Antiphishing Model with URL & Image based Webpage Matching

Madhuri S. Arade
Student
Shivaji University, Kolhapur
Maharashtra, India

P.C. Bhaskar
Professor
Department of Computer Science & Technology
Shivaji University, Kolhapur
Kolhapur, Maharashtra, India

R.K.Kamat
Professor
Department of Electronics
Shivaji University, Kolhapur
Kolhapur, Maharashtra, India

ABSTRACT
Phishing is a form of online identity theft associated with both social engineering and technical subterfuge and a major threat to information security and personal privacy. Many anti-phishing solutions, such as content analysis and HTML code analysis, rely on this property to detect fake web pages. However, these techniques failed, as phishers 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, keypoints 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.

Keywords
Phishing, networking, keypoints, string matching, image Matching

1 INTRODUCTION
Phishing is a brand spoofing a variation on "fishing," the idea being that bait is thrown out with the hopes that while most will ignore the bait, some will be tempted into biting.

Phishing is a form of online criminal trick of stealing victims' personal information by sending them spoofed emails urging them to visit a forged webpage that looks like a true one. Phishing is a form of online identity theft associated with both social engineering and technical subterfuge. Specifically, phishers attempt to trick Internet users into revealing sensitive or private information, such as their bank account, credit-card numbers and passwords. Users are often lured to browse these web sites through spoofed email, and they might easily be convinced that fake pages with hijacked brand names are authentic.

We ask that authors follow some simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace the content with your own material.

1.1 APWG-Phishing Activity Trends Summary (2009-2010)
The Anti-Phishing Working Group (APWG) [1] is the global pan-industrial and law enforcement association focused on eliminating the fraud and identity theft that result from phishing, pharming and email spoofing of all types.

Payment Services are the most targeted industry sector after Financial Services held top position during 2009 as shown in figure 1. However, the category of ‚Other rose from 13 percent to nearly 18 percent from Q4 2009 to Q1 2010, an increase of nearly 38 percent. The increase in the 'Other' category is attributed to the sharp increase in attacks against the online classifieds, social networking and gaming industries. The United States continued its position as the top country hosting phishing sites during the first quarter of 2010 with China maintaining a top three listing during the three month period.

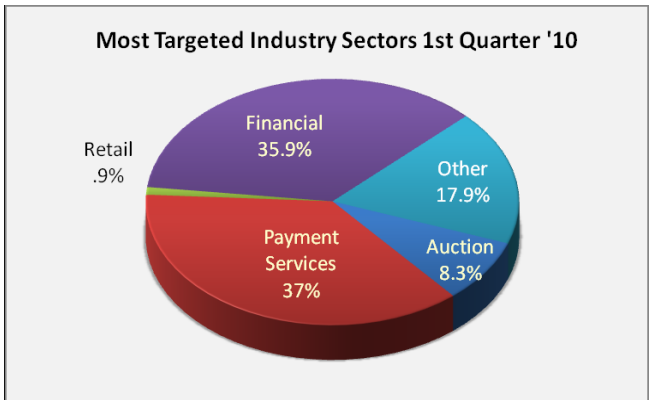

Figure 1. Most Targeted Industry Sectors 1st Quarter 10

2 PHISHING TECHNIQUE

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).

2.1 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.

2.1.1 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.

2.1.2 Use of JPEG images

Electronic mail rendered in HTML format is becoming more prevalent. Phishers 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.

2.1.3 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.

2.1.4 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. Phishers 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

2.2 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.

2.2.1 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 "а" instead of the ASCII "a").

2.2.2 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.

2.3 Specialized Malware

Over the past two years, there has been an emergence of malware being used for criminal activity against users of online banking and commerce sites. This type of specialized malware (which can be considered a class of spyware) greatly increases the potential return on investment for criminals, providing them with the ability to target information for as many or as few sites as they wish. One benefit for criminals is that most malware can easily be reconfigured to change targeted sites and add new ones. Malware also provides several mechanisms for stealing data that improve the potential for successfully compromising sensitive information.

3 LITRATURE REVIEW

3.1 Email-Level Approach

It includes authentication and content filtering. The email filtering techniques, in ref [3] 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.

3.2 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.

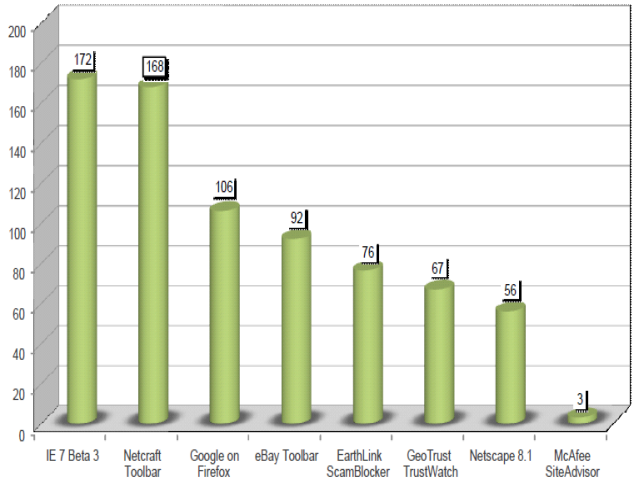

Figure 2.Composite Accuracy Score Result

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. Figure 4 shows the accuracy of various toolbars.

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.

3.3 Webpage Content Analysis

It analyzes a Web page's content [5], such as the HTML code, text, input fields, forms, links, and images. In the past, such content-based approaches proved effective in detecting phishing pages.

Phishers responded by compiling pages with non-HTML components, such as images, Flash objects, and Java applets. A phisher might design a fake page composed entirely of images, even if the original page contains only text information. In this case, content-based antiphishing tools can't analyze the suspect page because its HTML code contains nothing but HTML <img/> elements.

3.4 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 pages according to three metrics: block-level, layout, and style. If the visual similarity is higher than the corresponding threshold, the system issues a phishing report to the customer.

However, this approach is susceptible to significant changes in the Web page's aspect ratio and important colors used.

4 PROPOSED WORK

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

1. URL and Domain Identity

2. Image Based Webpage Matching

4.1 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.

4.2 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.

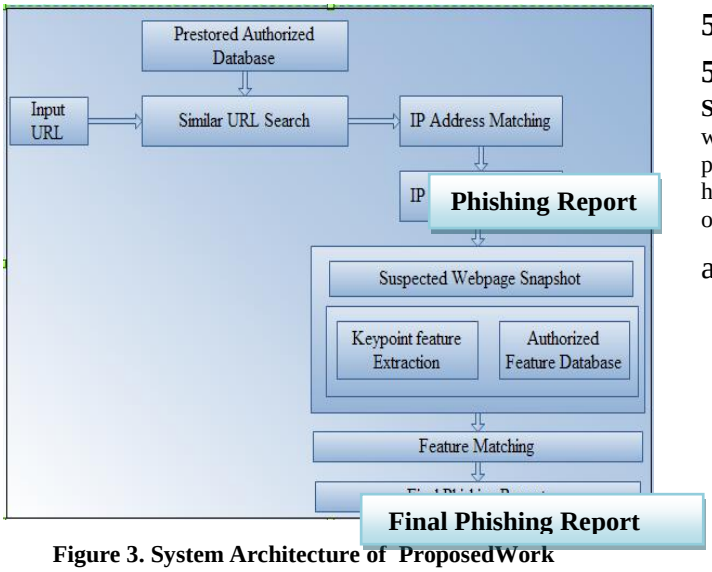

Figure 3. System Architecture of ProposedWork

This scheme from ref [8] first calculates certain number of keypoints in a suspected webpage image. (The keypoint 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 keypoints 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.

4.3 Methods of Data Collection

Proposed work mainly related to the financial services, payment services websites. In this project input data will be the URLs of websites and the snapshot of webpage's of these URLs. Databases consist of authorized URLs and their webpage's descriptors (features) as well as suspicious URLs and their webpage's descriptors. So there are following possible ways to collect data from different sources.

4.3.1 Databases

Some websites provides the available datasets for suspicious URLs and snapshots of webpage's for phishing detection e.g. phishtank database. This database has records of URL for the suspected website that has been reported & consists of the time of that report, and further detail such as the screenshots of the website.

Authorized financial related web sites URL and snapshot features of webpage which are often targeted by phishers will be taken as reference to prepare database which will be used for data analysis.

5 SYSTEM DESIGN

5.1 URL and Domain Identity

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 figure 4.

a) paypal website

1. http://topsmiles.ru/smilies/authen/**paypal_login**/secur_redirect/Processing.php?cmd=_Processing&dispatch=5885d80a13c0db1fb6947b0aeae66fdbfb2119927117e3a6f876e0fd34af4365494378e5d1704fcde593ec106fae5707494378e5d1704fcde593ec106fae5707
2. http://greensws.com/www.**paypel**.com/Fr/undispath=445qsd456qsd456q4d56q4sd564qsd56456f4s65g4df65465f4h654654fd56sq4df564qs65f4s6

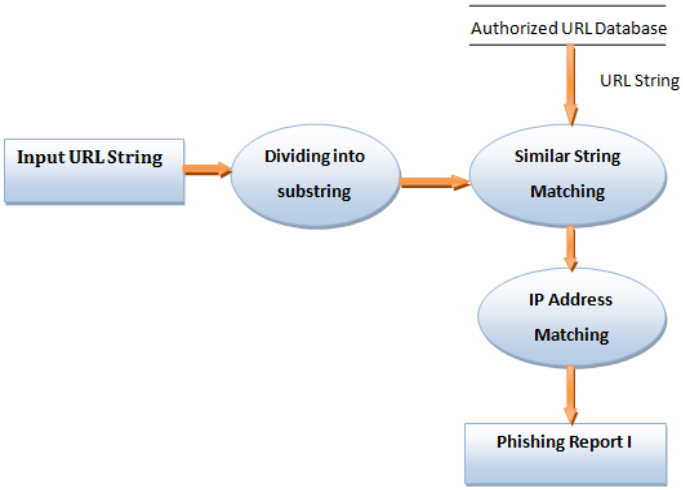

Figure 4. Data Flow Diagram of URL & Domain Identity phase

3. http://**paypal**.com.salsabiltravel.com/uk/cgi-bin/webscr/?cmd=_home-general&nav=0

b) ebay website

1. http://batangas.bhpi.com.ph/au/eBayISAPI.dll_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.

c) Example 1:

http://greensws.com/www.**paypel**.com/Fr/undispatsh=445qsd456qsd456q4d56q4sd564qsd56456f4s65g4df65465f4h654654fd56sq4df564qs65f4s65dqf465s4f65s4qdf6546548d7f65sqdf41sd4f654sqdf567s.

After applying divide and conquer algorithm approach, we get the result as below.

Repeat this algorithm steps until we get an individual word which is separated by all [**.** - _] these signs.

Then output will be as shown below.
http, greensws, com, www, paypel, com, Fr, undispatch.

Most of these URLs contain the word which is similar to at least one authorized URL domain name. Only it has some slight changes in spellings, some addition of characters or some deletion of characters.

E.g. paypal word replaced with paypail, paypel, paypal_login.

There are lots of algorithms for exact string matching algorithm such as Boyer Moore, Knuth-Morris-Pratt, Naïve Search, Quick Search algorithms. But proposed work requires string matching approximately. Hence in this proposed scheme new approximate string matching algorithm called as similarity ranking algorithm is used.

5.1.1 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

$$\text{Similarity}(s1, s2) = \frac{|\text{pairs}(s1) \,\Omega\, \text{pairs}(s2)| * 100}{\text{Pairs}(s2)}$$

Where s1= Input URL String,

s2=Authorized URL,

Pairs (s1) =Pairs for each substring of URL,

Pairs (s2) = Pairs for Authorized URL

Ω=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 input URL 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) ΩPairs (s2) = {pa, ay, yp}

|Pairs (s1) ΩPairs (s2)|=3

|Pairs (s2)| = 5

Similarity value= (3/5)*100=60

So, this input URL is related to paypal.

IP Address Matching

In this step, IP addresses of the input URL and related authorized URL result from above similar URL search method is calculated. Then compare these two IP addresses, if they don't match then this input URL could be phishing URL and it will be considered input for next module.

As the input URL is too long. Only the part of URL from http to first domain is considered for IP address calculation.

For above example the input URL considered for IP address is,

http://greensws.com. – IP Address is -**205.134.253.122**

Related authorized URL which is result from above similar URL search step will concatenated with all possible domains and then IP address for each will be calculated.

For above example 1 input URL is related to **paypal** website. Let's consider for two domains .com and co. in.

After concatenation result is **paypal.com, paypal.co.in**

IP addresses paypal.com - 64.4.241.45

paypal.co.in - 64.4.241.161

Compare the input URLs IP address with all related URL domains IP addresses from step 2 ,if these are different then this input URL could be phishing one.

5.2 Image Based Webpage Matching

Following is the dataflow of this module.

5.2.1 Image salient Point Detection

It calculates salient points in webpage image by corner detection methods [8]. Salient points in an image is a point considered a keypoint 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-Laplacian corners as the images keypoints.

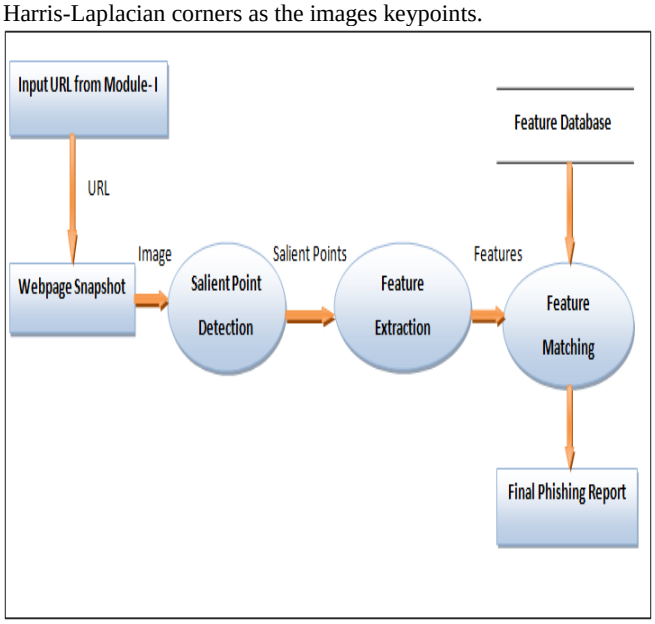

Figure 5. DFD of Image matching phase

5.2.2 Feature Extraction

Features of these salient points extracted by using any descriptor. Use the Contrast Context Histogram (CCH) descriptors to capture invariant information around discriminative keypoints on the suspect page.

5.2.2.1 Contrast Context Histogram

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 [10] 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. The proposed approach considers a histogram based representation of the contrast values in the local region around the salient corners.

5.2.3 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 [9].

A threshold is chosen, if similarity degree of two webpage images crosses threshold limit, then this webpage will be detected as phishing one.

6 CONCLUSION

Thus, Phishing has become a major threat to information security and personal privacy. This paper represents new antiphishing 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 keypoints 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.

7 ACKNOWLEDGEMENT

M.S.Arade Author thanks to Prof. Bhaskar P.C., project guide, Mr.Kamat, project co-guide, Department of Computer Science & Technology, Kolhapur, India, for their precious guidance, encouragement and continuous monitoring throughout the presented work.

8 REFERENCES

[1] The Anti-Phishing Working Group,APWG Phishing Trends-Reports, www.antiphishing.org/phishReports Archive.html

[2] Jason Milletary, Technical Trends in Phishing Attacks, Carnegie Mellon University,2005

[3] Sumit Siddharth, Anti Spamming Techniques.pdf.

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

[5] C. Ludl et al., On the Effectiveness of Techniques to Detect Phishing Sites, Proc. Detection of Intrusions and Malware, and Vulnerability Assessment, LNCS 4579, Springer, 2007, pp. 20–39

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

[7] 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.

[8] IEEE 2009 Paper on Fighting Phishing with Discriminative Keypoint Features by K. J. Lin, Yan Wang.

[9] J. Han and M. Kamber,Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.

[10] C.-R. Huang, C.-S. Chen, and P.-4. C. Chung, ContrastContext Histogram — An Efficient Discriminating Local Descriptor for Object Recognition and Image Matching,Pattern Recognition, vol. 41, no. 10, 2008, pp.3071–3077; http://imp.iis.sinica.edu.tw/CCH/CCH.htm.