Journal of Information Security and Applications 35 (2017) 75–84

Contents lists available at ScienceDirect

Journal of Information Security and Applications

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

Identification of phishing webpages and its target domains by analyzing the feign relationship

Gowtham Ramesh*, Jithendranath Gupta, P.G. Gamya

Dept. of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India

ARTICLE INFO

Article history:

Keywords:
Phishing
Anti-phishing framework
E-commerce security
Target Domain Identification

ABSTRACT

Phishing is the act of stealing personal information from the online users by impersonating as a statutory source in the cyberspace. Phishers often bait online users to visit their forged webpages to acquire users sensitive information. Most of the anti-phishing techniques today, endeavor to identify the legitimacy of the webpages the user visits and warn them with a phishing label when the webpage is a phish. But, these warnings generated by the anti-phishing tools are generic and does not provide any assistance for the users to safely navigate to the legitimate webpages. Any anti-phishing technique will be incomplete and incompetent without having a victimized domain identification in place. The method proposed in this paper addresses this lacuna by automatically identifying the victimized domain (target domain) of every successfully distinguished phishing webpage. This method initially identifies the possible target domains of the webpage by analyzing the feign relationships which exist between the webpage and its associated domains through the in-degree link associations. Further, a novel Target Validation (TVD) algorithm is used to ensure the correctness of the identified target domain which in turn helps in reducing the false target predictions of the system. The legitimacy of the webpage is further confirmed using the identified target domain. The experiment results show that this method is efficient in protecting users from the online identity attacks and also in identifying victimized domain with over 99% accuracy.

1. Introduction

Electronic spamming involves sending unsolicited bulk messages to users using electronic messaging systems. This minor annoyance has indiscriminately evolved into a new and dangerous form known as phishing where phishers try to lure internet users into revealing their personal and financially sensitive information such as credit card numbers, usernames, and passwords. Criminals spam thousands of online users with social engineered emails along with links to phished webpages. The phishing webpages imitate the look and feel of the popular legitimate websites to deceive the users into revealing their sensitive information [1].

The phishing attack has grown tremendously and has evolved in its strategies and targets over the period of time. Phishing attack in 2016 primarily targeted five industries which include financial industries, cloud storage services, webmail services, payment services, and e-commerce companies. The attack volume on these targets is increased by an average of 33% in 2016 compared to 2015. Similarly, the phishing attack has increased over 300% against government tax departments since 2014. Most of the phishing websites launched today are created using phishing kits. This makes it easier for the attackers to create fake webpages with little technical knowledge. According to PhishLab's research report, more than 29,000 phish kits are sold in the black market. Most of these phish kits create phishing webpages with techniques to evade the simple phishing detection methods. These kits are the primary reason for the bountiful increase in phishing webpages. More than 170,000 unique phishing domains are identified to have hosted phishing webpages in 2016 which are 23% higher than 2015 [2]. These attacks not only lead to financial losses of the organizations and individuals but also indirectly undermines the reputation of the organizations and trust on the Internet. This statistic shows the need for robust approaches to restrain and prevent such increasing phishing attacks [3].

Various countermeasures have been developed to mitigate the inimical effects of the phishing attack which are commonly categorized into technical and non-technical solutions. The non-technical solutions mostly emphasize on educating the novice online users through awareness programs, training, and workshops to correctly identify the phishing emails and websites [4]. The policies adopted by the Law Enforcement Agencies (LEA) against phishing attacks includes restrictions on the domain registrations and issue of punishments such as imprisonment and fines [5,6]. The technical

* Corresponding author.
E-mail address: r_gowtham@cb.amrita.edu (G. Ramesh).

http://dx.doi.org/10.1016/j.jisa.2017.06.001
2214-2126/

solutions are developed with the intention of better classification and also, to overcome the human flaws or ignorance in the view of detecting the phishing webpages. This is an important alternate to the non-technical solutions as it is not as expensive as compared to the human training and also due to the feasibility of implementation at all the times. The technical solutions are typically grouped into black-list and white-list based approaches, heuristics based approaches, visual similarity approaches and multifaceted methods. The black-list and white-list based methods maintain the list of URLs either locally or globally. These methods are inefficient in protecting users from zero-day phishing attacks mainly because 47% to 83% of phishing URL were blacklisted after 12 h [7], on the other hand, all the webpages which are not in the whitelist are irrationally labeled as suspicious. The heuristics based approaches detect the phishing webpages based on the set of characteristics present in it. This method is efficient in protecting users from a zero-day phishing attack, but are merely subjected to the presence of characteristics considered in the webpage. Visual similarity approaches identify phishing webpages by analyzing a set of visual features extracted from suspicious webpages [8]. These approaches require a robust method to retrieve website's visual content, any distortion in retrieving the content of the webpage leads to misclassification. The multifaceted approaches detect phishing webpages using any of the techniques or combinations of techniques in computational informatics. These methods commonly apply techniques like link relationship, ranking relationship, and text similarity relationship on the suspicious webpages to confirm its legitimacy. The anti-phishing techniques developed using multifaceted approaches guarantees relatively reliable results compared to any other anti-phishing techniques [9,10].

Most of the anti-phishing methods today primarily attempt to identify the legitimacy of the suspicious webpages, but lack techniques to identify the victimized legitimate webpage that the phishing webpages mimic, where, the legitimate webpage is referred to as the phishing target [11]. However, any anti-phishing technique would be incomplete without identification of the phishing target, as it plays a crucial role in assisting the users to safely navigate to the legitimate webpages. At times, when the phishers attack less popular or newly created webpages it becomes tough to find the target webpage. Also when phishers use masquerading techniques, detection of the target webpage becomes a challenge. Masquerading techniques include creating a webpage using only embedded objects like images and scripts without using any content that could provide us a clue.

The method proposed in this paper detects phishing webpages and its target domain efficiently by working on all the anticipated lacunas. Moreover, this method identifies legitimacy of the suspicious webpages without depending completely on the external information repositories such as search engines, and other third party data sources. Here, we take the suspicious webpage under scrutiny; visit its links up to level two to check for the possible number of domains that can be reached. This domain count value determines the method to be followed in generating the Target Domain Set. We then formulate a cost matrix based on the relationships that exist between the domains in the Target Domain Set. This cost matrix in turn exposes the strength of feigning relationships which exist between the domains in the Target Domain Set and webpage the user visits. The domain with higher in degree feign relationship will be considered as a target domain. The target domain is further validated using Target Validation (TVD) algorithm to ensure its correctness. Finally, the legitimacy of the webpage will be determined by comparing it with the confirmed target domain. Thus, as the content of the suspicious webpage is the only subject on which our proposed methodology is built on, neither prior knowledge about the site is required nor does it require the training data.

An overview of literature review and related work presented in Section 2. Section 3 covers the architecture of the overall system. In Section 4, we have explained the Target Domain Identification module of our system. Section 5 explains the Target Domain validation and Phishing detection methods. The implementation details, evaluation methodology, experimental results and the limitations of the proposed work are discussed in Section 6. Finally, conclusions are presented in Section 7.

2. Related work

In the recent years, many countermeasures have been developed to overcome the phishing attacks. With the development of phishing techniques over the years, meticulous efforts have been taken in the quest to find an efficient method to determine the legitimacy of the suspicious webpages and forewarn the users about the phishing attack. The current anti-phishing approaches depict many drawbacks that need to be addressed are highlighted in this section. This motivated us to propose an anti-phishing method which attempts to overcome some of the limitations of the existing schemes.

2.1. Whitelist based anti-phishing approaches

The whitelist based anti-phishing approaches maintain a list of safe websites along with necessary information. Any website which is not a part of the whitelist is considered as suspicious and such a website is further scrutinized to detect its legitimacy.

Han et al. [12] developed a whitelist based approach which records the well-known legitimate websites of the user rather than maintaining a universal legitimate sites list. In this approach, every URL which the user visits is recorded along with its LUI (Login User Interface) information and the IP addresses. The users are warned when the account information submitted to the website does not match with the corresponding details that are present in the whitelist. But, this method identifies every legitimate webpage as suspicious when the user visits the page for the first time.

2.2. Blacklist based approaches

In contrast to whitelist based approaches, blacklist approaches maintain a list of known phishing sites along with their corresponding necessary details. The blacklist entries are typically compiled from multiple data sources which include spam traps, user posts, or verified phishes compiled by third parties such as takedown vendors.

Prakash et al. [13] developed a system PhishNet which uses the approximate matching algorithm to check if the URL of the suspicious webpage is in the blacklist maintained. Along with this, five heuristics were also proposed to identify new phishing URLs from entries in the blacklist.

Zhang et al. [14] proposed a system which yields customized blacklists to individuals by using relevance ranking scheme and severity score generated for the user. The ranking scheme uses attackers history and users recent log data to measure how closely they are related, and this value is fused with the severity metric to construct the individualized blacklist. But these blacklist based approaches needs frequent updates from their sources and the rapid growth of the list demands need of massive system resources.

2.3. Heuristic-based approaches

The heuristic-based approaches decide the legitimacy of a suspicious webpage by analyzing the webpage content and using the information from the external and internal repositories.