Phishing Detection via Identification of Website Identity

Ee Hung Chang*, Kang Leng Chiew†, San Nah Sze‡ and Wei King Tiong§
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
94300 Kota Samarahan
Sarawak, Malaysia
*fancy_2309@yahoo.com
†klchiew@fit.unimas.my
‡snsze@fit.unimas.my
§wktiong@fit.unimas.my

***Abstract*—In this paper, we propose an anti-phishing method to protect Internet users from the phishing attacks. The scope of our study is on the Internet phishing, particularly focusing on the detection of phishing website. In order to do that, our proposed method will render a screenshot of the webpage and segment the region of interest, which consists of the website logo. Next, we will utilize Google image database to identify the website identity based on the segmented website logo. During the identification process, we employ the content-based image retrieval mechanism provided in Google Image Search engine to locate the most similar logo from Google image database. The returned results will reveal the real identity of the website. With the real identity, we can differentiate a phishing website from the legitimate website by assessing the domain name of the query website. The conducted experiments show promising results and our findings prove that we can effectively detect a phishing website when we manage to determine the real identity of a website.**

Keywords*—*phishing detection, website identity, Google search, logo segmentation, security

I. Introduction

Online Phishing is a criminal act of deception in obtaining the sensitive information such as username, passwords, credit card detail and etc. by masquerading as trustworthy entities in electronic communication. It usually gained users credence by proclaiming they are from the legitimate party, such as popular mail services providers (Gmail, Yahoo) or financial institution (Paypal, Brandesco Bank), and then directing user to a fraudulent website to harvest users credentials.

Due to the severity of the losses, phishing was recognized as one of a fully industrialized economy crime [1]. According to the reports released by APWG, the number of phishing attempts is on the rise and accelerating. The statistics has showed the total phishing attacks have increased from 48,244 to 123,486 cases from the first half of year 2010 until the second half of year 2012 [2]. As for monetary losses, the statistics released by RSA in July shows that the estimated worldwide losses from phishing attacks alone amounted to over US$687 million during the first half of year 2012 [3].

To prevent users from the phishing attacks, many anti-phishing solutions have been proposed. Basically, there are two major approaches in anti-phishing: the store-list based (i.e., blacklist and whitelist) and the heuristic based approaches. Stored-list approach is assessing the existence of a query website (e.g., the URL of a suspicious website) to the set of entries stored in the predefined list. The list can be blacklist, whitelist or both. On the other hand, heuristic based approach is based on the mechanism of extracting some distinctive features or characteristics from the website in query to facilitate the detection and identification of a phishing website.

II. Related Works

Due to the ever increasing incidents of phishing attacks as shown in the statistics listed above, there are many different new anti-phishing methods are proposed. One of the popular methods is browser-integrated solution. Chou et al. introduced one such tool called SpoofGuard, which will looks for phishing symptoms such as obfuscated URLs in web pages and raises alerts [4]. Gabber et al. also presented a tool to protect the clients identity and password information. They defined client persona in terms of username, password and email address and introduced a function which provides a client with different persona for different servers visited [7]. Ross et al. later introduced a similar concept by a tool called PwdHash. This tool will create a domain-specific passwords which will become useless if it is submitted to a different domain [10].

Another solution has been proposed by Dhamija et al. which called Dynamic Security Skins [5]. This technique uses a shared secret image that allows a remote server to prove its identity to an user in such a way that it is easy for an user to verification but hard for attackers to spoof. However, this protocol does not provide security for situations where the user login is from a public terminal.

Fu et al. proposed a visual similarity method which uses Earth Movers Distance (EMD) to measure the webpage visual similarity in distinguishing the phishing websites [6]. They first converted the involved webpages into low resolution images and then used color and coordinate features to represent the image signatures. After that, they used EMD to calculate the signature distances of the converted images. If the EMD-based visual similarity of a webpage exceeds the threshold of a protected webpage, that page will be classified as a phishing websites. Since this method is assuming the phishing website

978-1-4799-2845-3/13/$31.00 ©2013 IEEE