Scientific Journal Impact Factor (SJIF): 1.711

e-ISSN: 2349-9745
p-ISSN: 2393-8161

International Journal of Modern Trends in Engineering and Research
www.ijmter.com

Analyzing the effectualness of Phishing Algorithms in Web Applications

Inquest of LinkGuard and CANTINA Approach

Mr. Vinod Singh Kharsan[1], Bernadine Rekha Soman[2], Sonali Sharma[3]
Assistant Professor[1], UG Scholar[2,3]
Department of Computer Science, Chouksey Engineering College, Bilaspur (C.G.)[1,2,3]

Abstract— The initial and proficient loss of deception is belief. A wolf in sheep's clothing is tough to recognize, similar is the schema of a phishing website. Phishing is the emulsion of social engineering and technical exploits designed to persuade a victim to provide personal information, for the fiscal gain of the attacker. It is a new kind of network assault where the attacker creates a spitting image of an already existing Web Page to delude users. In this paper, we will study two anti-phishing algorithms, one an end-host based algorithm known as the LinkGuard Algorithm, while the other a content based approach known as the CANTINA.

Keywords- Phishing, toolbars, Subroutine, LinkGuard, CANTINA, heuristics, visual.

I. INTRODUCTION

Progression of internet has played a eloquent role in trade and custom activities. Unfortunately, indigent security on the internet and large monetary gains furnish a strong motivation for attackers to commit such seemingly low risk, yet high- return online scams. Web –based casualties are usually caused by Phishing attacks, malware, spoofing and other threats, costs organizations millions of dollars every year. Web fraud also strikes user's insights of e-businesses. It is estimated that between May 2004 and 2005, approx. 1.2 million computer users in U.S suffered losses amounting up to US$929 million. As a matter of fact United States businesses lose an estimated US$2 billion per year as their clients become victims.[1]

Recently, there has been a dramatic increment in phishing, a relative attack in which victims are deluded by spoofed emails and fraudulent web sites into giving up patented information. Phishing is a quick growing challemge, with 9,255 unique phishing sites reported in June 2006 alone[2]. It is not familiar exactly how much phishing costs each year since impacted industries are reluctant to release figures; estimates range from $1 billion[3] to 2.8 billion[4] per year. To respond to this peril, software brokers and organizations have released a variety of anti phishing toolbars. As of September 2006, the free software download site download.com, listed 84 anti –phishing toolbars. However a strong need for superior automated detection algorithms was required. Phishers employ e-mail as their major method to carry out phishing attacks. However, this algorithm can be applied to attacks that use other means such as instant messaging. In general phishing attacks are performed with the following steps:

1) Phishers set up a forged web site which looks exactly like the legal Web site. Including setting up a server, applying the dns server name, and creating the web pages similar to the destination Website etc.

2) Then they send large amounts of mimic emails to target users under the name of those legitimate companies and troupes, in order to convince the potential victims to visit their Web sites.

3) Users then click on the hyperlink given, and input the required information.

4) Next the phishers steal the information and prosecute their fraud such as transferring money from the victims account.

Despite efforts such as blocking phishing Web sites, enhancing their security, using spam filters etc. it is impossible to avoid phishing. As a last defense, users can install anti-phishing tools in their computers. Anti-phishing tools used these days can be divided into three categories: blacklist/whitelist based, rule based and content based.

- Category 1: When a user sojourns a Web site, the anti phishing tool searches the address of that sites in a blacklist stored in the database. If the visited site is on the list, then the anti-phishing tool then warns the users. Tools such as the Scam Blocker from the Earthlink Company, PhishGuard and Net craft, etc. Nevertheless newly emerged Web sites cannot be prevented with this.
- Category 2: This category of tools uses a certain set of rules for their software. Softwares such as the SpoofGuard and TrustWatch use this method, i.e. it checks the domain name, URL of the Web site and also checks whether the browser is directed to the current URL via the links in the contents of emails. If it finds that the domain name of the visited Website is similar to the well-known website or domain name, then it warns the users.
- Category 3: This category of tools is a recent discovery. A content based approach for detecting phishing web sites known as CANTINA. It examines the content of the webpage such as the common characteristics like the URL and the domain name and determines whether the site is legitimate or not.

II. ANTI-PHISHING TOOLS AND ALGORITHMS

Here we interrogate the adopted methodologies, working principles and concluding results of both the algorithms and try to infer which out of the two is better. Shifting onto the methodologies of the algorithms we begin with the procedure adopted in the LinkGuard Algorithm and then we will move on to CANTINA.

2.1 LINKGUARD ALGORITHM

The LinkGuard algorithm is one such algorithm which defines the characteristics of URL and Domain names. The elementary work of this algorithm is to dissect the actual links and the visual links. Some of the important systems contained in the architecture of this algorithm can be given as follows:

Communication: All the input course is gathered and facts are sent to the analyzer.

Database: User's inputs such as the domain name, Url, Blacklists etc. are stored in here.

Analyzer: Data which is provided by the communication and the database is sent to the analyzer for modification purpose. Hence the analyzer can also be called as the Link Guard of the system.

Alerter: Every time something is received by the analyzer which has to be displayed to the user the alerter generates a warning message.

Logger: Information related to this particular session is stored.

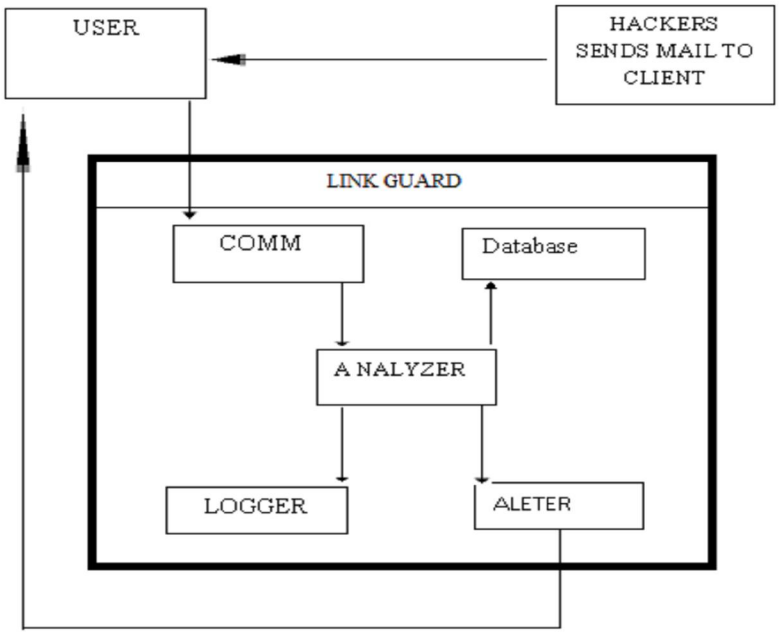

Fig.1. Architecture of LinkGuard Algorithm

Now we shall discuss procedure of the LinkGuard algorithm:

2.1.1 Class of the hyperlinks in the phishing e-mails:

To gather useful information from the user's phishers persuade the victims to click in the hyperlink which is embedded in the counterfeited site. A hyperlink has the following structure:

<aherf=”http://www.antiphishing.org/phishingarchive.html”>Phishing Archive

Where 'http://www.antiphishing.org/phishingarchive.html' is the "URI(universal resource identifiers)" which provides information to the user about the networked resource, and 'Phishing Archive' is the "Anchor text" which is the text displayed on the user's web browser.

2.1.2 The LinkGuard Algorithm:

Linkguard algorithm effectuates by analyizing the differences between the actual and the hyperlink. It also computes the resemblances of the URI with a legitimate site. The common terms used in the algorithms are:

v_link: visual_link;

a_link: actual_link;

v_dns: visual DNS name;

a_dns: actual DNS name;

sender_dns: sender's DNS name;

int LinkGuard (v_link, a_link} {


```
1 v_dns = GetDNSName (v_link);
2 a_dns = GetDNSName (a_link);
3 if ((v_dns and a_dns are not
4 empty) and (v_dns! = a_dns))
5 return PHISHING;
6 if (a_dns is dotted decimal)
7 return POSSIBLE_PHISHING;
8 if (a_link or v_link is encoded)
9 {
10 v_link2 = decode (v_link);
11 a_link2 = decode (a_link);
12 return LinkGuard (v_link2, a_link2);
13}
14 /* analyze the domain name for
15 possible phishing */
16 if (v_dns is NULL)
17 return AnalyzeDNS (a_link);
}


int AnalyzeDNS (actual_link) {
/* Anatomize the actual DNS name according to the blacklist and whitelist*/
18 if (actual_dns in blacklist)
19 return PHISHING;
20 if (actual_dns in whitelist)
21 return NOTPHISHING;
22 return PatternMatching(actual_link);
}

int PatternMatching (actual_link){
23 if (sender_dns and actual_dns are different)
24 return POSSIBLE_PHISHING;
25 for (each item prev_dns in seed_set)
26 {
27 bv = Similarity(prev_dns, actual_link);
28 if (bv == true)
29 return POSSIBLE_PHISHING;
30 }
31 return NO_PHISHING;
}
float Similarity (str, actual_link) {
32 if (str is part of actual_link)
33 return true;
34 int maxlen = the maximum string
35 lengths of str and actual_dns;
36 int minchange = the minimum number of
37 changes needed to transform str
38 to actual_dns (or vice verse);
39 if (thresh<(maxlen-minchange)/maxlen<1)
40 return true
```


41 return false;
} [5]

The Linkguard algorithm contains a main routine LinkGuard in which it extracts the DNS names from the actual and visual links. When a comparision of th names is links is made and it is found that these names are not same, or dotted decimal IP address is directly used in actual dns then phishing attack is recognized. On the other hand when there is no destination information(DNS or dotted IP address) in the visual link the subroutine AnalyzeDNS is called to handle phishing attack. While the pattern matching subroutine is used to identify other unkown attacks. In this method either the sender's email address is extracted from the e-mail or a collection of all the DNS names is done in a "seed set". After this the PatternMatching subroutine checks if the actual DNS name of a hyperlink is different from the DNS name in the sender's address.

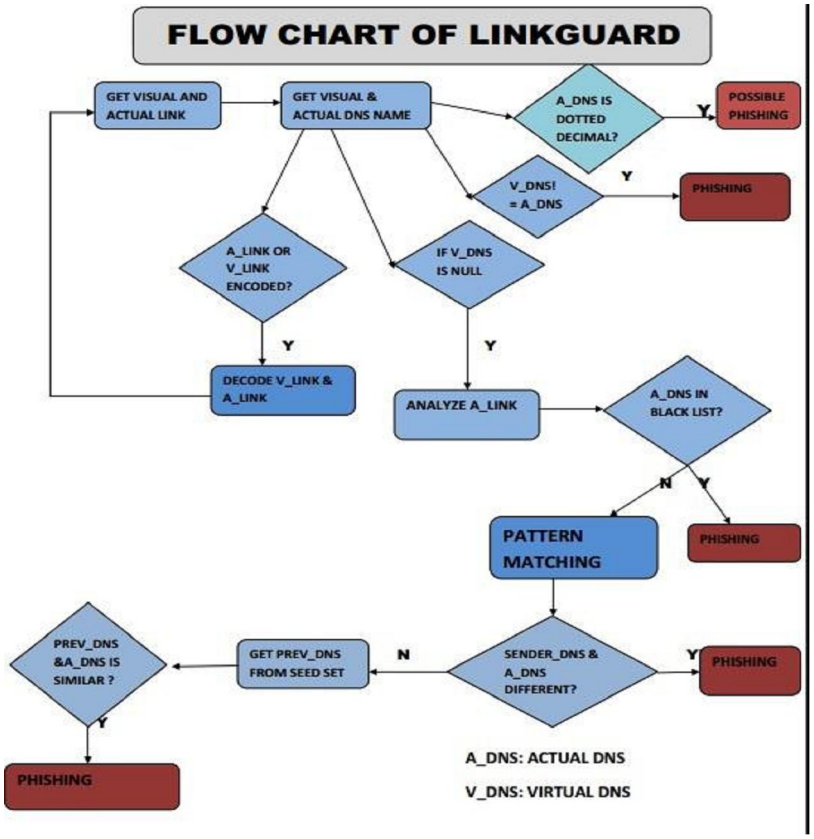

Fig. 2. Flow Chart of the LinkGuard Algorithm

Hence, in the above figure we have seen the workings and procedures involved in LinkGuard algorithm. Now we shall study the CANTINA algorithm.

2.2 CANTINA

The CANTINA is a unprecedented content-based passage of detecting phishing websites. What makes the cantina intriguing is that other methods look at the shell peculiarity of the web pages whereas the CANTINA studies the substance or content of that webpage. Not only this, the approach also analyzes

the text-based content of the page itself. This approach works on the principle of a heuristic based algorithm known as the TF-IDF algorithm.

Working of the TF-IDF:

The TF-IDF algorithm is used for the purpose of information retrieval and text-mining. It succumbs importance to the occurrence of a word in a document. As the number of word appears in the document the importance increases.

The term 'TF' stands for "term frequency" meaning the number of times a given term appears in a specific document. It usually helps in preventing a bias towards the longer documents and also measures the importance of that term within a particular document. On the other hand the term 'IDF' stands for "inverse document frequency" and is meant to define the general importance of the term i.e. its measures how common a particular word is in the entire document. Accordingly we can say that any term which has high TF-IDF weight means that it has high TF and low document frequency.

Alongside the TF-IDF algorithm Robust Hyperlinks are also required in the CANTINA approach. Phelps and Wilensky evolved the idea of Robust Hyperlink to overcome the problem of broken links[6]. They proposed on adding a small number of some chosen terms which they called as "lexical signature" to the URL's. Now here comes the role of the TF-IDF i.e. to generate a lexical signature.

In order to generate the signatures the TF-IDF value of each word of a document is calculated and words which have highest values are then selected. The term frequency defines wholesomeness and IDF provides rarity in documents.

So the work of the CANTINA can be supervened as:

- Calculate the TF-IDF scores of each term on a given web page.
- Clasp any five terms having the highest TF-IDF weights and generate a lexical signature.
- Provide this signature to any search engine of your choice.
- If the DNS name of the current webpage matches with the N top search results then we consider that there is no phishing or else it is a phish website.

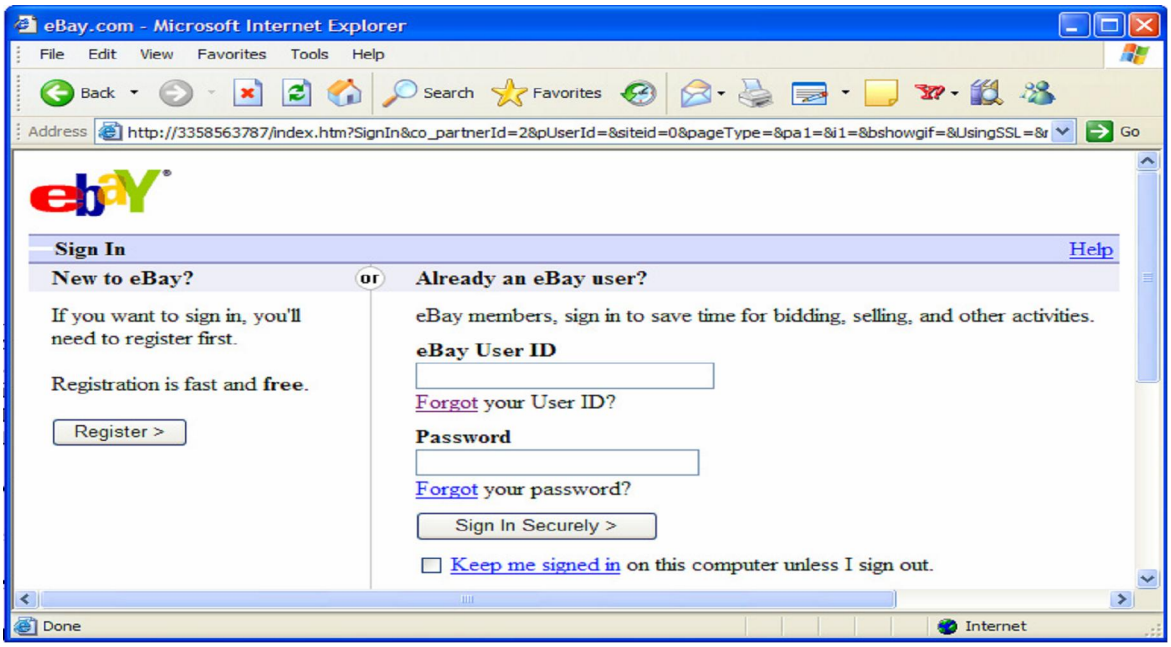

Fig. 3. Fake website which is uploaded to the internet.

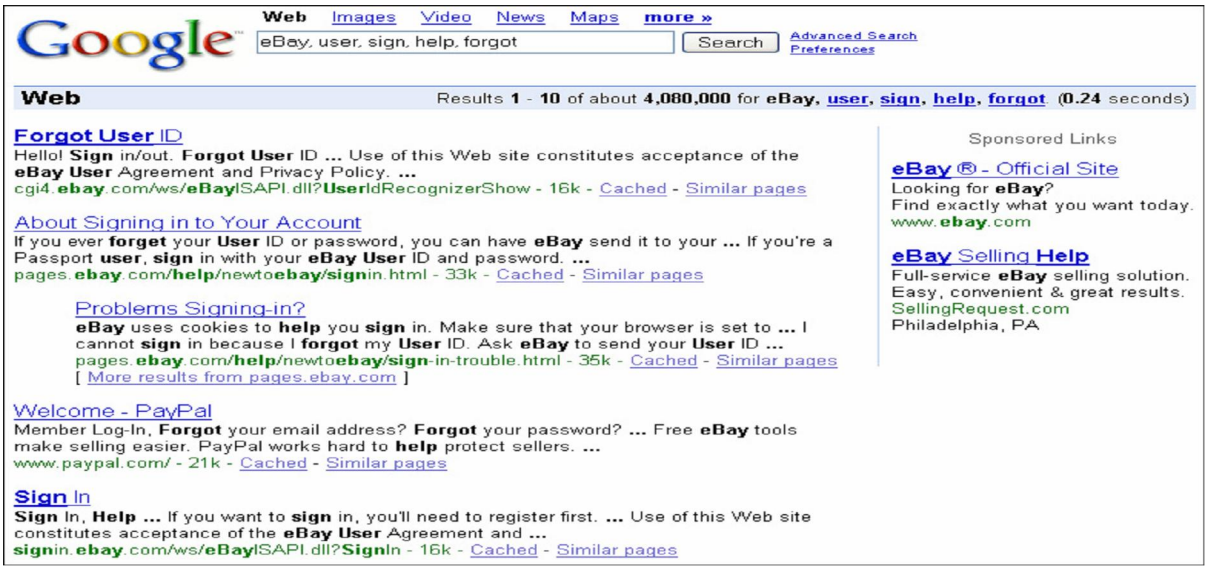

Fig. 4. The search engine showing results whether the site is legitimate with the help of the lexical signature.

III. COMPARING BENEFITS AND DRAWBACKS

Commencing with the Pros and cons of both the approaches discussed above would be given as-

Benefits of the Linkguard Algorithm can be given as:

- One great asset is that the LinkGuard algorithm recognizes known as well as unknown phishing sites.
- It is light-weighted and based on derived characteristics.

Drawbacks of the LinkGuard algorithm can be given as:

- Due its false negative differentiating nature the LinkGuard algorithm sometimes treats some phishing sites as normal websites(specifically those sites which have not been visted by the user at all).
- When the comparability index of the site is near close for example say 3/4 it often generates a false positive.

Now coming to the benefits of the CANTINA APPROACH:

- Fast in nature and is based on the content-based characteristics.
- No maintenance required by the user as most functionalities are performed by the search engines.

Drawbacks of the CANTINA approach can be given as:

- Language can be a drawback here as phishing occurs in all kinds of websites irrespective of their languages.
- In order to overcome some of the quering problems of the search engines the algorithms will have to be implemented as server based email filter rather than a browser tool.

IV. CONCLUSION

The frogery which has occurred due to phishing has taken a rigorous plunge in the recent years. In this paper, we have analogized the working principles of two approaches which are used for the detection of phishing Websites namely the LinkGuard and CANTINA. During this we implicited that the LinkGuard is an algorithm which differentiates on the basis of URL's and DNS and is effective upto 96% detection including the unknown websites. On the other hand the CANTINA approach is a novel designed method involving heuristics and content based methods thereby increasing 97% detection of phishing with 6% false positives. Therefore, in the future work both these approaches can act as great barriers for keeping away all kinds of phishing such that user's might not become victims to forgery or theft.

REFERENCES

[1] H.Kerstein, Paul (July 19,2005)

[2] Anti-Phishing Working Group, Phishing Activity Trends Report. 2006. http://www.antiphishing.org/reports/.

[3] Keizer, G., Phishing Costs Neraly $1 Billion, *TechWeb TechnologyNews*. http://www.techweb.com/wire/security/164902671.

[4] McMillian, R.,Gartner: Consumers to lose $2.8 billion to phishers in 2006, Networkworld,2006.

[5] U.Vidyasagar "Intelligent phishing website detection and prevention system by using LinkGuard algorithm"

[6] Phelps, T.A. and R. Wilensky, Robust Hyperlinks and Locations, *D-Lib Magazine*, vol. 6(7/8),2000. http://www.dlib.org/dlib/july00/wilensky/07wilensky.html.

