Available online at www.sciencedirect.com

ScienceDirect

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

Computers & Security

A comprehensive and efficacious architecture for detecting phishing webpages

R. Gowtham [a,*], Ilango Krishnamurthi [b]

[a] Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Ettimadai, Coimbatore, Tamilnadu, India

[b] Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Kuniamuthur, Coimbatore, Tamilnadu, India

ARTICLE INFO

Article history:
Received 21 October 2012
Received in revised form
15 October 2013
Accepted 31 October 2013

Keywords:
Phishing
Anti-phishing
Anti-phishing framework
E-commerce security
Machine learning

ABSTRACT

Phishing is a web-based criminal act. Phishing sites lure sensitive information from naive online users by camouflaging themselves as trustworthy entities. Phishing is considered an annoying threat in the field of electronic commerce. Due to the short lifespan of phishing webpages and the rapid advancement of phishing techniques, maintaining blacklists, white-lists or employing solely heuristics-based approaches are not particularly effective. The impact of phishing can be largely mitigated by adopting a suitable combination of all these techniques. In this study, the characteristics of legitimate and phishing webpages were investigated in depth, and based on this analysis, we proposed heuristics to extract 15 features from such webpages. These heuristic results were fed as an input to a trained machine learning algorithm to detect phishing sites. Before applying heuristics to the webpages, we used two preliminary screening modules in this system. The first module, the preapproved site identifier, checks webpages against a private white-list maintained by the user, and the second module, the Login Form Finder, classifies webpages as legitimate when there are no login forms present. These modules help to reduce superfluous computation in the system and in addition reducing the rate of false positives without compromising on the false negatives. By using all of these modules, we are able to classify webpages with 99.8% precision and a 0.4% of false positive rate. The experimental results indicate that this method is efficient for protecting users from online identity attacks.

1. Introduction

Victims of phishing scams often find their personal or financial information, such as their credit card numbers, health information, email address, login credentials, answers to security questions and other sensitive data, stolen. Once this information is acquired, these details can be used to create fake accounts in the victim's name that can have a severe impact on their credit ratings or even prevent the users from accessing their own accounts, leading to a lack of financial credibility. According to the RSA's online fraud report, there was a 19% increase in phishing attacks in the first half of 2012 compared to the second half of 2011. Additionally, the RSA

* *Corresponding author*. Tel.: +91 9842032323 (mobile).
E-mail address: rameshgowtham@gmail.com (R. Gowtham).

0167-4048/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cose.2013.10.004

estimates that more than $687,000,000 was lost by global organisations due to phishing attacks in the same period (RSA Anti-Fraud Command Centre, 2012).

Statistics from the Anti-Phishing Working Group (2012) shows that banks are the primary targets of phishing attacks. Recently, the targets have widened and have shifted from primarily regional banks to nationwide banks. Approximately 65% of phishing attacks target nationwide banks, 30% target the regional banks and only approximately 5% target credit unions. Though there has been some fluctuation in these numbers, each, except for national banks, stays within 5–10% of its base value. Social networking sites are now a prime target for phishing because the personal details provided on such sites can be used in identity theft. Today, there has been an increase in number of phishing websites that are created using phishing toolkits (e.g., Super Phisher, Rock Phish), as these tools simplify the creation of fraudulent websites by stealing the source code of legitimate webpages. Such tools make their phishing easier compared with the manual method of creating a phishing webpage.

The above-mentioned characteristics clearly demonstrate that there is a need for a robust anti-phishing solution for this continuously evolving internet threat. There are several different anti-phishing techniques that have been developed, but there is no single solution that can guarantee total protection against phishing. However, a properly applied technology, along with awareness, significantly reduces the risk of identity theft. There are a large number of possible countermeasures that can be employed to avoid a phishing scam. Most of these countermeasures fall under one of the following three categories.

- Governmental policies against online fraud, such as the Anti-Phishing Act of 2005 in United States to combat phishing and pharming. If a criminal is sentenced under this law, they could risk spending up to five years in prison and/or a fine of $250,000. Similarly, the China Internet Network Information Centre (CNNIC) announced the suspension of new overseas .cn domain registrations to protect against online frauds (Symantec Global Intelligence Network, 2010).
- Creating awareness among users with education and training. Anti-Phishing Phil (Kumaraguru et al., 2010) is an interactive anti-phishing training game that trains users to identify fraudulent and malicious URLs in 10 min; if they make any mistake in the game, they instantly learn why. The game has a Learning Management System (LMS) that targets further training based on the performance of the user.
- Technological countermeasures to detect phishing at the website level fall under one of the following categories: heuristic approaches, blacklist based methods, white-list based methods, hybrid approaches, visual similarity-based approaches and multifaceted approaches.

In our study, we proposed a hybrid anti-phishing approach with three modules to verify the legitimacy of webpages. The first two modules act as preliminary filters that help the system to reduce false positives (FP) and computation by eliminating pages that are preapproved by the user and do not contain login forms. In the third module, we use 15 pivotal heuristics that check the phishiness of pages by examining their structural and behavioural properties. These heuristic results are provided to the machine learning algorithm as a 15-dimensional vector for classification.

The rest of the paper is organised as follows: Section 2 presents an overview of related works. Section 3 illustrates the overall system architecture. In Section 4, 5, 6 and 7, we respectively explain the detailed design and methodology used in the preapproved site identifier; login form finder; webpage feature generator and phishing classifier modules. In Section 8, we compare our method to other anti-phishing methods. The experimental results are discussed in Section 9, and conclusions are presented in Section 10.

2. Related work

Kang and Lee (2007) proposed a global white-list-based approach that prevents access to explicit phishing sites with a URL similarity check. When a user visits a website, the website's URL and IP pair is passed to the Access Enforcement Facility (AEF) to check if the site is a phishing site. If the URL passed to the AEF matches an entry in the trusted site list, then the program checks the similarity of the IP address. If the address also matches, then the program allows the user to proceed; otherwise, it determines the type of phishing attack using other modules and warns the user. Han et al. (2012) developed an Automated Individual White-List (AIWL) approach to protect users' online credentials. In this method, users maintain an individual white-list that records the well-known legitimate websites that the user visits, rather than maintaining a list of all legitimate websites on the internet. In this white-list, along with URL, AIWL also maintains a list of the features of webpages (such as the legitimate IPs of the page and the paths of the input widgets on the page) where the user inputs his or her details. This additional information in the white-list helps AIWL to shield users from different types of online fraud. AIWL warns the user when the submitted account information does not match an entry in the white-list. In our system, we used a private white-list as a primary filtering module. The structure of the white-list was adopted from a study by Han et al., and its functions were adopted from a study by Kang et al.

Chou et al. (2004) developed the browser plug-in "SpoofGuard," which identifies phishing webpages based on a series of heuristics. These heuristics are grouped into stateless method and stateful method. The heuristics of the stateless method identify the suspiciousness of a webpage that is downloaded on a web browser (URL, image, link, and password), whereas the heuristics of the stateful method evaluate a webpage's credibility through the user's recent visits to the page (the user's history file). Fette et al. (2007) proposed a method to detect phishing emails by including features specific to phishing. They proposed 10 different features to identify a phishing email. Eight of these features can be extracted from an email itself. Of the other two features, the age of linked-to domain names has to be obtained with a WHOIS query at the time the email is received, and the spam-filter output feature incorporates the class assigned to the