A Phishing Sites Blacklist Generator

Mohsen Sharifi and Seyed Hossein Siadati
Computer Engineering Department
Iran University of Science & Technology
msharifi@iust.ac.ir, s_h_siadaty@comp.iust.ac.ir

Abstract

Phishing is an increasing web attack both in volume and techniques sophistication. Blacklists are used to resist this type of attack, but fail to make their lists up-to-date. This paper proposes a new technique and architecture for a blacklist generator that maintains an up-to-date blacklist of phishing sites. When a page claims that it belongs to a given company, the company's name is searched in a powerful search engine like Google. The domain of the page is then compared with the domain of each of the Google's top-10 searched results. If a matching domain is found, the page is considered as a legitimate page, and otherwise as a phishing site. Preliminary evaluation of our technique has shown an accuracy of 91% in detecting legitimate pages and 100% in detecting phishing sites.

1. Introduction

Phishing attack is a type of identity theft that aims to deceit users into revealing their personal information which could be exploited for illegal financial purposes.

A phishing attack begins with an email that claims it is from a legal company like eBay. The content of email motivates the user to click on a malicious link in the email. The link connects the user to an illegitimate page that mimics the outward appearance of original site. The phishing page then requests user's personal information, like online banking passwords and credit card information.

The number of phishing attacks has grown rapidly. According to trend reports by Anti-Phishing Working Group (APWG) [1], the number of unique phishing sites has been reported 37,444 sites in October 2006, increased from 4,367 sites in October 2005. Other statistics show the increase in the volume of the Phishing attack and their techniques are becoming much more advanced.

A number of techniques have been studied and practiced against phishing and a large number of them use phishing blacklists to battle against phishing. Blacklists of phishing sites are valuable sources that are in use by anti-phishing toolbars to notify users and deny their access to phishing sites, web and email filters to filter spam and phishing emails, and phishing termination communities to terminate the phishing sites.

Blacklist indicates whether a URL is good or bad. A bad URL means that it is known to be used by attackers to steel users' information. The blacklist publisher assigns the "goodness" (the URLs that are not in the list) and the "badness" (the URLs that are in the list) to all internet URLs [1].

Many browsers now check blacklist databases to address phishing problem and notify users when they browse phishing pages. Internet Explorer 7 [3], Netscape Browser 8.1[4], Google Safe Browsing (a feature of the Google Toolbar for Firefox)[5] are important browsers which use blacklists to protect users when they navigating phishing sites.

Due to the wide use of blacklists of phishing sites against phishing, it is very important to introduce techniques that generate the updated blacklists of phishing sites. The problem of the blacklist is that it is hard to keep the list up-to-date since it is easy to register new domains in the Internet. In this paper we propose a technique to detect deceptive phishing pages, as well as our proposed architecture for a blacklist of phishing sites generator.

The rest of paper is organized as follows. Section 2 discusses related works. Section 3 presents our proposed algorithm and the architecture of our blacklist generator. The evaluation of the approach is given in section 4. The paper is concluded in Section 5.

978-1-4244-1968-5/08/$25.00 ©2008 IEEE

2. Related Works

There are techniques that generate blacklist and some techniques that have the potential to be used as blacklist generator. We discuss these techniques here.

Users reports and community ratings is a technique to generate blacklists of phishing sites. Cloudmark Collaborative Security Network (CCSN) [6]consists of real-time users who are themselves targets and can distinguish a phishing attack and mark it. IE7 0 and Earthlink [7] browser toolbars also provide facilities for users to report the pages that they detect as phishing. A central repository gathers the reported URLs and makes a blacklist.

Spoofguard [8] is a heuristic technique that uses domain name, URL, link, and image checks to evaluate the likelihood that a given page is part of a spoof attack. Netcraft [6] is a toolbar that uses URL heuristic analysis to detect phishing pages. Netcraft traps suspicious URLs containing characters which have no common purpose. Earthlink [7] and McAfee SiteAdvisor [9] use information about owner, age and country of domain registration to estimate the likelihood of pages to be phished. McAfee SiteAdvisor also investigates a number of links to legitimate sites as a heuristic to detect phishing sites.

[12]Proposes a technique that detects phishing web pages based on visual similarity. The technique uses measures in three metrics to compare the similarity of pages: block level similarity, layout similarity, and overall style similarity.

CANTINA [13] is an automated approach to detect phishing pages. The main idea is similar to ours but its keyword extraction part differs. They use the TFIDF to extract keywords from the page. Instead, we use OCR and image processing techniques to extract keywords from the page. We will discuss more differences and tradeoffs related to these approaches in the next sections.

Although some of the above techniques merit generating a blacklist, each one uses a number of non-resisting properties of phishing pages. Phishers can avoid these properties and make phishing pages appear credible. A common property of all phishing pages is that they claim that they belong to a site using their outward appearance. We use this common property to detect phishing pages.

3. Proposed Blacklist Generator

Our proposed technique tries to generate an updated blacklist of phishing sites. Each web page belongs to a web site and most of them show this relation using the site's logo. Phishing pages also use legitimate site's logo to make their pages credible and claim that they belong to that site. Thus we can find which site a page claims to belong, using its logo. On the other hand, the domain of a legal site can be found by searching its name in a search engine like Google. Our technique is based on these two properties of the web pages and search engines to detect phishing pages. Figure 1 demonstrates our algorithm which gets a URL as input and returns True if the page is phishing and False if the page is a legitimate one.

```
1: Procedure Boolean ISPhishingPage(String varURL)
2: varCompanyName = ExtractComanyName(varURL)
3: varGoogleResults =
   GoogleSearch(varCompanyName , 10)
4: for i = 1 to 10
5: BEGIN
6:   varGResultURL   = varGoogleResults[i].URL
7:   if (AreInSameDomain(varURL, varGResultURL))
       return False
8: END
9: return True
```

Figure 1: Black List Generator Algorithm

Determining which pages belong to the same web site is an open problem, although some heuristic approaches have been proposed [14]. We approximate each web site by all the pages with the same host name. It is a useful technique but not quite accurate. The output of Blacklist Generator is an XML document. Table 1 shows an example XML file that our Blacklist Generator produces.

Our experiments indicates that in 74.4% of cases, searching a company name in Google brings its web site's link as the first search result item and it is beneficial to suggest this item as alternatives for users when their access to a phishing site is blocked.

Table 1: A blacklist generated by our Blacklist Generator

```
<IUSTBlacklist>
.
.
<Item id=102
 phishingurl="http://Phisher/path"
 alternativeurl=http://www.eBay.com />
.
.
</IUSTBlacklist>
```

In most cases, the starting point of a phishing attack is an email. Users receive emails that contain suspicious links that direct them to phishing sites. The

link to all phishing sites can be found in the body of the email, so emails are a valuable source to make a blacklist of phishing sites. In addition, since our algorithm is time-consuming, it can be useful to Internet users who can tolerate due delays for their increased safety. So the best place to apply our algorithm is to an email server. Figure 2 shows our proposed architecture for the blacklist generator.

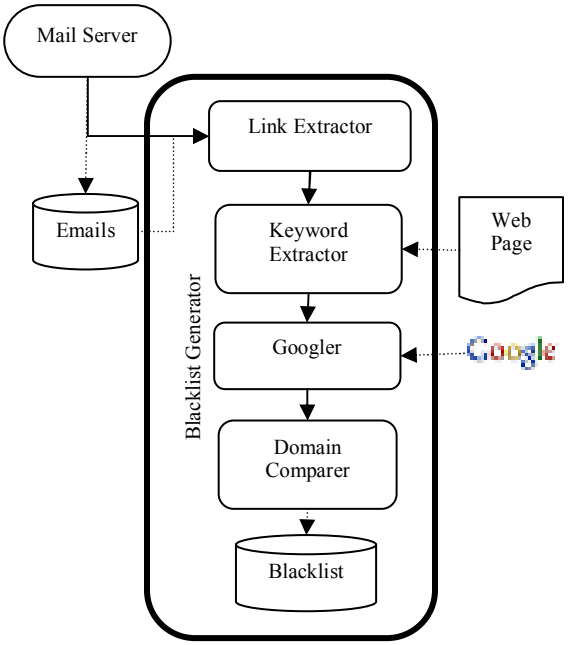

Figure 2: Architecture of Blacklist Generator

4. Evaluation

We fed our algorithm with datasets of legitimate and phishing pages. Keyword extraction and selection were done manually by human operators. So our test application is semi automated.

In order to have an accurate estimation of efficiency of our algorithm, the datasets must be selected accurately. In the case of legitimate pages, the dataset must be a uniform sample set of URLs. The legitimate pages dataset that we fed into our algorithm were selected using a near uniform web sampling technique. We took a dictionary of about 150,000 words and randomly selected 10 words. Then we recorded the link of top 50 items resulted from searching each randomized word in Google and generated a dataset of 500 samples of legitimate sites.

We fed these 500 samples in our test application and found that the algorithm wrongly classified 45 legitimate pages as phishing page. Thus the false positive rate of our algorithm was 9%.

We also fed 30 live phishing samples, selected from PIRT reports in September 2006, into our test application. As a result, our algorithm was able to detect all phishing pages correctly. Thus the false negative of our algorithm was 0% and our technique was successful in detecting phishing pages. The result is rational because none of the current phishing sites gain high rank in Google search results.

The number of items that we picked from Google results to compare with the input URL affects the false positive and false negative rates of our algorithm. Figure 2 show the relation between the number of top Google items that our algorithm had used and the false positive of our algorithm.

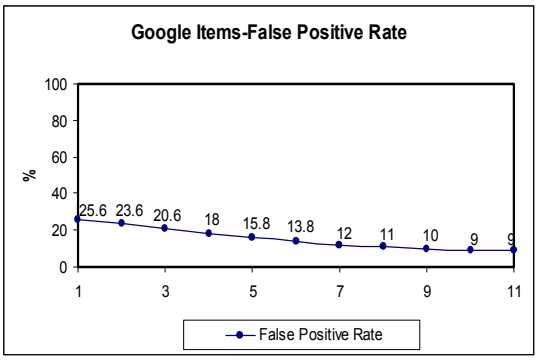

Figure 3: Google Items-False Positive Rate

There is a tradeoff between false positive rate and the chance of *Spamdexing* techniques to boost the rank of phishing pages up to our selected threshold.

A big threshold decreases the false positive rate but it raises the opportunity for phishers to enter into the top results in Google search. In contrast, a low threshold raises the false positive but it decreases the opportunity for phishers to boost their page rank up to our selected set of Google results.

According to the results shown in Figure 3, the false positive rate becomes nearly fixed to 9% for Google items more than 10 and it seems that 10 is a good threshold for our technique.

In addition, in 74.6% of the cases the link of legitimate site appears as the first item in Google search results. We use this item to make alternative URLs for a phishing URL. The tools that use the blacklist can use alternative URLs to provide alternatives for users and increase the usability of their tools.

CANTINA is successful in detecting phishing websites too, but it is vulnerable to Image-Instead-of-Text attacks which are increasingly used nowadays in phishing attacks. It disables the keyword extraction technique to work properly. In addition, phishers might

use the hidden text in HTML pages to evade the keyword extraction of this technique.

5. Conclusion

Currently many techniques use backlists to ignore users from navigating malicious URLs but the challenge is in generating an *updated* blacklist of phishing sites. The Blacklist Generator proposed in this paper aimed to provide an updated blacklist using page content and search engine results. Analysis of our technique showed an accuracy of 91% to detect legitimate pages and 100% to detect phishing sites.

Acknowledgement

We thank the Iran Telecommunication Research Center for partial financial support of the research project whose findings are reported in this publication.

6. References

[1] Anti-Phishing Working Group, "Phishing Activity Trends Report Combined", *Report*, Anti-Phishing Working Group, USA., October 2006.

[2] M. Wu, "Fighting Phishing at the User Interface", *PhD Thesis*, Mass. Inst. of Technology, 2004

[3] Microsoft, *Internet Explorer 7*, Last Access on December 14, 2006 http://www.microsoft.com/windows/ie/default.mspx.

[4] Netscape, Netscape browser, Last Access on December 14, 2006, http://browser.netscape.com/ns8/.

[5] Google Inc., "Google Safe Browsing", Last Access on December 14, 2006.http://www.google.com/tools/firefox/safebrowsing/.

[6] Cludmark Co., "Cludmark Unique Approach", Anti-Phishing Working Group, July 28, 2006, www.cloudmark.com/releases/docs/wp_unique_approach_10550406.pdf.

[7] EarthLink Inc., "EarthLink Toolbar" , Last Access on December 14, 2006, http://www.earthlink.net/software/free/toolbar/

[8] N. Chou, R. Ledesma, Y. Teraguchi, JC Mitchell, "Client-Side Defense Against Web-Based Identity Theft", The ISOC symposium on Network and Distributed System Security, San Diego, February 2004.

[9] McAfee, Inc., "McAfee SiteAdvisor", Last Access on December 14, 2006. http://www.siteadvisor.com/.

[10] Netcraft, Netcraft Browser, Last Access on December 14 2006, http://www.netcraft.com/

[11] Cloudmark, Last Access on December 14, 2006,http://www.cloudmark.com/serviceproviders/authority/phishing/

[12] L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, X. Deng, "Detection of Phishing Web Pages based on Visual Similarity", WWW (Special interest tracks and posters), pp. 1060-1061, ACM, 2005.

[13] Y. Zhang, J. Hong, and L. Cranor. "CANTINA: A Content-Based Approach to Detecting Phishing Web Sites", The 16th International conference on World Wide Web, Banff, Alberta, Canada, May 8-12, 2007.

[14] K Bharat, BW Chang, M Henzinger, M Ruhl, "Who Links to Whom: Mining Linkage between Web Sites", IEEE International Conference on Data Mining, San Jose, CA, USA, 2001.