Expert Systems With Applications 220 (2023) 119723

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Hybrid phishing detection using joint visual and textual identity

Colin Choon Lin Tan [a], Kang Leng Chiew [b,*], Kelvin S.C. Yong [a], Yakub Sebastian [c], Joel Chia Ming Than [a], Wei King Tiong [b]

[a] Faculty of Engineering, Computing and Science, Swinburne University of Technology, Sarawak Campus, Jalan Simpang Tiga, 93350 Kuching, Sarawak, Malaysia
[b] Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[c] College of Engineering, IT and Environment, Charles Darwin University, Ellengowan Drive, Casuarina, NT 0810, Australia

ARTICLE INFO

Keywords:
Phishing detection
Web security
Website identity
Computer vision
Logo detection
Brand names

ABSTRACT

In recent years, phishing attacks have evolved considerably, causing existing adversarial features that were widely utilised for detecting phishing websites to become less discriminative. These developments have fuelled growing interests among security researchers towards an anti-phishing strategy known as the identity-based detection technique. Identity-based detection techniques have consistently achieved high true positive rates in a rapidly changing phishing landscape, owing to its capitalisation on fundamental brand identity relations that are inherent in most legitimate webpages. However, existing identity-based techniques often suffer higher false positive rates due to complexities and challenges in establishing the webpage's brand identity. To close the existing performance gap, this paper proposes a new hybrid identity-based phishing detection technique that leverages webpage visual and textual identity. Extending earlier anti-phishing work based on the website logo as visual identity, our method incorporates novel image features that mimic human vision to enhance the logo detection accuracy. The proposed hybrid technique integrates the visual identity with a textual identity, namely, brand-specific keywords derived from the webpage content using textual analysis methods. We empirically demonstrated on multiple benchmark datasets that this joint visual-textual identity detection approach significantly improves phishing detection performance with an overall accuracy of 98.6%. Benchmarking results against an existing technique showed comparable true positive rates and a reduction of up to 3.4% in false positive rates, thus affirming our objective of reducing the misclassification of legitimate webpages without sacrificing the phishing detection performance. The proposed hybrid identity-based technique is proven to be a significant and practical contribution that will enrich the anti-phishing community with improved defence strategies against rapidly evolving phishing schemes.

1. Introduction

Phishing is any criminal act of using genuine-looking websites to deceive users into disclosing confidential information such as account passwords, credit card numbers, etc. A typical phishing website is constructed by replicating visual cues and textual content of an existing legitimate website (Aleroud & Zhou, 2017; Chiew, Yong, & Tan, 2018). Recent statistics indicated that phishing attacks are still blooming, where an average of 54,924 unique phishing websites were reported monthly between January and March 2020 (Anti-Phishing Working Group, 2020).

Studies have shown that manually recognising phishing websites can be difficult even for users who previously underwent anti-phishing education training, therefore users should not be expected to perform this risky task on their own (Alsharnouby et al., 2015; Arachchilage et al., 2016; Ubing et al., 2019). As such, blacklist-based automated detection mechanisms have been built into mainstream browsers to assist in warning users when known phishing websites are loaded, while unknown or zero-day phishing websites are still able to slip past the detection. Despite the limitations of blacklist-based detection, they remain commercially successful owing to its lightweight computational overhead and low rate of misclassifying legitimate websites (da Silva et al., 2020; Gupta et al., 2017; Yang et al., 2021). Therefore, in designing practical phishing detection solutions, achieving a low rate of misclassifying legitimate websites is an important requirement (Gupta et al., 2017). Anti-phishing researchers have been focusing more on enhancing the phishing detection rate, while adverse issues that impact the detection of legitimate webpages remain largely unattended. To address this performance gap, we attempt to study the possible

* Corresponding author.
E-mail addresses: ctan@swinburne.edu.my (C.C.L. Tan), klchiew@unimas.my (K.L. Chiew), kscyong@swinburne.edu.my (K.S.C. Yong), yakub.sebastian@cdu.edu.au (Y. Sebastian), jcmthan@swinburne.edu.my (J.C.M. Than), wktiong@unimas.my (W.K. Tiong).

https://doi.org/10.1016/j.eswa.2023.119723
Received 20 July 2022; Received in revised form 8 December 2022; Accepted 19 February 2023
Available online 23 February 2023

improvement of one particular type of phishing detection technique called the *identity-based technique*.

Identity-based phishing detection techniques ascertain whether the portrayed identity (e.g., brand name, logo) presented in the webpage content corresponds to the actual page identity observed through the URL (Jain & Gupta, 2021). This is typically done by querying the search engine using the website logo or brand keywords to retrieve search results that can pinpoint the brand's original website. Since identity-based techniques do not depend on dynamically evolving adversarial patterns in the phishing webpage, it has consistently achieved superior performance in terms of phishing identification (i.e., high true positive rate) (Rao & Pais, 2019).

Nevertheless, identity-based techniques are more susceptible to false positives due to complications and challenges in extracting the website's portrayed identity and obtaining relevant search results for identity verification (Jain & Gupta, 2018; Rao & Pais, 2019; Van Dooremaal et al., 2021). Due to the diverse presentation styles and visual layout of webpages, it is challenging to derive efficient techniques that can reliably detect and extract previously unseen graphical brand elements such as the website logo (Gupta & Jain, 2020). Although state of the art machine learning techniques such as deep learning can be used to detect the website logo (Bozkir & Aydos, 2020), its detection capability is primarily limited to previously trained logos, thus making it less practical for phishing detection applications. Meanwhile, other techniques that strive to detect previously unseen logos in webpages are lacking the ability to model discriminative features utilised by human vision, thus limiting the logo detection accuracy (Chiew et al., 2015). On the other hand, existing identity-based techniques relying solely on textual identity may not yield accurate brand keywords as the portrayed identity when employed against webpages with limited textual content or contain mainly non-ASCII text (Tan et al., 2016).

In this paper, we advance the state-of-the-art of identity-based phishing detection techniques by enhancing the detection and extraction accuracy of the portrayed identities, as well as integrating visual and textual identities in a hybrid framework to improve the efficacy of website identity verification. Through a series of experiments, we empirically demonstrated that our method promises greater practicality and robustness than singular identity-based techniques. The main contributions from this research include the following:

(a) Novel features that mimic human vision are proposed to improve the detection and extraction of the website logo, which is vital for phishing detection techniques utilising visual identities. For example, one of the proposed features is colourfulness, which has yet to be capitalised in existing techniques (Chiew et al., 2015). We argue that colourfulness is an essential consideration that guides humans in performing logo identification in webpages and, hence must be included in any computer vision-based system for logo detection (Van Dooremaal et al., 2021).

(b) The proposed technique advances the emerging body of knowledge on the hybridisation of visual and textual identities for phishing detection. As evidenced in Van Dooremaal et al. (2021), the hybridisation of visual and textual identities is one of the emerging research directions aimed at overcoming the existing performance limitations. Therefore, our work establishes the foundation for other anti-phishing researchers to further develop improved methods based on hybrid identities.

(c) The proposed hybrid identity-based technique achieves a lower rate of misclassifying legitimate webpages without sacrificing the high detection rate for phishing webpages. This performance improvement is attributed to the proposed technique's capability in adapting to a wider range of websites with varying design quality which can complicate the extraction of brand identity elements such as website logos and keywords. To achieve such level of robustness, the proposed technique integrates complementary visual and textual identity discovery components that work hand-in-hand to establish the webpage identity and achieve lower false positives.

The remainder of this paper is organised as follows: Section 2 reviews and discusses identity-based phishing detection techniques. Section 3 puts forward the proposed method. In Section 4, we describe the experimental setup, present the results, and discuss our research findings. Finally, Section 5 concludes the paper and suggests some future research directions.

2. Related work

Based on the scope of our research, the literature presented in this section focuses on identity-based phishing detection techniques. Identity-based techniques for phishing detection can be accomplished using textual or visual brand identities. We first discuss techniques that utilise textual brand identities, followed by visual brand identities and a combination of both.

In regular web browsing, users that intend to visit a particular website but are unsure of its actual URL will normally engage the help of a search engine. Tan et al. (2016) leveraged the very same concept to find the target domain by querying the search engine with identity keywords. A weighted URL tokens system was proposed, which utilises the structure of URLs in a webpage to extract identity keywords in the form of single keywords (uni-gram) or multiple coexisting keywords (multi-gram). These keywords are then searched using a search engine, where the target domain is derived from the search results. Based on experimental results, the proposed technique achieved competitive true positive and true negative rates of 99.68% and 92.52%, respectively. Although the proposed technique can also work on non-English webpages, it is still unable to accommodate webpage textual content written in non-ASCII languages.

Rao and Pais (2019) assume the domain name and page title as textual brand descriptors. They proposed an identity-based method by formulating dynamic search queries using the domain name and page title before feeding them to the search engine. If the potential target website is not found in the search results, the suspicious webpage will be declared as phishing. Otherwise, a similarity measure between the suspicious webpage and the potential target website will be used to determine the page's legitimacy. The proposed method achieved a true positive rate of 97.77% and a promising true negative rate of 99.36%. In an earlier technique by Jain and Gupta (2018), the identity verification process was also accomplished similarly. However, false negative detection may occur if the phishing webpage utilises mostly internal hyperlinks that point to the same phishing domain.

In the work of Peng et al. (2019), the behaviour and properties of phishing websites were studied in-depth. One of their analysis involved finding the target brand for 1500 samples of phishing webpages. Their analysis is carried out by extracting visible text from the webpage screenshot using optical character recognition (OCR) techniques. The less important terms were then filtered from the extracted text using a text-mining algorithm called RAKE (Rapid Automatic Keyword Extraction). Lastly, the remaining terms were searched on Google, and the top result is taken as the target brand. The authors observed that the target for phishing webpages targeting popular brands such as PayPal and Microsoft can be resolved with accuracy up to 99.8%, while the accuracy dropped to 88% for less popular brands. If actual legitimate webpage samples are tested using this technique, the less popular brands will suffer misclassifications as well. As such, more studies in the area of website identity verification is desirable to close the current performance gap.

A more recent identity-based anti-phishing approach came from Liu and Fu (2020). Given a query webpage, hyperlinks were obtained from the HTML source and used to retrieve the corresponding linked webpages. Based on this initial pool of webpages, the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm was employed to extract keywords, which were then submitted to Google Search to retrieve webpages listed in the top-10 search results. This expanded set of webpages and their corresponding page linking structure were processed