ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 12, December 2013

A Comparative Study on Various Mechanisms to Detect Phishing Websites

Radhe Shyam Panda, Rajesh Tiwari

***Abstract*— Phishing attacks have been a significant threat to the Internet users and business organizations across the globe causing billions of dollars loss. Phishing is a type of internet fraud that uses emails or websites, which is designed to look like from legitimate organizations, to take in users into disclosing their personal or financial information. This information can be used by hostile party for scandalous and criminal purposes, such as identity theft and fraud. The inability of an end user to accurately sense legitimate and fraudulent emails or websites apart results into a successful phishing attack .The current paper focuses on a comparative study and analysis of various phishing detection mechanisms. This survey is a structured guide to support the current status of the literature that is based on the anti-phishing mechanism.**

***Index Terms*— Blacklisting, MIME, Phishers, Phishing, SSL, UBPD, Whitelisting.**

I. INTRODUCTION

Phishing attacks are the biggest threat in today's web world. A lot of these attacks tempt people into visiting fraudulent websites that pretend to be trusted entities, and convince them to disclose their sensitive information like passwords; personal identification numbers (PINs), etc. Despite of having many advances in anti-Phishing solutions in recent years, phishing still causes tremendous losses every year. The exact definition of phish varies from paper to paper. We define phish to be a webpage satisfying the following criteria:

- **A well-known website is impersonated by replicating the whole or part of the target site, showing high visual similarity to its target.**
- **The sensitive information such as a password is demanded by using login form.**

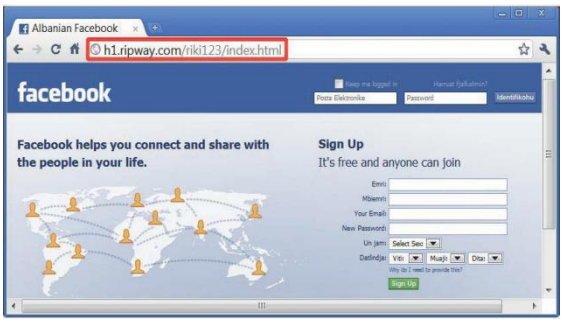

Figure 1: Phishing Website

Figure 1 shows an example of a phishing website posing as Facebook; notice how the URL in the red box (http://h1.ripway.com/riki123/index.html) is completely different to the original URL.

An analysis has been done by Dhamija et al. [9] on 200 phishing attacks from the Anti-Phishing Work Group database and it identified several reasons, ranging from pure lack of computer system awareness, to visual deception tricks used by opponents, due to which phishing attacks were successfully executed. They further performed a usability study with 22 participants and observed that 23% of the participants failed to look at security indicators against phishing attacks and as a result 40% of the time they were subjected to a phishing attack.

A common practice that is followed among all phishing sites is that they maliciously deceive and mislead users to trust that they are legitimate sites. Therefore, detecting a phishing page is basically an authentication problem between servers and humans. Ideally, when a website is visited by any user, he or she is able to authenticate the web server, even if the server is accessed for the first time. No existing technical mechanism fully solves this problem. For example, SSL only verifies a web server's IP address or hostname to a browser and protects the communication channel as well. However, it provides no guarantee the HTML files sent by the web server are not misleading. We report our study in this paper.

A. Type of Phishing

Phishing can be broadly classified into two categories such as deceptive phishing and malware-based phishing. This classification is done on the basis of techniques that are employed for phishing. The first technique is related to social engineering schemes, which depend on fake email claims that originate from a legitimate organization. By using an embedded link within the email, the phishers attempt to redirect users to fake Web sites. These Web sites are designed to obtain confidential data from victims deceptively, including credit card numbers, usernames and passwords or personal information. The second technique uses technical deception schemes that rely on malicious software programs spread through deceptive emails or by detecting and using security holes in the user's computer to obtain the victim's online account information directly. The current study focuses on deceptive phishing using social engineering schemes. Figure 2 explains the place of phishing email in phishing attack techniques.

Several common tricks are used for deceptive phishing:

1) **Imitation:** The email and the linked website closely resemble official emails and the official websites of the target. This includes the use of genuine design elements, images and trademarks logos.

***Radhe Shyam Panda**, Computer Science and Engineering, CSVTU, Bhilai, India, 09981588740*
***Rajesh Tiwari**, Computer science and Engineering, CSVTU, Bhilai, India, 09407780026.*

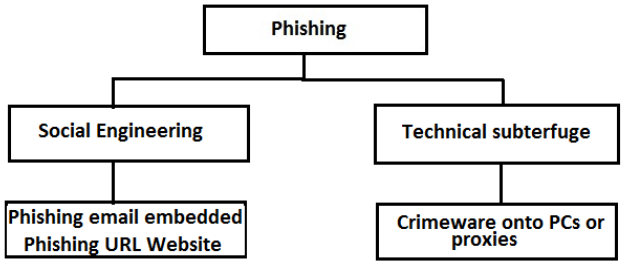

Figure 2: Types of Phishing Attacks

2) **URL hiding:** Phishers attempt to make the URLs in the email and the linked website to appear official and legitimate and hide the actual link addresses.
3) **Social engineering:** The invention of plausible stories, scenarios and methodologies to produce a convincing context and in addition the use of personalized information.
4) **Image content:** Phishers create images that contain the text of the message only in graphical form.
5) **Email spoofing:** Phishers hide the actual sender's identity and show a faked sender address to the user.
6) **Hidden content:** Phishers insert information into the phishing mail or website, which is invisible to the user and aimed at fooling automatic filtering approaches

II. LITARATURE REVIEW

According to one survey by the Gartner Group [McCall 2007], phishing attacks in the United States caused $3.2 billion loss in 2007, with about 3.6 million victims falling for the attacks, a huge increase from the 2.3 million the year before. Moore et al [Moore and Clayton 2007] reported that the loss to consumers and businesses in 2007 in the US alone was around $2 billion. A significant proportion of those losses were caused by one particularly infamous group, known as the "rock phish gang" that uses toolkits to create a large number of unique phishing URLs [1]. There are some existing countermeasures against phishing that have been proposed-

1) Network and Encryption-Based Measures.
2) Blacklisting and Whitelisting.
3) Content-Based Filtering for Emails and Websites.
4) Phishing Page Detection based on Anomaly.
5) User Behavior Based Phishing Websites Detection(UBPD)

III. COMPARATIVE STUDY

In this section we focus on the comparative study of various anti-phishing mechanisms that have already been described in the literature review section.

A. *Network and Encryption-Based Measures*

Communication-oriented measures aim at establishing a secure communication. A first protection against malware-based phishing attacks is the installation of virus scanners and regular software updates two factor authentications, where two independent credentials are used for authentication, e.g., smartcards and passwords. Recently transaction authentication numbers (TANs) are being sent to the user on demand by some banks during transaction through Short Message Service (SMS) using the mobile phone network. This reduces the phishing risk to a large extent as the TAN encodes the amount to be transferred and the receiving account number and hence is not affected by man-in-the-middle attacks.

B. *Blacklisting and Whitelisting*

In Blacklisting [2] a database for the phishing web site is maintained and the URL of the phishing website is blacklisted. By this technique only the known -bad phishing websites can be effectively detected, it can be weak against detecting new ones. It is also difficult to update efficiently and verify the website entries of central databases.

Whitelists manage a list of known-good websites [3]. This technique follows the counter process of the Blacklisting technique. It maintains the database of a legitimate websites. To some extent it can be effective against new phishing websites since only those considered 'trusted' are accessed by users. However, it is somewhat difficult to know exactly which new legitimate websites a user will visit next and have these entries added to the white list prior to their visit.

Whitelisting or Blacklisting, both the techniques of phishing prevention concentrates on checking web addresses when a page is rendered in a web browser. For example, In the Mozilla Firefox browser, each web page requested by a user is checked against a blacklist of known phishing sites. This list is automatically downloaded to the local machine and is updated in regular intervals. It takes some time until a new phishing site is reported and added to the blacklist. Therefore the effectiveness of blacklisting is limited because new phishing sites appear frequently e.g., in December 2007 about 35 new phishing sites were detected per hour [4]. The average time for phishing sites to be online is only 3 days; many sites disappear within hours. An additional problem is caused by distributed phishing attacks [6] where the links in a phishing email point to a large number of different servers are hosted on a botnet. It is found that well-crafted spam and phishing messages regularly manage to pass the blacklist filters [7].

C. *Content-Based Filtering for Websites*

These examine the content of an email or the associated website and try to identify different tricks for producing a phishing attack, e.g.

1) The study of images which may contain the message text.
2) The detection of invisible content inserted to deceive automatic filtering approaches.
3) The detection of design elements, logos and trademarks known brands. The identification of fraudulent sender addresses and URLs [4].
4) The detection of typical formulations urging the user to enter confidential information.

One very common strategy that is frequently used for phishing is to entice the users to any website through emails that seems to be legitimate where they are urged to disclose confidential data. It is therefore expected to look at the email structure and external links, respectively. For the detection of phishing emails several sets of basic features have been proposed, some of them are as follows

1) **Structural Features**: The numbers of possible message formats are defined by MIME (Multipurpose Internet Mail Extensions) standard to analyze the structure of an email. The body part

structure of an email can be captured from the HTML tree. We observe three features: the total number of body parts, the number of distinct and compound body parts and the number of alternative body parts, which are different representations of the same content.

2) **Link Features:** Various properties of links contained in an email can be reflected by observing following features:
   a) The total number of links.
   b) The number of internal and external links,
   c) The number of links with IP-numbers,
   d) The number of deceiving links (links where the URL visible to the user is different from the URL the link is pointing to).
   e) The number of links behind an image,
   f) The maximum number of dots in a link
   g) A Boolean indicating whether there is a link

3) **Element Features:** The kind of web technology that is used in an email can be revealed by observing the element features of an email. In order to achieve this we record three Boolean features of whether HTML, scripting and in particular JavaScript, and forms are used.
4) **Spam Filter Features:** Some off-line version of Spam Assassin can be used to generate two features, the score and a Boolean of whether or not an email is considered as spam. A message is considered spam if its score is greater than 5.0. Black- or whitelisting can be embedded for the performance improvements in real world. Some other industry standard SPAM email detection techniques are used to identify SPAM emails by Correlating links in known SPAM email to phishing sites [8].

D. Phishing Page Detection based on Anomaly

Every website claims an identity in the cyberspace either explicitly or implicitly. When a phishing site maliciously claims a false identity, it always exhibits anomalous behaviors compared to an honest site, which are indicated by some web DOM objects in the page and HTTP transactions and a phishing attempt can be detected by capturing those anomalies. An anomaly based phishing detector comprises two components: an identity extractor and a web page classifier described below:

1) **Identity Extractor**: Identity extraction is to acquire an abstraction of the ownership of a web site. An identity of a web site is defined as a set of words (i.e. character strings) which uniquely identify the web site's ownership in the cyberspace. Typically, the identity is an abbreviation of the organization's full name and/or a unique string appearing in its domain name. The identity is indicated in a number of objects or properties in a web page, the identity extractor uses an algorithm by making use of keyword extraction techniques from information retrieval literature. Identity extractor determines the web identity by considering the following DOM objects [5]:
   a) **Title:** the title of one web pages namely, the text between the tag <title> and < \title>).
   b) **Description:** the content property of the META whose name or http-equiv is "description".
   c) **Copyright:** the content property of the META whose name or http–equiv is "copyright".
   d) **ALT/title:** the alt and title properties of the DOM objects such as IMG, AREA, INPUT, APPLET, OBJECT.
   e) **Address:** the text of address objects.
   f) **Body:** the text in the main body or the images in a web page. There are many technologies to recognize texts from one image, such as Optical Character Recognition (OCR)

2) **Page Classifier:** Page classifier employs Support Vector Machine, a well-known algorithm for classification. Since a web page is either faked or authentic, phishing detection is by nature a binary classification problem. We make use of SVM as the page classifier. Its input is a 10- dimension vector representing a web page's 10 structural features. It outputs a label 1 indicating a phishing page or a label -1 indicating an authentic one.

E. User Behavior Based Phishing Websites Detection

The user is warned when he/she is about to submit credential information to phishing websites (i.e. when other existing anti-phishing mechanism fail), and protects user as the last line of resistance. Its detection algorithm is independent from how phishing attacks are implemented, and it can easily detect sophisticated phishing websites that other mechanisms find tough to deal with.

In this mechanism the authentication credentials, which phishers try to extract, must be shared only between user and legitimate organizations. Such (authentication credential, legitimate website) pairs are treated as the user's binding relationships. In legitimate web authentication interactions, the authentication credentials are sent to the website they have been bound to. In a phishing attack the mismatches cause the user to unintentionally break binding relationships by sending credentials to a phishing website. Hence phishing websites can be detected when both of the following two conditions are met:

1) The current website has rarely or never been visited before by the user.
2) The data, which the user is about to submit, is bound to website other than the current one.

Figure 3 explains the work flow of this mechanism.

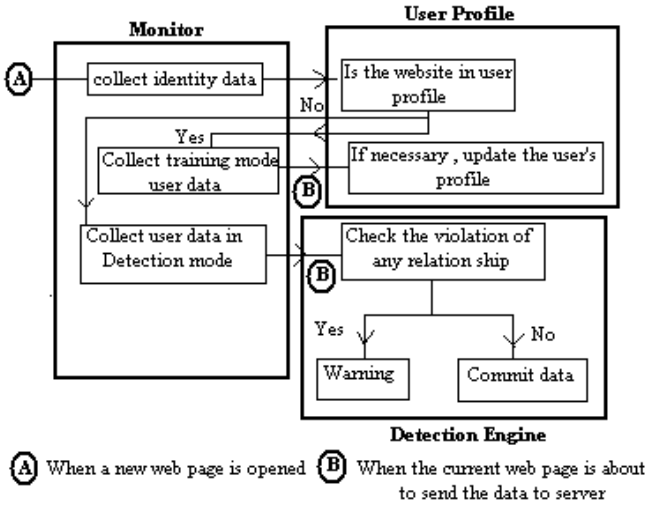

Figure 3: Detection Process Work Flow

In this section we have gone through various countermeasures that can be used to detect the phishing attempt. Each has its own advantages and disadvantages.

IV. Conclusions and Future Work

Network and Encryption-Based Measures are secure but use mobile TAN that requires a considerable infrastructure and are time consuming and incur costs. Black listing and White listing techniques are very easy to implement but requires considerable data storage and may produce false information against a newly created website. In case of anomaly based phishing detection technique, it remains as an open problem how to extract the identity from web pages with an overwhelming success probability. A further study on the phishing attacks and web structural features is desired to improve the effectiveness of the phishing detector. UBPD is able to consistently detect phishing WebPages regardless of how they are implemented as long as they ask for authentication credentials; however UBPD cannot handle all types of authentication credentials. It can handle static type authentication credentials such as user name, password, security questions, etc, but dynamic authentication credentials shared between users and the legitimate websites cannot be handled by the current implementation (e.g. one-time passwords). Although Content-Based filtering for Websites seems to be a bit complex but can be used to overcome problem associated with other techniques discussed above. Another dimension of future work is to combine the Content-Based filtering approach with other techniques. For instance, it may co-work with a search engine and anomaly based phishing detection technique. Given a web page, the phishing detector first analyzes its contents then using the Identity Extractor, URL is fed into a search engine as a keyword. If the search engine returns a URL with the same domain, this web page under checking is genuine with an overwhelmingly high probability.

Acknowledgment

Working on this survey paper has been great learning experience for me. It will be my pleasure to acknowledge, utmost cooperation & valuable suggestion time to time given by staff members of Department. My greatest gratitude is reserved for my project guide Prof. Rajesh Tiwari. He always helped me by introducing me to the world of research. I am thankful for his benevolence, valuable suggestion, constructive criticism & active interest in successfully making of this paper.

References

[1] T. Moore and R. Clayton, 2007, "Examining the impact of website take-down on phishing," In proceedings of Anti-Phishing Working Group eCrime Researcher's Summit (APWG crime), ACM, pp. 1-13.

[2] S Sheng, B Wardman , Warner , G Cranor, L F Hong, J Zhang , 2009, "An- empirical analysis of phishing blacklists" In: CEAS 2009: Proceedings of the 6th Conference on Email and Anti-Spam

[3] J H Huh and H Kim, 2012, "Phishing Detection with popular search engine: Simple and effective", In Proceeding FPS'11 Proceedings of the 4th Canada-France MITACS conference on Foundations and Practice of Security, pp 194-207.

[4] Anti-Phishing Working Group. Phishing activity trends report for the month of December 2007, 2008.

[5] Y Pan, X Ding,2006, "Anomaly Based Web Phishing Page Detection", In Proceeding of the 22nd Annual Computer Security Applications Conference (ACSAC'06), pp 381-392.

[6] M. Jakobsson and A. Tsow, 2007, "Making takedown difficult", In M. Jakobsson and Steven, editors, Phishing and Countermeasures, pp461–467.

[7] M. Gupta, "Spoofing and countermeasures", 2007, In M. Jakobsson and S. Myers, editors, Phishing and Countermeasures, pp 65–104.

[8] M Tyler, C Richard and S Henry, 2009, "Temporal correlations between spam and phishing websites", In Proceedings (LEET'09) of the 2nd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more, pp5-5.

[9] R Dhamija, D Tygar, and M Hearst, 2006, "Why phishing works" In CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, pages 581–590.

[10] R.Dhamija and J. D. Tygar, 2005, "The battle against phishing: Dynamic security skins". In SOUPS '05: Proceedings of the 2005 symposium on Usable privacy and security, pages 77–888.

[11] M. Rossignol and P.sebillot, 2005 , "Combining statistical data analysis techniques to extract topical keyword classes from corpora", in Intell .Data Anal,9(10), pp- 105-127.

Mr. Radhe Shyam Panda has completed his B.E in Computer Science and Engineering from Pt. Ravishankar Shukla University Raipur and he is Pursuing M.E in Computer Technology and Application from CSVTU Bhilai. Currently he is working on various anti-phishing mechanisms.

Mr. Rajesh Tiwari has completed his A.M.I.E in Computer Engineering from IE (India) and received master degree in Computer Technology and Application from CSVTU Bhilai. He is pursuing PhD from CSVTU Bhilai in parallel computing. There is 15 research paper of him in National and **International** Journal. He has 15 years of teaching experience both in Under graduate as well as Post graduate. His research areas are parallel programming and parallel computing.