based on data returned from search engines.
- Rule 2: If the email is formatted as HTML, and an included URL uses Transport Layer Security (TLS) while the actual Hypertext Reference (HREF) attribute does not use TLS, then the result is positive.
- Rule 3: If the host-name portion of a URL is an IP address, the result is positive.
- Rule 4: If a URL mentions an organization's name (e.g. PayPal) in a URL path but not in the domain name, the result is positive.
- Rule 5: If URL's displayed domain does not match the domain name as specified in HREF attribute, the result is positive.
- Rule 6: If the received SMTP header does not include the organization's domain name, the result is positive.
- Rule 7: If inconsistencies are found in a non-image URL's domain portion, the result is positive.
- Rule 8: If inconsistencies are found in Whois records of non-image URL's domain portion, the result is positive.
- Rule 9: If inconsistencies are found in image URL's domain portion, the result is positive.

- Rule 10: If inconsistencies are found in Whois records of image URL's domain portion, the result is positive.
- Rule 11: If the page is not accessible, the result is positive.

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).

CONCLUSION:

User education or training is an attempt to increase the technical awareness level of users to reduce their susceptibility to phishing attacks. However, the human factor is broad and education alone may not guarantee a positive behavioral response.

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 Machine Learning-based detection techniques achieved high classification accuracy for analyzing similar data parts to those of rule-based heuristic techniques.