Mitigating Online Fraud by Ant phishing Model with URL & Image based Webpage Matching

T.BALAMURALIKRISHNA, N.RAGHAVENDRASAI, M.SATYA SUKUMAR

Abstract : Phishing is a malicious form of Internet fraud with the aim to steal valuable information such as credit cards, social security numbers, and account information. This is accomplished primarily by crafting a faux online presence to masquerade as a legitimate institution and soliciting information from unsuspecting customers. Phishing is a form of online fraud that aims to steal a user's sensitive information, such as online banking passwords or credit card numbers. The victim is tricked into entering such information on a web page that is crafted by the attacker so that it mimics a legitimate page.

In this paper, we present a novel technique to visually compare a suspected phishing page with the legitimate one. The goal is to determine whether the two pages are suspiciously similar. We identify and consider three page features that play a key role in making a phishing page look similar to a legitimate one. These features are text pieces and their style, images embedded in the page, and the overall visual appearance of the page as rendered by the browser. Due to the malicious nature of phishing attacks, identifying them bears higher demands in detection than filtering spam or other nuisance content. This paper establishes some requirements for phishing identification and explains various approaches to detection by looking for copying of web site layout and structure through source code (and optionally image) fingerprinting.

Keywords- Anti-Phishing, Web document analysis, Security, Visual Similarity, Webpage Matching

1

. INTRODUCTION

Successful phishing attacks are based on a form of copying, or reengineering, a website's design and layout in order to pass themselves off as a genuine (targeted) website. A malicious website is crafted which looks and feels like the original site, convincing unsuspecting users that they are giving personal information to a trusted organization. Users are frequently drawn to the sites by forged emails designed to look like legitimate correspondence and may even copy the body from real email, but when the user clicks a link to visit the website, they will be directed to the malicious site instead. The more convincing a phishing attack appears - or rather, the more genuine a malicious website looks - the more success the attack will have in extracting personal information.

2. The Growing Threat of Email Fraud

According to the Anti-Phishing Working Group (APWG) – an industry association dedicated to eliminating fraud and identity theft via phishing and spoofing – the number of unique phishing attacks has risen dramatically. Attacks monitored by APWG rose from 116 in December 2003 to 1, 422 in June 2004 – a 12-fold increase in this six-month period. Customers of companies in the financial sector (led by attacks on Citibank and U.S. Bank) and the online retail sector (led by attacks on eBay and PayPal) are most often attacked. Phishing Web sites are not overly concentrated in one country. The U.S. was the country that hosted the most phishing Web sites, with 27 percent of the total, followed by Korea with 20 percent, and China with 16 percent.

• Phishing Web sites has short lives, with an average lifespan of 2.25 days.

• About one quarter of phishing Web sites are hosted on hacked Web servers.

• Almost all phishing Web sites (94 percent) enable their developers to remotely download captured Personal data.

3. A Typical Example of Online Fraud

Figure 1 illustrates how the one-two punch of brand-spoofing and phishing can threaten the brand quity of an organization, in this case a fictitious bank. In the latter stages of the fraud, the trust and Confidence of customers is undermined.

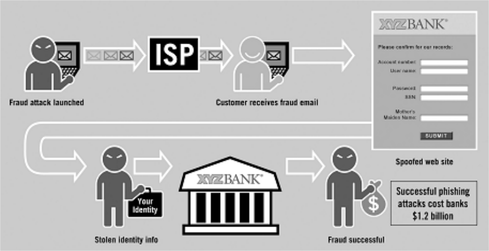

Figure 1. The process of email fraud includes customer receipt of an initial brand-spoofing email, content in the email that lures unsuspecting customers to a Web site that appears to be a legitimate financial institution site, and a request for customer information in an online form.

4. Effective Methods of Mitigating Online Fraud

To date, no complete solution that mitigates online fraud has been available. Symantec is working with several industry groups, including APWG and the Financial Services Technology Consortium– Counter Phishing Initiative, to develop technologies that will address this problem in both the short-term and long-term. In the meantime, the FTC and other organizations advocate consumer education on the risks of online fraud, and legal measures can be pursued if the perpetrators of the crime can be identified.

• An email fraud detection, filtering, and alerting network • On-line customer education

• A desktop security assessment capability for customers of financial institutions

• An infrastructure and means for financial services customers to acquire the products and services needed to improve their level of protection.

• Consulting and assessment services.

The fraud detection network detects and blocks fraudulent email before it reaches financial services customers. In parallel, a single online destination – co-branded with the financial institution –allows customers to better understand security-related and fraud avoidance issues, test their exposure to online threats, and identify and address their security needs.

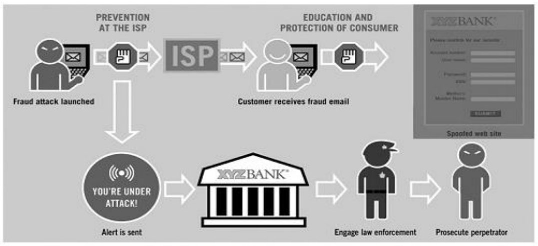

Figure 3. Symantec's Online Fraud Management Solution incorporates fraudulent email blocking, fraud attack alerting, customer education, assessment, and desktop protection combined to prevent online fraud and identity theft.

However, these techniques failed, as phishes are now composing phishing pages with non-analyzable elements, such as images and flash objects. This paper proposes a new phishing detection scheme based on an URL domain identity & webpage image matching. At first, it identifies the similar authorized URL, using divide rule approach and approximate string matching algorithm. For this similar URL and input URL, the IP addresses will be identified. If their IP addresses doesn't match with each other, then it could be phishing URL and phase-I phishing report will be generated. Then, this suspected URL's webpage snapshot will be treated as an image during phase-II.

In phase-II, key points will be detected and their features will be extracted. These features will be extracted using CCH descriptor. Then, match this suspected image features with the features of authorized webpage. If this matching crosses threshold value, then this webpage is phishing one. At last, final phishing report will be generated. As the combined approach of URL domain identity and webpage image matching used, it performs better than other existing tools.

5. Phishing Techniques

In a typical attack, the phisher sends a large number of spoofed (i.e. fake) e-mails to random Internet users that seem to be coming from a legitimate and well-known business organization (e.g. financial institutions, credit card companies, etc).A. Basic URL Obfuscation Ref [2], URL obfuscation misleads the victims into thinking that a link and/or web site displayed in their web browser or HTML-capable email client is that of a trusted site. These methods tend to be technically simple yet highly effective, and are still used to some extent in phishing emails today.

1. Simple HTML redirection: One of the simplest techniques for obscuring the actual destination of a hyperlink is to use a legitimate URL within an anchor element but have its href attribute point to a malicious site. Thus clicking on a legitimate-looking URL actually sends the user to a phishing site.

2. Use of JPEG images

Electronic mail rendered in HTML format is becoming more prevalent. Phishes are taking advantage of this by constructing phishing emails that contain a single image in JPEG format. When displayed, this image appears to be legitimate email from an online bank or merchant site. The image often includes official logos and text to add to the deception. However, when users click on this image, they are directed to a phishing site.

3. Use of alternate encoding schemes

Hostnames and IP addresses can be represented in alternate formats that are less likely to be recognizable to most people. Alphanumeric characters can be changed to their hexadecimal representations.

4. **Registration of similar domain names**

At initial glance, users may attempt to verify that the address displayed in the address or status bar of their web browser is the one for a legitimate site. Phishes often register domain names that contain the name of their target institution to trick customers who are satisfied by just seeing a legitimate name appear in a URL.A widely implemented version of this attack uses parts of a legitimate URL to form a new domain name as demonstrated below:

Legitimate URL http://login.example.com

Malicious URL http://login-example.com

6. Web Browser Spoofing Vulnerabilities

Over the past two years, several vulnerabilities in web browsers have provided phishers with the ability to obfuscate URLs and/or install malware on victim machines.

1. International Domain Names (IDN) Abuse:

International Domain Names in Applications (IDNA) is a mechanism by which domain names with Unicode characters can be supported in the ASCII format used by the existing DNS infrastructure. IDNA uses an encoding syntax called puny code to represent Unicode characters in ASCII format. A web browser that supports IDNA would interpret this syntax to display the Unicode characters when appropriate. Users of web browsers that support IDNA could be susceptible to phishing via homograph attacks, where an attacker could register a domain that contains a Unicode character that appears identical to an ASCII character in a legitimate site (for example, a site containing the word "bank" that uses the Cyrillic character "a" instead of the ASCII "a").

2. Web Browser Cross-Zone Vulnerabilities:

Most web browsers implement the concept of security zones, where the security settings of a web browser can vary based on the location of the web page being viewed. We have observed phishing emails that attempt to lure users to a web site attempting to install spyware and/or malware onto the victim's computer. These web sites usually rely on vulnerabilities in web browsers to install and execute programs on a victim's computer, even when these sites are located in a security zone that is not trusted and normally would not allow those actions.

7. Literature Review

A. Email-Level Approach

It includes authentication and content filtering. The email filtering techniques commonly used to prevent phishing. These are quite popular in antispam solutions because they try to stop email scams from reaching target users by analyzing email contents. Phishing messages are usually sent as spoofed emails; therefore, researchers have proposed numerous path-based verification methods. Current mechanisms, such as Microsoft's Sender ID or Yahoo's Domain Key, are designed by looking up mail sources in DNS tables. The challenge in designing such techniques lies in how to construct efficient filter rules and simultaneously reduce the probability of false alarms.

B. Browser Integrated Tool Approach

A browser-integrated tool [4,5] usually relies on a blacklist containing the URLs of malicious sites to determine whether a URL corresponds to a phishing page. In Microsoft Internet Explorer (IE) 7, for example, the address bar turns red when a malicious page loads.

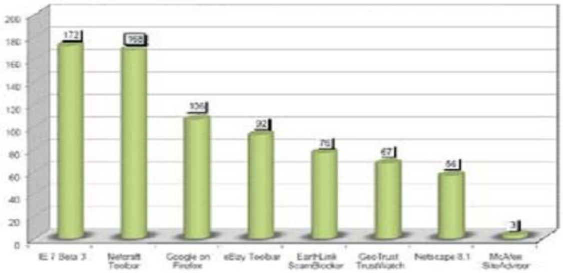

Fig. 2: Composite Accuracy Score Result

A blacklist's effectiveness is strongly influenced by its coverage, credibility, and update frequency .Currently, the most well-known blacklists are those Google and Microsoft maintain for the popular browsers Mozilla Firefox and IE, respectively. However, experiments show that neither database can achieve a correct detection rate greater than 90 percent, and the worst-case scenario can be less than 60 percent.

C. Visual similarity based analysis

New solution [6-7] is proposed by Anthony Fu and his colleagues, detecting phishing pages based on the similarity between the phishing and authentic pages at the visual appearance level, rather than using text-based analysis. An important feature of a phishing webpage is its visual similarity to its target (true) webpage. Hence, a legitimate webpage owner or its agent can detect suspicious URLs and compare the corresponding WebPages with the true one in visual aspects. If the visual similarity of a webpage to the true webpage is high, the owner will be alerted and can then take whatever actions to immediately prevent potential phishing attacks and hence protect its brand and reputation. This module extracts the Web pages' features and measures the similarity to the true

8. Proposed Work

This system proposes a new scheme for phishing page detection based on two phases as shown in fig. 3.

1. URL and Domain Identity

2. Image Based Webpage Matching

A. URL and Domain Identity Verification

Normally phishing is done via sending mails to thousands of users urging them to visit the fake website through the link or URL present in it. The input for proposed project is URLs for the detection process. These URLs are mostly similar to authorized URLs, with very minor variation which couldn't observed by normal users. Using approximate string search algorithm similar authorized URLs will be searched which are stored in database that is often targeted by phishers. Then calculate the IP addresses of the similar URLs. If IP addresses of the Authorized URLs do not match with the IP address of entered (input) URL then this URL could be phishing one. This URL will be considered as input for next phase which are based on the webpage's image matching.

B. Image Based Webpage Matching

In this phase, take a snapshot of a suspect webpage whose URL is detected as a suspected phishing URL in previous phase and treat it as an image throughout the detection process. The suspected webpage's snapshot is taken from the URL detected as phishing in earlier URL and Domain identity phase.

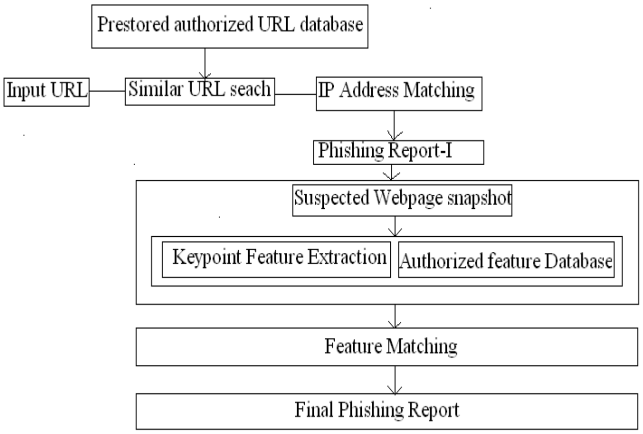
Prestored authorized URL database
Input URL
Similar URL seach
IP Address Matching
Phishing Report-I
Suspected Webpage snapshot
Keypoint Feature Extraction
Authorized feature Database
Feature Matching
Final Phishing Report

Fig:3 System Architecture of ProposedWork

This scheme from ref first calculates certain number of key points in a suspected webpage image. (The key point is a point, it can be detected, though image undergoes through various changes, such as shifting, lighting variation etc.). Use descriptors to capture invariant information around discriminative key points on the suspect page. Then match the descriptors with those of authentic page's descriptors' which are already stored in descriptors database. Matching descriptors yields a similarity degree for a suspect page and an authentic page. Finally, we use the similarity degree between the two pages to determine whether the suspected page is a counterfeit. If the similarity degree between a suspected page and an authentic one is greater than a certain threshold, we consider the suspected page is a phishing page.

9. System Design

A. URL and Domain Identity

1. Similar URL Search

The input URL is entered by user which is normally received by emails. Some sample URLs of payment services websites are enlisted as below. These URLs have taken from website of phish tank database. The data flow of this phase is shown in fig.4.

a. PayPal website

1.http://topsmiles.ru/smilies/authen/paypal_login/secur_redirect/Processing.php?cmd=_ing&dispatch=5885d80a13c0db1fb6947b0aeae66fdbfb2119927117e3a6f876e0fd34af4365494378e5d1704fcde593ec106fae5707494378e5d1704fcde593ec106fae5707

2.http://greensws.com/www.paypel.com/Fr/undispatsh=445qsd456qsd456q4d56q4sd564qsd56456f4s65g4df65465f4h654654fd56sq4df564qs65f4s6

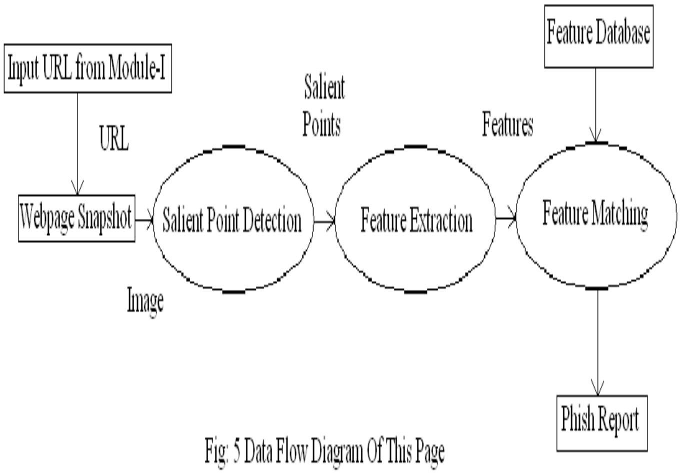

Fig: 5 Data Flow Diagram Of This Page

3. http: / /paypal .com.salsabi l t ravel .com/uk/cgi -bin/webscr/?cmd=_home-general&nav=0

b) EBay website

1.http://batangas.bhpi.com.ph/au/eBaySAPIdll_SignIn&=8&pUserId=&co_partnerId=2&siteid=15&pageType=1&pa1=&i1=1&UsingSSL=1&bshowgif=0&favoritenav=.ht

2. http://realavisor.com/version.php

3. http://mir3241.far.ru/signin.dll.html

From above examples it is observed that there are some common patterns of URLs structure exists. Techniques for generating input for approximate string matching algorithm are as follows. These URLs are too long. So, it is very difficult to analyze it sequentially and also its time consuming process. So we applied divide rule approach. These input URL is separated by slashes (/) .It will looks like below.

2. Similarity Ranking Algorithm: The steps of this algorithm are as follows. Input: Input URL substrings formed by above step i.e. through divide rule. Authorized URLs domain name stored in database.

Steps: Find out pairs of each string. Pair is formed of adjacent characters of string. E.g. Let authorized URL domain is paypal, then pairs= {pa, ay, yp, pa, al}.Then similarity between two pairs calculated by following formula.

Similarity (s1, s2) = | pairs (s1) Ώ pairs (s2)|*100 Pairs (s2) .Where s1= Input URL String, s2=Authorized URL, Pairs (s1) =Pairs for each substring of URL, Pairs (s2) = Pairs for Authorized URL Ώ=Intersection of pairs for authorized URL & input URL.

Output: Similarity Value

If the similarity value is equal or greater than 60 then the input URL substring is related to authorized URLs used for pairs which are stored in database. It becomes related authorized URL. If similarity value is less than 60 % then there may be possibility that no single word of input URL string related to any authorized.

URL in database.

In this case we have to extract html source content. From these html content source we will consider only <href> content i.e. the link to other WebPages. Then treat this reference URL as inputURL string and repeat above steps as like an input URL. In above example, pairs for each substring are as follows.

http= {ht, tt, tp} greensws= {gr, re, en, sw, ws}com= {co, om}www= {ww, ww} paypel= {pa, ay, yp, pe, el}

Repeat the above step until all words pairs are find out. For authorized URLs, let's take two financial organizations' URLs. Pairs for them are as follows.

paypal= {pa, ay, yp, pa, al} ebay= {ab, ba, ay}

For each authorized URLs and input URL substring calculate similarity value.Similarity value for paypel and paypal is

Pairs(s1)={pa,ay,yp,pe,el} Pairs(s2)={pa,ay,yp,pa,al}

Pairs (s1) ΏPairs (s2) = {pa, ay, yp}

|Pairs (s1) ΏPairs (s2)|=3 ,|Pairs (s2)| = 5 Similarity value= (3/5)*100=60 .So, this input URL is related to paypal

B. Image Based Webpage Matching

Following is the dataflow of this module.

1. Image salient Point Detection

It calculates salient points in webpage image by corner detection methods. Salient points in an image is a point considered a key point if it can still be detected after the image undergoes various changes, such as shifting, lighting variation, color transformation, or format conversion. Use the Harris-Palladian corners as the images points.

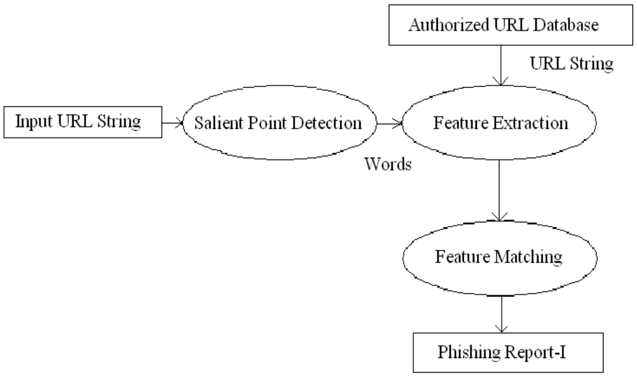

Fig:4 Data Flow Diagram of URL & Domain Identity Phase

2. Feature Extraction and Contrast Context Histogram

Features of these salient points extracted by using any descriptor. Use the Contrast Context Histogram (CCH) descriptors to capture invariant information around discriminative key points on the suspect page. To determine whether two images are similar, a common approach involves extracting a vector of salient features from each image and computing the distance between those vectors. We take this distance as the degree of visual difference between the two images. To construct CCH descriptors for an image, we use only gray-level information, which we obtain by averaging the red, green, and blue values of each pixel in the image.

3. Feature Matching

To determine whether a suspected web page is a phishing page or not the evaluation of its similarity to the potential target based on features extracted in above step. Ideally, the number of successful feature matches the descriptor finds will indicate the degree of similarity between the two pages A threshold is chosen, if similarity degree of two webpage images crosses threshold limit, then this webpage will be detected as phishing one.

10.Conclusion:

Phishing differs from traditional scams primarily in the scale of the fraud that can be committed. Con artists have been around for centuries, but E-mail and the World Wide Web provide them with the tools to reach thousands or millions of potential victims in minutes at almost no expense. Since there is no face-to-face contact between the attacker and the consumer, the consumer has very little information to work with in order to decide if an E-mail or web site is legitimate.

Thus, Phishing has become a major threat to information security and personal privacy. This paper represents new ant phishing technique based on URL domain identity and image matching mechanism. It first identifies the related authorized URL. We used approximate string matching algorithm. The image matching mechanism uses key point's detection and feature extraction methods. Two techniques i.e. URL domain identity and image webpage matching are combined, so this proposed work performs better than other existing tools. .The phase-II implementation is in progress. Further research will extend the system to increase performance by parallel executing these two modules (phases). This will reduce latency period of detection of phishing URLs.

References:

[1] Identity Theft Resource Center, Facts & Statistics, http://www.idtheftcenter.org /facts.shtml.

[2]"Phishing Attack Trends Report, June 2004," Anti-Phishing Working Group, http://www.antiphishing. org APWG_Phishing_Attack_Report-Jun2004.pdf.

[3]The Anti-Phishing Working Group, "APWG Phishing Trends Reports, [Online] Available: www.antiphishing.org/phishReports Archive.html

[4]P. Robichaux, D.L. Ganger, "Gone Phishing: Evaluating Antiphishing Tools for Windows," 3Sharp Project Report, Sept. 2006; [Online] Available : www.3sharp.com/ projects/antiphishing/.

[5] L. Wenyin et al., "Detection of Phishing Webpages Based on Visual Similarity," Proc. World Wide Web Conf. (special interest tracks and posters), A. Ellis and T. Hagino, eds., ACM Press, 2005, pp. 1060–1061.

[6]W. Liu et al., "An Antiphishing Strategy Based on Visual Similarity Assessment", IEEE Internet Computing, vol. 10, no. 2, 2006, pp. 58–65