Phishing Attack Detection using a Search Engine and Heuristics-based Technique

Brij B. Gupta, National Institute of Technology, India
Ankit Kumar Jain, National Institute of Technology, India
https://orcid.org/0000-0002-9482-6991

ABSTRACT
The language used in the textual content of the webpage is the barrier in most of the existing anti-phishing methods. Most of the existing anti-phishing methods can identify the fake webpages written in the English language only. Therefore, we present a search engine-based method in this article, which identifies phishing webpages accurately regardless of the textual language used within the webpage. The proposed search engine-based method uses a lightweight, consistent and language independent search query to detect the legality of the suspicious URL. We have also integrated five heuristics with the search engine-based mechanism to improve the detection accuracy, as some newly created legitimate sites may not appear in the search engine. The proposed method can also correctly classify the newly created legitimate sites that are not classified by available search engine-based methods. Evaluation results show that our method outperforms the available search-based techniques and achieves 98.15% TPR of and only 0.05% FPR.

KEYWORDS
Google, Heuristic, Login Form, Phishing Attack, Search Engine

1. INTRODUCTION
Nowadays, the Internet has rapidly grown and become a critical resource for a variety of services such as WWW, e-mail, e-commerce, etc. As the Internet continues to grow, more and more organizations and industries are integrating their existing system with the Internet to increase the revenue. Our daily lives have also become more reliant on Internet-based services for delivery of data services. Moreover, these services store user’s sensitive and personal credentials, and attract cybercriminal for stealing of user’s credential (Gupta et al., 2017; Pejic-Bach, 2010). These events have prompted many cybercriminals who try to take advantage of easy access to the WWW and associated user-centric approaches (Konradt et al., 2016).

Phishing is one of the most dangerous cybercrimes in which attackers steal sensitive credentials from the Internet users such as, bank account details, credit card number, protected password, etc. Phishing is similar to the word ‘fishing’. In fishing, we use bait in order to catch a fish same thing happens in phishing but instead, victim is a normal user using the Internet. The victim of this attack thinks that information is going to the trustworthy party, instead, it is going to the cybercriminals. Cybercriminals use this information for its own advantage not for its intended purpose. Phishing is one of the major problems faced by cyberspace. The statistics of most recent web vulnerability reports revealed that phishing attack stands at peak position among all other cyber-attacks (APWG, 2016).
In this attack, the cybercriminals exploit the vulnerability of innocent Internet users, as many users are not aware of phishing attack (Arachchilagea et al., 2016). Most of the Internet users avoid the warning messages (i.e. security indicators) shown by the web browsers. Moreover, users also do not know the secure sockets layer (SSL) certificate and they are not able to differentiate between genuine and fake uniform resource locators (URLs). The visual deception of the fake website also plays an important factor to catch innocent victims. The appearance of fake website deceives users to believe that they are visiting the correct website. Phishing messages are spread over e-mail, SMS, instant messengers, social networking sites, etc, but e-mail is the popular way to perform this attack.

The motive behind phishing is not only stealing the user’s information, but to install other types of malware, like, Trojan horse, ransomware, etc., in the victim machine (Phishingpro, 2018). Moreover, this brutal attack also creates the negative business impact on electronic commerce, payment gateways, and social media websites (Konradt et al., 2016). Google raised 10 million phishing warning messages every day to their users who wish to visit a fake website and added 10000+ websites in its blacklist (Orman, 2013). According to Anti-Phishing Working Group (APWG) cyber report, 662795 unique phishing websites were detected in 2017 (APWG, 2018). Cybercriminals stole a total of $172bn from consumers in 2017 (Norton Cyber Security, 2017). Moreover, it was found that 76% of organisations had experienced phishing attacks in 2017 (State of the Phish, 2018). Furthermore, 53% of InfoSec professionals reported that they had experienced spear phishing attacks in 2017. (State of the Phish, 2018)

In the recent years, researchers have developed various anti-phishing techniques (Aleroud & Zhou, 2017). However, the detection of phishing attacks still lacks a complete solution, and the construction of efficient phishing detection method has become a challenging task. Various existing methods are unable to detect newly created malicious webpages, and the detection rate of anti-phishing methods is also low (Gupta et al., 2017). The existing anti-phishing methods are not effective due to the small lifetime of a phishing webpage. Most of the existing techniques are language dependent and are only useful for the detection of language specific webpages (Marchal et al., 2017). Moreover, the attackers may also use the embedded objects (i.e. images, scripts, Flash, ActiveX, Java Applet, etc.) in the fake webpages to bypass phishing detection techniques.

Various automated phishing detection tools also developed in the past few years (e.g., GeoTrust, McAfee SiteAdvisor, Netscape Toolbar, Web of Trust extension, etc.). However, no browser toolbar can successfully stop all phishing attacks. The browser security toolbar related studies show that there are critical concerns about the trust of the conventional tools (Purkait, 2014).

This paper presents search engine-based method that uses a lightweight, and consistent query to check the legitimacy of the webpage. In the search engine, legitimate websites appear in top search results whereas malicious webpage does not present in top search results due to the short lifetime of phishing websites. Since search engine takes the considerable time to appear the websites in top ‘n’ search results due to which these methods may misclassify the newly created legitimate webpages (Jain & Gupta, 2017). Therefore, we have also integrated five heuristics with the search-based mechanism to improve the detection accuracy. Proposed method verifies that a genuine webpage should not be classified as malicious. Experimental results illustrate that the proposed method has high classification accuracy, low computation, and delivers rapid response time. Moreover, our method outperforms the existing search based anti-phishing methods and achieves 98.15% TTR and 0.05% FPR.

The rest of article is organized as follows. Section 2 presents the related work. We describe the research objective of our method in section 3. Section 4 provides the detailed overview of our proposed anti-phishing method. Section 5 provides the details of the identity features associated with a webpage. Section 6 describes the experimental outcomes of our technique. Section 7 presents the limitation of proposed method. Finally, section 8 concludes the paper.
Related Content

Questioning the Key Techniques Underlying the Iterative and Incremental Approach to Information Systems Development
[www.igi-global.com/article/questioning-key-techniques-underlying-iterative/40337?camid=4v1a](www.igi-global.com/article/questioning-key-techniques-underlying-iterative/40337?camid=4v1a)

Wheelchairs Embedded Control System Design for Secure Navigation with RF Signal Triangulation
[www.igi-global.com/article/wheelchairs-embedded-control-system-design-for-secure-navigation-with-rf-signal-triangulation/86273?camid=4v1a](www.igi-global.com/article/wheelchairs-embedded-control-system-design-for-secure-navigation-with-rf-signal-triangulation/86273?camid=4v1a)
A Systemic Framework for Business Process Modeling Combining Soft Systems Methodology and UML
www.igi-global.com/article/systemic-framework-business-process-modeling/1320?camid=4v1a

See-Through-Sound: Transforming Images into Sonic Representations to Help the Blind
www.igi-global.com/article/see-through-sound/111252?camid=4v1a