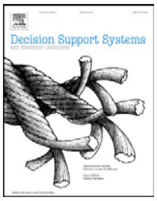
Contents lists available at ScienceDirect

Decision Support Systems

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



PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder

Choon Lin Tan[a], Kang Leng Chiew[a,*], KokSheik Wong[b], San Nah Sze[a]

[a]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[b]Faculty of Computer Science and Information Technology,University of Malaya, 50603 Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 20 October 2015
Received in revised form 28 March 2016
Accepted 22 May 2016
Available online 1 June 2016

Keywords:
Phishing detection
Identity keywords
N-gram
Weighted URL tokens
Search engine

ABSTRACT

This paper proposes a phishing detection technique based on the difference between the target and actual identities of a webpage. The proposed phishing detection approach, called PhishWHO, can be divided into three phases. The first phase extracts identity keywords from the textual contents of the website, where a novel weighted URL tokens system based on the N-gram model is proposed. The second phase finds the target domain name by using a search engine, and the target domain name is selected based on identity-relevant features. In the final phase, a 3-tier identity matching system is proposed to determine the legitimacy of the query webpage. The overall experimental results suggest that the proposed system outperforms the conventional phishing detection methods considered.

1. Introduction

In this modern age of information technology, consumers are dealing with more products and services through the online channel. Therefore, having multiple online accounts (e.g., email account, banking account, social networking account) have become a norm for most people. This technological trend is exposing internet users to a rising threat of online identity theft known as phishing [17].

Phishing websites are counterfeit websites designed to deceive victims and steal their account login credentials, credit card numbers or other personal secrets. Phishers usually entice victims to the phishing website by sending emails containing the fraudulent URL and some threatening messages such as possible account termination, and fake alert on illegal transaction [9]. At the phishing website, the phishers will capture sensitive information submitted by the victims.

The severity of phishing threats in recent years continues to escalate, based on statistics gathered from security organizations. For instance, a total of 42,212 unique phishing websites was reported in June 2014 by the Anti-Phishing Working Group [2], whereas the financial loss inflicted upon worldwide organization in December 2014 was estimated to be $453 million [10]. These alarming trends have resulted in the loss of consumers' trust in using E-commerce websites because they are feared to become fraud victims [6]. In summary, phishing attacks have resulted in widespread leakage of sensitive information, monetary loss and crippled businesses' reputation.

The key factor that makes phishing possible is the human behaviour when interacting with electronic communication channels. Dhamija et al. [8] identified several user tendencies that are exploited by phishing attacks. For instance, a typical user is often unaware of the significance of common security indicators such as the Secure Sockets Layer (SSL) icon and digital certificate on the browser address bar. As a result, these useful indicators are often ignored. In addition, some users are confused on how a legitimate URL is supposed to resemble, thus they rely on the webpage contents to determine its genuineness [18]. A recent assessment by Alsharnouby et al. [1] reveals that participants with phishing awareness can only achieve 53% of average success rate in identifying phishing websites. These studies have proven that both normal and technical users can be easily deceived by phishing webpages. Hence, it is crucial to have an efficient phishing detection system, where users can be effectively safeguarded from phishing attacks.

To compensate for the human limitations in detecting phishing websites, automated solutions have been introduced in conventional web browsers and security applications. Most solutions rely on blacklists (e.g., Google Safe Browsing list, PhishTank list) that

* Corresponding author. Tel.: +6082 58 3762.
E-mail addresses: colin89lin@gmail.com (C. Tan), klchiew@unimas.my (K. Chiew), koksheik@um.edu.my (K. Wong), snsze@unimas.my (S. Sze).

http://dx.doi.org/10.1016/j.dss.2016.05.005
0167-9236/© 2016 Elsevier B.V. All rights reserved.

are compiled from automated link analysis or manual submission by volunteers. Zhang et al. [24] and Sheng et al. [21] tested several browser-based phishing detection tools, and concluded that most tools fail to detect phishing websites that are yet to be added into the blacklist or otherwise known as zero-day phishing websites. Recently, Purkait [19] tested the latest commercial browsers and security applications, where 12 out of 14 tools have failed to detect any of the phishing websites that existed for only a few minutes old. In short, the related studies in [19, 21, 24] have raised critical concerns on the reliability of the conventional tools. Furthermore, the existing tools also face challenges from increasingly sophisticated phishing threats, such as phishing webpages injected into existing legitimate websites and cloning of legitimate webpages using phishing toolkit [3]. Despite the existence of conventional phishing detection tools, the general public remains exposed to high risk of becoming phishing victims.

The detection of phishing webpages is extremely important, since it is the core mechanism for the mitigation of phishing attacks. When the detection phase is functional and accurate, appropriate warnings can be issued and actions can be applied to protect the victim or minimise the effects of the phishing attack. As such, we propose PhishWHO in this paper, which is an extension of our proposed phishing detection system in [22]. PhishWHO is based on a permanent phishing characteristic that stays intact over time, namely the discrepancies between the target and the actual identities of the query webpage. Here, target identity is defined as the domain name belonging to a legitimate brand that the phishing webpage deceptively represents, while actual identity refers to the query webpage's domain name. For legitimate webpages, the target identity often point to its own domain name, while phishing webpage does not. As such, the proposed method checks whether the query webpage is promoting itself, or promoting another existing legitimate webpage. By applying the proposed information processing techniques, both the target identity and the actual identity can be systematically derived from the query webpage. Hereinafter, the term "identity" and "domain name" shall be used interchangeably.

Several other enhancements are also incorporated in this paper, which distinguish it from our previous work in [22]. Our contributions include: (a) Exploiting the differences between the target and actual identities of a webpage to detect zero-day phishing webpages; (b) proposing a novel weighted URL tokens system based on N-gram model to overcome language limitations in phishing webpage detection; (c) exploiting indirect identity relationships to reduce false positives, and; (d) offering long-term effectiveness by leveraging on permanent phishing characteristic.

The remainder of this paper is organised as follows: Section 2 briefly reviews the scholarly works related to this research. Section 3 puts forward the proposed method. Section 4 describes the experiment setup and summarises the results. Section 5 discusses the merits and limitations of the proposed method. Finally, Section 6 concludes this paper.

2. Related works

A broad range of automated anti-phishing techniques have been introduced over the years. This section reviews the conventional anti-phishing techniques available.

2.1. Text-based detection

Zhang et al. [25] proposed CANTINA, a text-based phishing detection technique that extracts keywords from a webpage using the term frequency-inverse document frequency (TF-IDF) algorithm. The keywords are searched on Google, where the query webpage will be classified as legitimate if its domain name exists among the search results. Enhancements to CANTINA are proposed in [13], including an improved HTML parsing method and text-handling. The main problem is the reliance of TF-IDF on language-specific word list, thus Zhang et al. [25] and Komiyama et al. [13] are ineffective in classifying non-English webpages. Similar language limitation is found in [20].

2.2. Visual-based detection

Fu et al. [11] proposed using the Earth Mover's Distance (EMD) to assess visual similarities between suspected webpages and legitimate webpages. To calculate visual similarities, the webpages are sampled into low resolution images and represented by image features, i.e., dominant colour category and corresponding centroid coordinate. However, phishers can evade their method by making phishing webpages that appears less similar with the legitimate webpages. In another work, the same EMD algorithm is combined with a text classifier based on the naive Bayes rules to detect phishing webpages [23]. The bottleneck in visual-based techniques is the need for a large image database of legitimate websites. Since it is costly to maintain an up-to-date image database, visual-based techniques are impractical for widespread adoption. Chiew et al. [7] attempts to address this weakness by proposing an approach to extract logo from the webpage and submit it to Google image search, where the target identity can be determined.

2.3. Feature-based detection

Several researchers have shown that URL features are viable in determining whether a website is a phishing website. Le et al. [14] proposed an approach based on URL features such as length of full URL, domain name, directory, file and the number of symbols. These features are then fed into a classifier to compute the phishing probability. Another subset of feature-based techniques detects phishing based on search result features. For instance, Huh and Kim [12] queried Google, Yahoo, and Bing with the query webpage URL to obtain the number of search results and website ranking. These features are forwarded to a K-Nearest Neighbour (KNN) based classifier to compute the legitimacy of a webpage. Although feature-based techniques has the advantage of detecting zero-day phishing webpages, they often suffer from high false positive rates.

2.4. Identity-based detection

Liu et al. [15] focus on webpage identity analysis to detect phishing, where the Semantic Link Network (SLN) is proposed. SLN consists of weighted paths linking a set of webpages associated with the query webpage. Several metrics of the SLN are calculated to find the target identity and determine the legitimacy of the query webpage. Their proposed method are further improved in [16], resulting in the replacement of the SLN module by a parasitic community module. Motivated by [16], a similar strategy is proposed by [20] to identify the target domain name. However, [15, 16, 20] may fail to find the target domain name on phishing webpages that do not contain any URL belonging to the targeted legitimate website, thus resulting in false negative detections. In other words, phishers can easily bypass these techniques by hosting all the webpage resources (e.g., images, JavaScript, CSS) on the same phishing website.

2.5. Prevention-based techniques

Prevention-based techniques such as password management tools and 2-factor authentication focus on distorting the phishing infrastructure to protect the user's credentials [5]. For instance, a password management tool may add IP address string to the password whenever the user creates a new account at a legitimate website. When the user tries to log into a phishing website, a wrong