e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

A COMPREHENSIVE STUDY ON PHISHING ATTACKS AND DIFFERENT AVAILABLE DETECTION APPORACHES

Janhavi V*1, Neha R*2, Anagha B C*3, Megha Ambi*4, Bharathesh B S*5

*1Associate Professor, Department Of Computer Science, VVCE, Mysuru, Karnataka, India

*2,3,4,5Student, CSE, VVCE, Mysuru 570002, Karnataka, India.

ABSTRACT

This paper mainly concentrates on the fraudulent phishing attacks along with their detection techniques. As the technological advancements are happening at a large scale the world is becoming small. Due to these technological advancements the whole world is going towards a digital society. Due to these changes, there comes a plethora of cybercrimes such as Phishing, Cyberextortion, Ransomware attack, Crypto jacking, Cyberespionage, and many more Phishing attack being one of the most fraudulent one, extreme care and awareness must be taken to safeguard our data.

Keywords: Phishing Attacks, Fraudulent Method, Plethora, Data Mining, Visual Similarity, Heuristics.

I. INTRODUCTION

Phishing is a social engineering attack which mainly happens by lack of knowledge and awareness of an end user. Even though the systems are physically secured with passwords, face detection or fingerprint detection it is incredibly important to make sure that it has system security efficiency. As the number of people using system and software have increased there is an equal increase in the number of attackers who uses the innocence of the victims for their personal benefits.

Phishing attack being the major one has different kinds in it namely clone phishing, domain spoofing, evil, twin phishing, HTTPS phishing, smishing, spear phishing, wishing, watering hole phishing. Looking into all these types it is tremendously important to know in detail about their detection types.

This paper has a detailed information about all necessary facts and figures along with methods to tackle the attack. This study-paper begins with the basic definition of phishing in Section I, moving to the background and history of the phishing in Section II the alleviation of the phishing attacks in Section IV, different evaluation units and metrics in Section V, different detection techniques in Section VI, VII, VIII, IX, X. The evaluations of human training approaches are presented in Section XI. In addition to that the lessons learnt from the detailed study of the concept is ascertained in Section XII. Ultimately the conclusion of the study is putdown in Section XIII.

II. DEFINITION

Since phishing has many shades and happens in a wide range, a proper definition of phishing attack is still skeptical. To cite few:

1. As per PhishTank [1] - "Phishing is a fraudulent attempt, usual maid through Email, to steal one's personal information."

Even though this definition from PhishTank holds good in most cases, it is not limited to stealing personal information from the victim. To substantiate this constraint, we shall look into the example of a socially engineered message, which induces the victim to install MITB (Man In the Browser) malware. This MITB will help the attacker transfer enough amount from the victim's account whenever the sufferer logs in to perform any financial tasks. This loot happens ultimately without stealing any of the victim's personal information.

2. As per the definition by Colin Whittaker et.al. [4]- "We define a phishing page as any webpage but, without permission, alleges to act on behalf of a third party with the intention of the confusing viewer with which the viewer would only trust a true agent of the third party."

The above definition defines more broadly when compared to the PhishTank description. This mentions that a phishing attack is not restricted to stealing personal information but instead mentions the attacker's cunning mentality. Even though this is an improvised definition, it fails to explain the phishing attacks on behalf of third parties.

To overcome all these limitations from the already available definitions, by referring and understanding numerous phishing cases, we have defined phishing as – "Phishing attack is a fraudulent type of computer

attack that helps the attacker to perform specific tasks for his/her benefit by influencing the victim to share the necessary information or by hacking the systems to fulfill their greedy goals.

III. BACKGROUND

History

In 1996, there was an attack on America On-line by online attackers, due to which, for the first time, world phishing came into existence. The phrase phishing is derived from fishing, i.e., hypothetically, the fishers (attackers) use a lure (malicious messages) for fishing (gather personal information of the user).

Basically, ph substitute for attribute f in the word fishing because the earlier form of hacking is adverse to telecommunication networks, labelled as Phone Phreaking. So ph became the substitutions of f. In 1997 the hackers started selling the hacking programs in exchange for the thieved account.

Thieving AOL accounts initially started phishing attacks. As years passed, they entered attacking, more profitable prospects like e-commerce and online banking services.

At present, attackers focus on end-user and target technical employees and may bring more complex techniques like MITB attacks.

Phishing motives

The primary motives of phishing attacks in from an according to attacker's point of view:

1. Financial gain: Attackers can use thieved banking details for their capital gain.
2. Identity hiding: Attackers might tell thieved identities to others (criminals) who find ways to hide their identities and activities.
3. Fame and notoriety: Attackers will attack end-users to get peer limelight.

Importance

As per the APWG[2], till August 2009, phishing attacks are detected in a high ratio; then, there was sudden rise of 40,621 different phishing attack reports, which was submitted to APWG. There were 56,362 various phishing sites related to the 40,621 phishing tasks submitted in 2009 August. As defended by APWG the decrease in phishing attack reports in 2010 and 2011 likely to that of in 2009, due to the removal of the Avalanche gang. According to the 2nd of 2010 report by APWG, this was accountable for 66.6% of phishing campaigns in the 2nd of 2009. There were 35% lower phishing campaign reports submitted to APWG (i.e., 26,402) in the 1st half of the year 2011. However, a decrease in phishing attacks was because of the shift in the Avalanche gang's actions from conventional phishing campaigns to malware phishing campaigns.

Figure 1: Total Number of Submitted Unique Phishing Reports in the Second Half of 2011 to APWG

Trojan horses software is one of the famous models of malware, i.e., used in phishing attacks. According to APWG, 72% of the entire malware recognized in 1st half of 2011 was caused by Trojans software compared to the previous statistics of 55% in 2nd half of 2010. Even though phishing campaign reports decreased since 2009, the number of phishing campaign reports is still at the peak in contrast to that of 2nd half of 2008, which had a mean of 28,916 distinctive which vary between 22,000 and 26,000 special reports month to month in 1st half of 2011.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

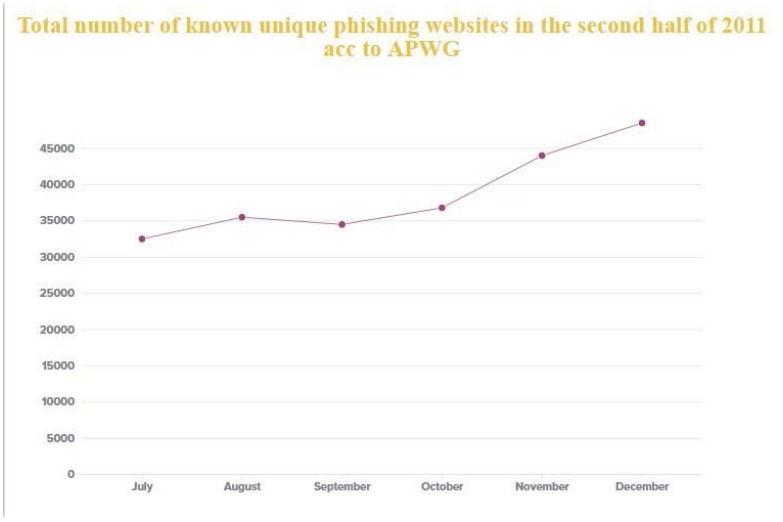

Figure 2: Total Number of Known Unique Phishing Websites in the Second Half of 2011 acc to APWG

On the other side of the coin, there was an increase in phishing campaigns and phishing websites in 2nd half of 2011, which looked to correspond with the holiday season as illustrated in Figure 1 and Figure 2. This can be expanded by intending that each phishing campaign can be done through electronic communication channels and sent to many users. In 2011, there was a phishing attack averse to RSA and HB Gary, which were popular security agency and caused further hacks in opposition to their clients like RSA's client Lockheed Martin which reveal the risk of phishing campaign or security weakness caused by human factor is not restricted to the immaturity of end-users, seeing that engineers can be the victims.

Cutting down the effect of phishing campaign is exceptionally important and enhance the comprehensive security of a firm.

CHALLENGES

Phishing campaigns grab the benefit of human inexperience and ignorance concerning their connection with electronic communication channels. Hence it is not an easy issue to solve. Here some of the suggested solutions that aim to reduce the effect of a phishing campaign. From a significant point of view, there are two proposed solutions to reduce phishing attacks in general.

1. User education: The user is trained to identify the phishing messages and reports such attacks to the system owner.
2. Software enhancement: A program is developed which helps the user to classify phishing messages more accurately.

The difficulties faced with both above mention methods are.

1. Non-technical individuals oppose learning, and if they acquire knowledge, which is not holding back their knowledge forever, they must be trained continuously.
2. Some of the software quick fix like security warning and confirmation/Authentication are still reliant on user behaviour if the individual neglect safety warning or authentication the finding can be made useless.

Phishing is a semantic attack that conveys the content in natural languages (e.g., English French) through electronic communication channels to convince the victim or user to carry out certain activities. Computers have a disadvantage in precisely grasping natural languages' semantics, which is one of the limitations. Email-Based Intrusion Detection System (BIDS) is a notable attempt that utilizes the NLP method to identify phishing attacks, yet its functional assessment displayed a rate of phishing detection is 75%.

IV. MITIGATION OF PHISHING ATTACKS: AN OVERVIEW

The anti-phishing solutions are mainly differentiated by knowing the life cycle of the phishing-attacks. Picturing these attacks' life cycle is tremendously crucial due to the wide range of types of phishing attacks. Gleaned from the literature survey, we have outlined a flowchart that efficiently gives us insight into the phishing campaigns' actual working, keeping the frame of reference as anti-phishing techniques. The flowchart with depicted as shown in the figure 3.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

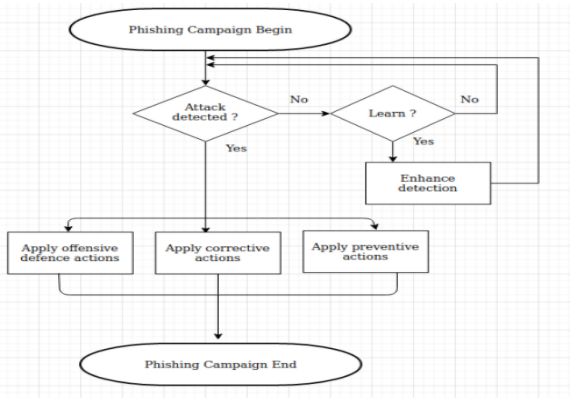

Figure 3: Life-cycle of Phishing attack

As we see in the figure 3, the attack starts from the very first phase-Phishing campaigns begin. Once the campaigns start, the first layer of protection is to check and validate the attack. If the attack is encountered, different detection techniques are used, such as offensive defense actions, corrective actions, and preventive actions to overcome the attacks. In further sections, we will know in detail about each of these detection methods. Thus the phishing-campaigns end.

1. Offensive defense actions: This particular technique helps abate the consequences of the phishing attacks, especially when the victim has shared personal details with the attacker.
2. Corrective actions: This particular method will help curtail the attacker's activities by suspending the hosting accounts or permanently deleting the phishing files.
3. Preventive actions: As there are a lot of phishing types, it is of no shock that there are numerous preventive actions. In this survey paper, most of the preventive measures are covered

Nevertheless, we should also take care of the detection phase because if the phishing-campaign is not detected correctly in the initial stage, then the whole concept of the life cycle of the phishing attack is of no importance. Thus primary priority must be given to the detection phase.

Detection Approaches

We have considered the ant phishing solutions as the detection solutions for identifying and classifying the phishing-attacks

1. User Training approaches: As most phishing-attacks involve manipulating the victim's mind and getting their details, it is always vital to educate the end-user, which will help them differentiate correctly between phishing and non-phishing messages. Even though user awareness is usually treated as a preventative approach for the phishing attack, identifying and detecting the phishing attack by the user is fundamentally essential. We categorize it as a detection technique.
2. Software Classification approaches: Usually, human errors and ignorance can cost a considerable loss. Different mitigation approaches aim to classify the legitimate and the fraudulent messages to overcome this issue. This acts as a nexus between the carelessness of the user and the protection of sensitive details.

As the user training can happen either by their online experiences or by external programs, it is not highly efficient to prevent phishing attacks. Whereas in the software classification approach, the classifiers are trained by machine learning methods to enhance detection techniques' efficiency.

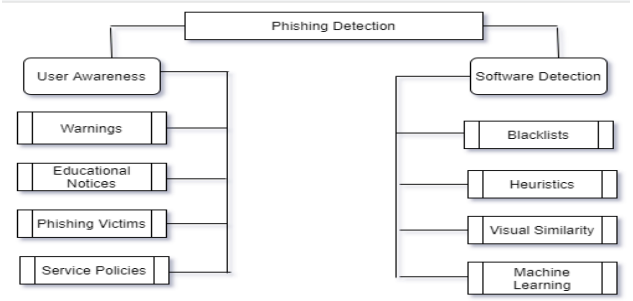

Figure 4: Different Phishing Detection Techniques

As seen in the above figure, the starting point of mitigating a particular phishing attack is detecting it. If the phishing-campaign is not detected, then the mitigation approaches are of no use. This survey paper is mainly focused on the detection stage.

Offensive Defense Approaches

As the defensive method will mostly give a positive result, this approach is highly recommended. This is usually carried away by flooding the attackers' website with plenty of fake credentials, which will make it the attacker very difficult to find legitimate details.

Even though this approach seems to work theoretically, it will usually be challenging to carry out this technique. It includes uploading plenty of fake data, which is time-consuming. Also, wise attackers might create some algorithms and loops to overcome the burden of evaluating the plethora of data.

Correction Approaches

Correction approaches begins once the phishing attack is discovered. Here modification is an action taken to bring down the phishing resources. This is usually accomplished by reporting phishing campaigns to Service Provider.

Resources on which phishing campaign usually depend on,

1. *Websites:* These websites can be a legal website synced with phishing content shared hosting website possessed by the phisher or a botnet that consists of several affected end-user workstations.
2. *E-mail messages:* E-mail messages can be sent through various platforms such as open SMTP relays, free E-mail Service Provider (ESP) (e.g., Hotmail, Gmail, etc.) or affected end-user computers, which belongs to part of the botnet.
3. *Social Networking services:* Phishing messages can be sent through web 2.0 services like Facebook and Twitter to influence users to disclose their passports.
4. *Public Switched Telephone Network (PSTN) and Voice over IP(VoIP):* Here, attackers prevail on dupe to carry out specific actions like another form of a phishing campaign. Although the difference is attacker tries to utilize spoken dialogues to gather the data. Besides, through the way VoIP protocols functions and VoIP provider system are designed, spoofing Caller ID's are being used by attackers used as an aid to raise their persuasion

In the aim of correcting such practices, service providers strive to draw the resources down.

For Example:

1. Deletion of phishing content from websites or adjournment of posting services.

2. Suspension of e-mail accounts, SMTP relays, VoIP services

3. Trace back and shutdown of botnets

This also leads to the closure of many enterprises that regularly aid phishing attackers. The organizations like banking and financial firms that give brand security services to their user can initiate the closure of enterprises. When the phishing attack is detected, they can inform or report to the website owner for instant closure. The punishment and procedures depend on in which country the phishers and phishing campaigns are identified.

Prevention Approaches

The prevention of phishing attacks can have a different meaning depending on its contexts:

1. Prevention of users from falling victim: According to this, even phishing detection methods are also regarded as a prevention technique.
2. Prevention of attackers from starting phishing campaigns: According to this, litigation and penalties against phishers by LEAs are considered one prevention approach.

Generally, LEA takes many weeks to finish their inquiry and reply procedures. Hence

Apply prevention approach after all other reduction techniques due to costly behaviour of LEA inquiries that make them take comparatively more time.

Once the phishing campaign is identified, LEA can then file litigation, which may issue penalties such as captivity, liability, and forfeiture of equipment used to make the attacks.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

V. EVALUATION METRICS

In this survey paper, we will be using evaluation metrics which is used in in phishing literature to compare and contrast between a numbers of detection technique.

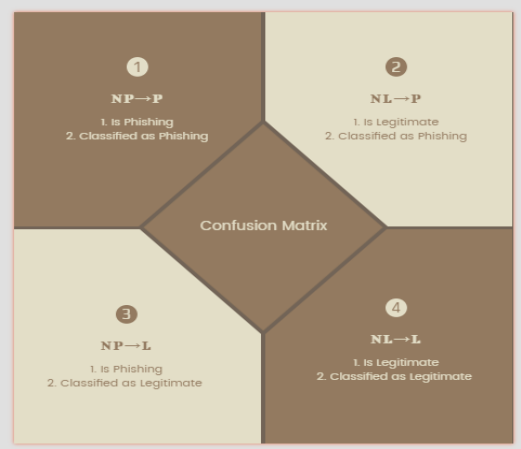

Figure 5: Conclusion Matrix

Here, $N_P \rightarrow {}_P$: Number of phishing instances correctly classified as phishing.

$N_L \rightarrow {}_P$: Number of non-phishing instances incorrectly classified as phishing.

$N_P \rightarrow {}_L$: Number of phishing instances incorrectly classified as legitimate.

$N_L \rightarrow {}_L$: Number of non-phishing instances correctly classified as legitimate.

Table 1. Different Evaluation Units along with their Mathematical Formulae

TP	$\frac{N_{P\rightarrow P}}{N_{P\rightarrow P}+N_{P\rightarrow L}}$
FP	$\frac{N_{L\rightarrow P}}{N_{L\rightarrow L}+N_{L\rightarrow P}}$
TN	$\frac{N_{L\rightarrow L}}{N_{L\rightarrow L}+N_{L\rightarrow P}}$
FN	$\frac{N_{P\rightarrow L}}{N_{P\rightarrow P}+N_{P\rightarrow L}}$
P	$\frac{N_{P\rightarrow P}}{N_{L\rightarrow P}+N_{P\rightarrow P}}$
R	TP
f_1	$\frac{2PR}{P+R}$
ACC	$\frac{N_{L\rightarrow L}+N_{P\rightarrow P}}{N_{L\rightarrow L}+N_{L\rightarrow P}+N_{P\rightarrow L}+N_{P\rightarrow P}}$
W_{err}	$-1\frac{\lambda\cdot N_{L\rightarrow L}+N_{P\rightarrow P}}{\lambda\cdot N_{L\rightarrow L}+\lambda\cdot N_{L\rightarrow P}+N_{P\rightarrow L}+N_{P\rightarrow P}}$

Here, TP: True Positive

FP: False Positive

TN: True Negative

FN: False Negative

P: Precision

R: Recall

ACC: Accuracy

W_{err}: Weighted error

VI. DETECTION OF PHISHING ATTACKS: THE HUMAN FACTOR

Considering phishing attacks that try to benefit unscaled users, an obvious solution is educating the end-users, which would decrease their vulnerability to falling dupe/victims of phishing attacks. Just training the end-users alone does not control their practices. This segment will discuss some of the deed put up in preparing the end-users concerning phishing attacks.

Phishing Victims

A survey was done by Julie S Downs et al. to analyse what are the benchmark that can forecast the vulnerability of an end-user to fall as a victim of a phishing email[15]. The result of the study was that the end-users who had a good understanding of "phishing definition were unlikely to fall for phishing emails, while the insight about cookies, spyware, and viruses didn't help to cut the vulnerability to phishing emails. The study concluded that educating end-users about phishing attacks should be focused more on warning end-users about the negative outcome.

One more study that was done by Hua Jun Huang et al. concluded that the primary reasons that lead technology users to fall as victims for phishing attacks are:

1. Users ignore passive warnings (e.g., toolbar indicators).

2. Many end users cannot determine the websites as phishing or genuine sites.

Another study done by Steve Shen et al. revealed that many indirect attributes connect victims and their vulnerability to phishing attacks. As per the study, gender and age intensely connect with phishing vulnerability. They conclude that.

1. Females tend to click on email links more often than males.
2. Users between 18 and 25 years of age more apparently to fall victims to a phishing attack.

Service Policies

Different models have been developed to notify the end-users, but this will be useful only when the end users have little understanding of phishing attacks.

Less knowledge about phishing attacks may expose end-users' personal credentials regardless of hardware and software protection layers s hence end-users should be educated about various types of phishing attacks like sending an email, periodic email, and SMS.

Firms should have strict rules against giving out confidential data over email, SMS, etc. The organizations must be coupled with various awareness programs to make sure that end-users know this policy. By doing so, users have higher chances of identifying variability within the phishing message, for example, asking about users' personal information.

Service providers should rigidly impose rules against the illegal use of their services. As y T. Moore and R. Clayton studied, many hosting providers take their services down if they are abused; hence the services takedown is highly common in the way security problems.

Users must be educated about the environment they use to ignore notifications such as X.509 certificate verification failure messages, which could lead users towards permanently trusting illegal certificates. Breaking the trust model of a Public Key Infrastructure (PKI), making Hypertext Transfer Protocol Secure (HTTPS) or Secure/Multipurpose Internet Mail Extensions (S/MIME) such behaviour is cost by Man in the Middle (MITM) attack which is rusted certificate.

User education policies and applied practices should be cleared with their objective. Otherwise, the educational policies and procedures could converse to the actual message. For example, the approach that appraises HTTP sites' risk with an invalid X.509 certificate publishes local sites with self-signed credentials, resulting in an indirect message that the security policies, practices, and notification are not of great significance. This leads to the way of opposing the real intentions of security policies.

An additional challenge is using computers in various environments other than the organizations where appropriate IT policies do not exit. Frequent ill-use of technology can lead end-users who ignore security warnings or notifications as the norm that might lead gradually taken into their workstations as well.

Due to the generic nature of the challenge, a highly predictable solution remains challenging to achieve. Even the phishing attack problem's psychological outlook is comprehensive and has a vital role in end-user behaviour.

Passive and Active Warnings

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

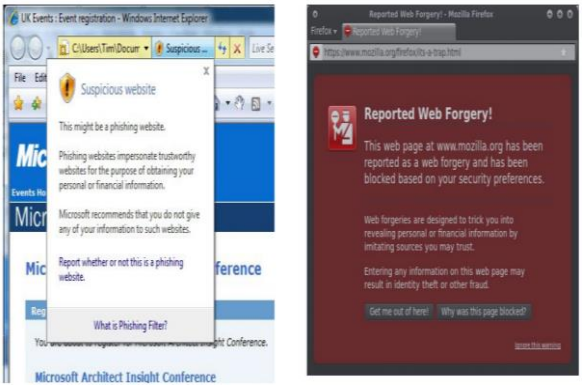

Figure 6a: Passive Warning **Figure 6b**: Active Warning

This detection approach is carried by notifying the user on their interface by giving warning messages whenever they start using or accidentally opening the phishing website marked as fraudulent webpages by the web browsers. The warnings can be shown in two different ways.

• Passive Warnings: Here, The user can view the fraudulent web page's content even after the warning is given. This will not block the content area.

• Active Warnings: Here, The user cannot view the fraudulent web page's content once the warning is given. This will block the content area.

Educational Notices

This detection technique is often done in the regular interval by sending awareness messages which warn the end-users regarding phishing threats. This is usually done by sending periodic SMS and Email. This is an ineffective approach as per the study conducted by Kumaraguru et al. [11], which says that sending messages will only improve the client's theoretical knowledge. In contrast, in a practical situation, the implementation of the same expertise is hard to find.

Kumaraguru et al. [11] have given an alternate approach to tackle this inefficiency. This is to use an embedded training system, which is far better than the periodic notices. This has been proved by the experimental statistics, which is – I. 30% of the comics embedded training participants were the victims of phishing attacks. II. 89% of the periodic notices training participants were the victims of the phishing attack.

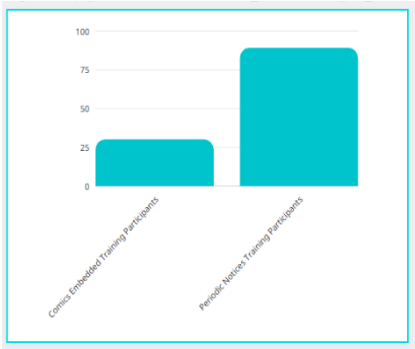

Figure 7: Statistical Representation of the different training sets used for the Educational Notices Detection Technique

The above statistics depict that the embedded training systems are more efficient than the regular educational notices.

This proposed system work as follows:

A set of dummy phishing emails is sent to the user. Whenever the user interacts with that email (by clicking on the link), he/she is directed to a page that gives them information about the phishing attack and its consequences. This helps the end-user to be more cautious while dealing with the real phishing mail.

VII. PHISHING DETECTION BY BLACKLISTS

PhishNet – Predictive Blacklisting

The fishnet [17] will overcome the limitations of any changes in the phishing URL, which would result in no match. To solve this issue, the other method will process the URL's that are present in the blacklist and make different variations of those URL's and store them in the Database for resolving the limitations. Five different variation approaches do this are:

- Replace Top Level Domain (TLD).
- Directory Structure Similarity.
- IP Address Equivalence.
- Query String Substitution.
- Brand Name Equivalence.

Automated Individual Whitelist

The Automated Individual Whitelist (AIWL) [21] is opposite to the concept of the blacklist. When the blacklist has the list of fraudulent URL's, the whitelist will be having the list of trusted Login User Interfaces (LUI's). So, whenever the user login to a page, a warning is given if that page is not in the whitelist. The two primary components of AIWL are:

- Whitelist- A list of trusted LUI's.
- Automated Whitelist Maintainer.

VIII. PHISHING DETECTION BY HEURISTICS

The software which scrutinizes payloads of various protocols via different algorithm can be installed on client or server-side. The multiple protocols could be HTTP, SMTP, or any arbitrary protocol algorithms could be any methods that can detect or prevent phishing attacks. Phishing heuristics are observed to exist in phishing attacks. However, the characteristics are not guaranteed to always live in such an attack. A set of heuristics test attacks that were not seen previously were detected (zero hour phishing attacks), an advantage over the backlist. Prime web browser and mail clients are built with a heuristic phishing protection mechanism that detects phishing campaigns. The clients consist of Mozilla Firefox, Internet Explorer, Mozilla Thunderbird, and MS Outlook.

Spoof Guard

It is a web browser plug-in developed by Stanford University. It detects HTTP(S)-based phishing attempts as a web browser toolbar by weighting certain irregularity found in the HTML content against a defined threshold value.

Listing 4 shows samples of HTML heuristically detectable by Spoof Guard

Click Here

Click Here

www.gmail.com

<input type=" password" />

CANTINA: A Content-Based Approach

To determine whether phishing page we have a toolbar in an internet explorer and this toolbar is called the cantina. To decrease FP cantina, heuristics methods, search engines, and Term Frequency-Inverse Document Frequency (TF-IDF). Cantina uses the below procedures to identify phishing sites.

1. TF-IDF of every word of a fishy website is computed.

2. To demonstrate the document, we considered Top 5 words with a high TF-IDF rate.

3. Using google search query, a search engine where we report the five results we obtained in the previous step names q=t1, t2, t3, t4, t5and it returns which yield first n entries which are stored in domain.

4. The website is genuine or lawful if the doubted domain name is obtained in return entries

TF and IDF are formulated as below,

1) In the document j, n Iij is the frequency of i, n $m_{,j}$ is the frequency of m, and k is total count.
2) Here |D| is the diversity of documents in each corpus, |Di| is the diversity of forms with term i in the same corpus. The corpus used to calculate IDF. Therefore, TF-IDF for every word i is computed as below.
3) Each page state is computed $S = f(w_i \cdot h_i)$

4) Where h_i is every heuristic's output, wi is the weight of every heuristic, and threshold function f returns -1 and 1 to denotes phishing sites and legitimate.

Phishing Sites Blacklist Generator

Many mechanisms are foot forward to generate blacklists using search engines dynamically. Like Google. The method identifies phishing websites and then stores them in DB.

The estimation of the dataset consists of 500 randomly selected websites. When loaded 13 random search keyword and 30 phishing sites which are gathered using PIRT.

The suggested heuristics are given below:

1. Using suspicious URLs, we obtain the name of the company.
2. To fetch the first ten results, the name of the company is searched on google.
3. The page is genuine, and then the suspicious URL is returned from the first ten returned results.
4. If the questionable URL is detected as phishing, then the dubious URL does not return from the early ten returned products.
5. Then phishing URL is stored in DB.

When the above test was conducted, phishing websites were accurately classified. The blacklist generator is placed in a mail server that analyses all the URLs present in messages. Researchers noticed that the heuristics method introduced additional delay on the internet browsing experience. This allows the mail server to process URLs before the user requests them, and if the result from the blacklist generator is cached, delay on the user side can be noticeably reduced.

IX. PHISHING DETECTION BY VISUAL SIMILARITY

This detection technique works on the Visual appearance by opposing or any network-level information.

Classification with Discriminative Key point Features

This proposed solution [37] concentrates on the screen's content rather than the underlying content code. This detection technique follows specific steps:

1. Take a snapshot of the suspected site.
2. Match the snapshot with the legitimate sites that are present in the whitelist.
3. Convert the RGB Channel into GrayScale channel.
4. This resulted in GrayScale Channel is then used to analyse the critical feature.

Visual Similarity-based detection without victim's site information

According to the work presented in [39], this detection technique will not consider any victim's site information from the whitelist. This detection technique has visual similarity between the fake and the legitimate ones as their crucial aspect.

This technique usually happens in the following way:

1. A list of know fake (WP) and legitimate websites (WL) will already be in a Database, which will help as a baseline for the classifier.
2. **2**. Whenever the suspected website's screenshot is evaluated, the screenshot is compared with the Database to know the visual similarity.
3. Based on the comparison, the above algorithm is executed
4. A domain name whitelist is used for more efficiency.

X. PHISHING DETECTION BY HEURISTICS PHISHING DETECTION BY DATA MINING

In this part, the methods explain Detect phishing attacks as file segregation or an issue with assembly, where systems are built by receiving the advantage of machine learning and clustering algorithms, like, Density-Based Spatial Assembly of Applications with Noise (DBSCAN),k-means, and k-nearest neighbours. Let us assume that KNN stocks program examples in drive characterized as multidimensional parameters. In this, every section signifies the value vector obtained from a specific feature. The classification task is then achieved by the likewise Process test instances, compute the distance between the test instance and the other Training situations. When K = 3, the neighbours' ranks are considered three nearest (as obtained during the training phase), when the task is classified, a majority vote can be used to verify the test instance category. Algorithms

like C4.5 and SVM obey different methods in that they simplify the classification model (as opposed to the K-NN, which does not generalize the model). Let us assume C4.5 constructs a decision tree that should be broad enough to classify invisible instances appropriately. The decision tree contains nodes with divided branches. The division is commonly achieved to maximize temporary information after the split. Other than that, SVM aims to search for an adequate division level in the vector space by testing training cases. It should be the level of division be general enough so that they should still separate Invisible issues. Though, aggregation algorithms like k-means and DBSCAN are classless partition examples (i.e., Knowledge of the class label is not required to create Groups.) K-means algorithm designed to build k partitions by randomly placing the initial K-partition centres, followed by repeated instances set for a smaller section distance (e.g., the minority distance) near its centre Apprise the centre of the section to be the average instances in the same section. This repetitive process is replicated until the blocks converge. Other than that, DBSCAN can split data built on density (i.e., using a space measuring the function, such as the minority distance) from instances. Conflicting to k-means, DBSCAN does not essential to know before the number of partitions that should be obtained, which is accomplished by the idea of density reachability.

Detecting DNS-poisoning-based phishing attacks from their network performance characteristics

Kim, H. et al. intend an instrument for noticing DNS poisoning which examines network-level belongings for user transmission. The gathered data consists of 10,000 data items, 50% of which were about phishing sites, and 50% were directed to Genuine Websites.

The rating data set consisted of routing data representing 10,000 web sites, with 1:1 Phishing to genuine website ratio. The data set was created by gathering the routing data for 5,000 instances for 50 significantly aimed websites (10 cases per website) and 5,000 Phishing instances (100 points per web site). By that way of an initial decomposition, only 19% of Phishing sites were behindhand firewalls, whereas 79% of the genuine sites were behindhand firewalls. The authors aligned. That's the fact that phishing sites are hosted in fewer secure, shared web hosts than their real counterparts, which allows attackers to host phishing websites illegitimately.

The authors then handled routing data obtained by several learning algorithms, like SVM and kNN. Their findings presented that the best execution class K-NN was by k = 1. In the upcoming work, the authors propose inserting more Features, like RTT, between the end-user networks and, respectively, a 3-hop layer is on track to the target service.

Large-Scale Automatic Classification of Pages

The author explains Google's anti-phishing solution to quickly and quickly categorize pages Keeping FP low. Once the ranking is done, the output is grouped like a Posted backlist via Google Safe Browsing API (which Apple Safari, Firefox, and Google Chrome) uses that results from the glitches of delay in human-driven phishing. Rated like what Phish Tank does, Phish Tank takes 50 hours (as Broker) to verify the URL. The approach followed by Google mainly decreases user participation to issue blacklists urgently.

The fourth foo classifier functions as given below.

1. Users of Gmail are not mechanically organized unsolicited mails as spam.
2. URLs which fall in unwanted mails as applicants merely if he presented himself by the more users (to reduce abuse).
3. Filtered URLs are handled to extract some features.
4. A filtered URL is directed to extract properties of the page.
5. Classified (one of the learners of the hoped-for advertising machine the workbook) and then uses the output from the extracted Features. The result is measured from 1.0 to 0.0 (0.0 being very trolling, while 1.0 is being very legitimate). Practically, as Google tested it, the outputs were typically in the two parties are extreme, making the classification clearer. The same workbook is also running in contradiction of the pages he visited Google to crawl to notice more phishing attacks.
6. The above-class science classifies as it classifies more documents.

Natural-Language Processing for Intrusion Detection

Allen Stone intends the use of NLP for imposition detection systems[10], which also identify phishing attacks. IDS can detect style matches, but its strangeness finds out Messages depends on their connotations. It is called a system. BIDS, an overview of its process as below,

1. Message bodies are obtained and give back to identifying technique.
2. The identifying approach trusts on OntoSem to classify the message — the section that publishes NLP models.
3. IDS names are built on craft string matches in contradiction of ontoSem results.
4. Depends on the matches, an e-mail is categorized as other phishing or genuine.

BIDS contains four sets of naming rules,

- **Account compromise:** Base expects to reveal Phishing attacks were designed as e-mails that semantically claimed that the victim's account had been hacked.
- **Account change:** A rule which is anticipated to reveal phishing attacks that have been designed as semantic e-mails Demand that the victim's account be amended**.**
- **Financial opportunity:** Base expects to reveal Phishing attacks were designed as e-mails semantically claiming economic opportunity**,** like cash Gains.
- **Opportunity:** Base expected to reveal deception. The attacks formed as e-mails claiming semantic possibilities that do not accurately signify monetary gains, like free vacations**.**

XI. EVALUATION OF HUMAN TRAINING APPROACHES

Steve Sheng et al. [1] have conducted enough study regarding the various educational material evaluated to check their effectiveness. They are as follows:

• *Popular Training Materials*- There are three popular training materials.

1. OnGuardOnline phishing tips [9]

2. Microsoft Online Safety [8]

3. National Consumer League Fraud tips [5]

• *Antiphishing Phil*- Carnegie Mellon University has developed an interactive educational game that can be played on a web browser.

• PhishGuru Cartoon- This technique is usually helpful to teach the users about phishing by using the training system embedded into the users' mail system.

• *Antiphishing Phil with PhishGuru Cartoon*- As the name itself says, it is a combination of the above two techniques.

These evaluations are done by experimenting on different categories of people, summing up to 1001 participants in total. The other types are as below.

1. Control- They are not given with any educational materials.

2. Popular Training- This set of people is given the only materials described above.

3. Antiphish Phil– these people are given the Anti-phish Phil training material.

4. PhishGuru- these people are given the PhishGuru training mater.

5. Combined- The participants here are given both Anti-phish Phil and PhishGuru training material.

XII. LEARNED LESSONS

This section will present our views and the lessons learned from the survey that we have carried out.

User Education and awareness

It is always better to train the victims/ end users to be aware of fraudulent phishing. The quote "Prevention is better than cure" is still better to have awareness even before getting trapped in the attacker's nets. This is possible by using better user interfaces with passive and active warnings, by enhancing the behaviour of the systems by automatically deleting the harmful and the phishing email by using specific software on behalf of the end-user.

Blacklists

The blacklists are constructed to achieve efficiency, which is possible only when there are minimal FP rates. However, as every system has both pros and cons, even blacklists have their disadvantages. The main factor is.

• High retrieval time- As the blacklists contain a massive amount of data (URL or IP Address), it takes a longtime to search for the required data.

Heuristic tests and visual similarity

However, methods identify attacks which include zero-hour It also has high FP values which reduce user efficiency methods which depend on visual similarity have same property They need an image of how the website looks and stores the image which they gather some key points which in turn increases of method that depend on visual similarity.

Machine Learning-based classifiers

ML-based methods are all over the blacklist as it can reduce zero-hour attacks and build this own model.ML methods have an advantage when compared to heuristics tests.

- Can build efficient models without mechanically examining data.
- When phishing attack increases, ML methods builds new models through reinforcement learning.

XIII. CONCLUSION

Looking into the phishing attack and its detection technique in detain it is mostly evident that the motive of the phishing attack is to have personal benefits by tricking the innocent people to give their personal information and sensitive data into the hands of an attacker. The most efficient and the fundamental way to tackle this fraudulent attack is by giving awareness and educational sessions to the end users. This can be done effectively in various methods as mentioned in this paper. If we don't act immediately to this problem, then it might be very costly to recover after the attack.so let us all work together to save our data and thus our privacy.

XIV. REFERENCES

[1] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, "Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions," in Proceedings of the 28th international conference on Human factors in computing systems, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 373–382.

[2] B. Krebs, "HBGary Federal hacked by Anonymous," http://krebsonsecurity.com/2011/02/hbgary-federal-hacked-by-anonymous/, 2011, accessed December 2011.

[3] B. Schneier, "Lockheed Martin hack linked to RSA's SecurID breach," http://www.schneier.com/blog/archives/2011/05/lockheed martin.html, 2011, accessed December 2011.

[4] C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages," in NDSS '10, 2010.

[5] X. Dong, J. Clark, and J. Jacob, "Modelling user-phishing interaction," in Human System Interactions, 2008 Conference on, may 2008, pp. 627–632.

[6] W. D. Yu, S. Nargundkar, and N. Tiruthani, "A phishing vulnerability analysis of web based systems," in Proceedings of the 13th IEEE Symposium on Computers and Communications (ISCC 2008). Marrakech, Morocco: IEEE, July 2008, pp. 326–331.

[7] Anti-Phishing Working Group (APWG), "Phishing activity trends report — second half 2010," http://apwg.org/reports/apwg report h2 2010.pdf, 2010, accessed December 2011. 28http://www.buhooth.ae/

[8] Anti-Phishing Working Group (APWG), "Phishing activity trends report — first half 2011," http://apwg.org/reports/apwg trends report h1 2011.pdf, 2011, accessed December 2011.

[9] Anti-Phishing Working Group (APWG), "Phishing activity trends report — second half 2011," http://apwg.org/reports/apwg trends report h2 2011.pdf, 2011, accessed July 2012.

[10] B. Schneier, "Details of the RSA hack," http://www.schneier.com/blog/archives/2011/08/details of the.html, 2011, accessed December 2011.

[11] P. Kumaraguru, Y. Rhee, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge, "Protecting people from phishing: the design and evaluation of an embedded training email system," in Proceedings of the

SIGCHI conference on Human factors in computing systems, ser. CHI '07. New York, NY, USA: ACM, 2007, pp. 905–914.

[12] A. Alnajim and M. Munro, "An anti-phishing approach that uses training intervention for phishing websites detection," in Proceedings of the 2009 Sixth International Conference on Information Technology: New Generations. Washington, DC, USA: IEEE Computer Society,2009, pp. 405–410.

[13] S. Gorling, "The Myth of User Education," Proceedings of the 16th Virus Bulletin International Conference, 2006.

[14] G. Gaffney, "The myth of the stupid user,"

[15] http://www.infodesign.com.au/articles/themythofthestupiduser, accessed March 2011.

[16] A. Stone, "Natural-language processing for intrusion detection," Computer, vol. 40, no. 12, pp. 103 –105, dec. 2007.

[17] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit, ser. eCrime '07. New York, NY, USA: ACM, 2007, pp. 60– 69.

[18] C. Yue and H. Wang, "Anti-phishing in offense and defense," in Computer Security Applications Conference, 2008. ACSAC 2008. Annual, 8-12 2008, pp. 345 –354.

[19] [18] P. Knickerbocker, D. Yu, and J. Li, "Humboldt: A distributed phishing disruption system," in eCrime Researchers Summit, 2009, pp. 1–12.

[20] L. James, Phishing Exposed. Syngress Publishing, 2005.

[21] J. S. Downs, M. Holbrook, and L. F. Cranor, "Behavioral response to phishing risk," in Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit, ser. eCrime '07. New York, NY, USA: ACM, 2007, pp. 37–44.

[22] H. Huang, J. Tan, and L. Liu, "Countermeasure techniques for deceptive phishing attack," in International Conference on New Trends in Information and Service Science, 2009. NISS '09, 2009, pp. 636 – 641.

[23] T. Moore and R. Clayton, "Examining the impact of website take-down on phishing," in eCrime '07: Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit. New York, NY, USA: