COMPUTERS & SECURITY ■■ (2015) ■■–■■

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Utilisation of website logo for phishing detection

Kang Leng Chiew *, *Ee Hung Chang, San Nah Sze, Wei King Tiong*

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

ARTICLE INFO

Article history:
Received 2 February 2015
Received in revised form 4 July 2015
Accepted 31 July 2015
Available online

Keywords:
Anti-phishing
Website logo
Website identity
Google image search
Identity consistency
Logo extraction

ABSTRACT

Phishing is a security threat which combines social engineering and website spoofing techniques to deceive users into revealing confidential information. In this paper, we propose a phishing detection method to protect Internet users from the phishing attacks. In particular, given a website, our proposed method will be able to detect if it is a phishing website. We use a logo image to determine the identity consistency between the real and the portrayed identity of a website. Consistent identity indicates a legitimate website and inconsistent identity indicates a phishing website. The proposed method consists of two processes, namely logo extraction and identity verification. The first process will detect and extract the logo image from all the downloaded image resources of a webpage. In order to detect the right logo image, we utilise a machine learning technique. Based on the extracted logo image, the second process will employ the Google image search to retrieve the portrayed identity. Since the relationship between the logo and domain name is exclusive, it is reasonable to treat the domain name as the identity. Hence, a comparison between the domain name returned by Google with the one from the query website will enable us to differentiate a phishing from a legitimate website. The conducted experiments show reliable and promising results. This proves the effectiveness and feasibility of using a graphical element such as a logo to detect a phishing website.

1. Introduction

The emergence of information technology has made our life easier. For example, we no longer need to be available in front of the bank counter to make a transaction. We can do this by using a personal computer through the Internet infrastructure. The online application platform is huge and encompasses a vast variety, ranging from casual information sharing (e.g., social networking) to a more monetary intensive related application (e.g., Internet banking, bill payment, E-commerce, etc.).

The popularity of this infrastructure has attracted immoral parties to gain profit illegally. This illegal profit oriented threat is considered an online crime. One of the simple yet effective threats is a phishing attack. Usually in the phishing attack, the phisher will send a huge number of emails that impersonates as it was sent from a genuine party. Typically, the email content is crafted to create a sense of urgency, worry, or offer some great incentive and asks the victims to take action. For example, the email will urge victims to update their confidential information (e.g., login password) before his or her account is suspended. Once the victim innocently updates the confidential information, the phisher will gain all necessary details. They will utilise them for illegal access purposes, as the information the user transmits is sent to the phisher's counterfeit website rather than the genuine one.

* *Corresponding author.* Tel.: +60 82 583762/3758.
E-mail address: klchiew@fit.unimas.my (K.L. Chiew).

http://dx.doi.org/10.1016/j.cose.2015.07.006

Please cite this article in press as: Kang Leng Chiew, Ee Hung Chang, San Nah Sze, Wei King Tiong, Utilisation of website logo for phishing detection, computers & security (2015), doi: 10.1016/j.cose.2015.07.006

Phishing is a very serious security threat, and it causes a multitude of pitfalls which include identity theft, stolen money, unauthorised account access, and credit card fraud. The impact of this threat is dreadful, and causes tremendous financial losses every year. Besides the tangible losses, phishing also causes long term damage (i.e., reputation, credibility and confidence losses) to the customer–company relationship.

The main reason that makes phishing attacks possible to perpetrate is the lack of computer knowledge among Internet users. Many Internet users do not know how the web applications work. For example, they do not understand the syntax or the meaning of the Uniform Resource Locators (URLs), and cannot differentiate a fraudulent website from a legitimate one. Lack of computer knowledge also blindfolds the Internet users from utilising the security indicators, which are normally available in the Internet browser. With the advancement of web technology, it is possible for the phishing attacks to exploit visual deception. This technique ranges from manipulating textual to graphical form. For example, the phishers may choose the number one to substitute the alphabet one in http://www.paypal.com (Dhamija et al., 2006), hence creating a fraudulent website that looks similar to the legitimate one. Whereas for the graphical form, the phishers may use Javascript to load the secure padlock icon in the address bar to indicate that the website is secure and deceive the users to believe that it is a legitimate website. In addition, security is often treated unconsciously as a secondary goal by most Internet users when focusing on their primary tasks (i.e., performing online banking, transactions, etc.) (Dhamija et al., 2006).

Clearly, we can see that the attackers exploit the human factors (e.g., computer illiteracy and carelessness) as the loophole. While improving public awareness is important in fighting against phishing attacks, it is even more crucial and necessary to equip the users with a more automated security mechanism. Currently, the security mechanisms can be broadly divided into list-based and heuristic-based approaches. A list-based 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 a blacklist, a whitelist, or both. On the other hand, a heuristic-based approach is based on the mechanism of extracting some distinctive features or characteristics from the query website to facilitate the detection and identification of a phishing website. The list-based approach is fast and produces low false positives, but its effectiveness is only as good as its up-to-date list. While the heuristic-based approach comparatively takes more computational power, it is more preferable due to its flexibility to detect a new phishing website.

In this paper, we propose a method that belongs to the heuristic-based approach, and it is an extension of our work proposed in Chang et al. (2013). We claim that in order to detect a phishing website, we must first be able to determine the consistency of the website identity. With the consistency, we will be able to assess the legitimacy of a website. Among the many elements within a website, a logo is the most suitable candidate because it is the official trademark and representative of a website. As we all know, the relationship between a logo and the domain name of a website is exclusive; any mismatch is an indication of a phishing attack. In Chang et al. (2013), we had illustrated and proved that it is possible to identify the associated legitimate website from querying the logo image through a Google image search. We applied fixed segmentation with manual best-fit cropping to extract the logo. We obtained four different sizes of segmentation from a rendered website screenshot, and utilised the Google image database to identify the website identity. We performed Google image searches using the segmented images, and used the returned keywords to perform a second search by using Google text Search. The top 30 URLs from the second search results were recorded and the legitimacy of the query websites were determined based on the comparison between the domain name of a query website with the URLs from the search results.

Differing from Chang et al. (2013), in this paper, we employ image processing and machine learning techniques to locate the logo. From the logo, we utilise Google Images as a source of a knowledge database to determine the website identity. Immediate comparison between the domain name of the determined identity with the one from the query website will enable us to differentiate a phishing from a legitimate website.

While there are many heuristic-based methods in the literature, they are different from the one proposed in this paper. They are either determining the identity of a website based on textual elements or direct evaluation based on some form of phishing characteristic without knowing the identity. The former is similar to ours, but different in the context, and the main weakness is its textual semantic gap. Whereas the latter is like finding some evidence from the wild without any baseline, and it has many uncertainties. While it might be effective to detect existing phishing (i.e., with a known phishing characteristic), it is certainly not effective for an unknown phishing (zero-day phishing) attack. For example, evaluating the URL for domain name obfuscation (a phishing characteristic) will fail when a legitimate website is injected with a phishing webpage. For another example, by assessing the structure of an HTML page (e.g., DOM) for abnormality, it is difficult to judge the legitimacy of a website, simply because the phisher can have the exact clone of the website. Further discussion of the existing techniques will be given in Section 2.

The remainder of the paper is structured as follow. In the next section, we will discuss some of the related works. We will then discuss in detail the proposed method in Section 3. In Section 4, we present the experimental results. We will later give the analysis in Section 5. Section 6 concludes the paper.

2. Related work

While there exists numerous different techniques in phishing detection, they can be divided into list-based and heuristic-based approaches. According to Huh and Kim (2012), one of the popular techniques is blacklisting. Many popular web browsers are using this approach to detect phishing website (Abrams et al., 2013; Schneider et al., 2008). In this technique, a query website is checked with a list (i.e., a list of known phishing URLs), which is compiled and maintained by some consortium or organisation. If the checking returns a match, then the website will be labeled as phishing. On the contrary, instead of maintaining the blacklist, one can compile a list of legitimate URLs. This technique is known as whitelisting, and it is also a type of list-based approach. An example of a whitelisting

Please cite this article in press as: Kang Leng Chiew, Ee Hung Chang, San Nah Sze, Wei King Tiong, Utilisation of website logo for phishing detection, computers & security (2015), doi: 10.1016/j.cose.2015.07.006