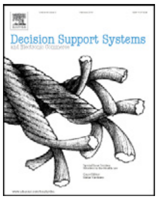
Decision Support Systems 61 (2014) 12–22

Contents lists available at ScienceDirect

Decision Support Systems

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



An efficacious method for detecting phishing webpages through target domain identification

Gowtham Ramesh [a,*], Ilango Krishnamurthi [b], K. Sampath Sree Kumar [a]

[a] *Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, Tamilnadu, India*
[b] *Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Kuniamuthur, Coimbatore, Tamilnadu, India*

ARTICLE INFO

Article history:
Received 26 February 2013
Received in revised form 1 December 2013
Accepted 7 January 2014
Available online 17 January 2014

Keywords:
Phishing
Anti-phishing
E-commerce security
Target domain detection

ABSTRACT

Phishing is a fraudulent act to acquire sensitive information from unsuspecting users by masking as a trustworthy entity in an electronic commerce. Several mechanisms such as spoofed e-mails, DNS spoofing and chat rooms which contain links to phishing websites are used to trick the victims. Though there are many existing anti-phishing solutions, phishers continue to lure the victims. In this paper, we present a novel approach that not only overcomes many of the difficulties in detecting phishing websites but also identifies the phishing target that is being mimicked. We have proposed an anti-phishing technique that groups the domains from hyperlinks having direct or indirect association with the given suspicious webpage. The domains gathered from the directly associated webpages are compared with the domains gathered from the indirectly associated webpages to arrive at a target domain set. On applying Target Identification (TID) algorithm on this set, we zero-in the target domain. We then perform third-party DNS lookup of the suspicious domain and the target domain and on comparison we identify the legitimacy of the suspicious page.

1. Introduction

Phishing encourages end users to visit fake webpages which have similar look and feel of a webpage with the malicious intention to steal user credentials and identities. This identity theft is used for many illegal activities like online money laundering. The losses created as a result of these activities run into billions of dollars [1]. According to the RSA's online fraud report, in the year 2012 there is 59% increase in phishing attacks as compared to 2011 and an estimated loss of more than $1.5 billion due to phishing attacks on the global organizations in the same period, which is 22% higher than in 2011 [2]. This results in people losing faith over the e-commerce industry and leads to significant loss in their market value [3]. Thereby there is a strong demand for an effective measure to curb such phishing attacks.

Any phishing attack usually involves first, creation of a fake website which looks similar to a legitimate website and then lures the users to these fake websites instead of the legitimate webpage, to provide the required authentication and other personal details. These details are extracted from the user without his knowledge. To mitigate this attack many possible counter measures have been developed by the researchers such as white-list based methods, blacklist based methods, heuristic approaches, hybrid approaches or multifaceted mechanisms.

All the aforesaid anti-phishing methods attempt to identify the phishing webpage, but lack techniques to identify the legitimate webpage that the phishing webpage mimics (phishing target [4]). However, any anti-phishing technique becomes incomplete without identification of the phishing target, as it plays a vital role in confirming that there is a phishing attack on a legitimate webpage. Unfortunately, finding the target webpage can be tedious at times when phishers attack the less popular or new webpages. Sometimes, it is also difficult to identify the target because of the masquerading techniques used by the phishers. For example, if the phishers create the webpage using only embedded objects like images and scripts then identifying the target becomes tedious with the existing methods.

Hence, there is a need for a holistic approach that can identify the right phishing target even when attackers use any masquerading techniques. Such a method would gain significant importance among anti-phishing techniques as it alerts the target owners to take necessary counter measures and enhance security.

In this paper, we propose a novel approach to detect the phishing webpages. In this process, we take the webpage under scrutiny and identify all the direct and indirect links associated with the page and generate domain group sets S1 and S2 respectively. From these sets we identify the target domain set, which is given as input to Target Identification (TID) algorithm to identify the phishing target. Using DNS lookup, we map the domains of suspicious webpage and phishing target to corresponding IP addresses. On comparing both the IP addresses, we conclude the authenticity of the suspicious webpage. As our approach depends only on content of the suspicious webpage it requires neither a prior knowledge about the site nor requires the training data.

An overview of literature review and related work is presented in Section 2. Section 3 covers the system overview. In Section 4, we have explained the target domain set construction followed by the target

[*] Corresponding author.
E-mail address: rameshgowtham@gmail.com (G. Ramesh).

0167-9236/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2014.01.002

domain identification using TID-algorithm and phishing detection procedure in Sections 5 and 6 respectively. The implementation details, evaluation methodology and experimental results are discussed in Section 7. We conclude by highlighting the key features of our technique and its limitations in Section 8.

2. Related work

Various solutions to phishing have been developed during the past years. In this section, we briefly review some of the notable anti-phishing works and empirical studies based on it. The studies of different approaches have motivated us to propose a method that overcomes these limitations of the existing schemes.

The white-list approach maintains a list of all safe websites and their associated information. Any website that does not appear in the list is treated as a suspicious website. The current white-list tools usually use a universal white-list of all legitimate websites that need to be constantly updated. In order to simplify this, Han et al. [7] developed an approach to maintain an individual white-list which records the well-known legitimate websites of the user rather than maintaining a universal legitimate site list. In this approach, the Automated Individual White-List (AIWL) records every URL along with its LUI (Login User Interface) information and the legitimate IP addresses mapping to these URLs. Here the AIWL warns the user when the account information submitted to the website does not match with the entry in the white-list. This helps the user to distinguish a pharming website. This technique is adopted and suitably improvised for our work.

The blacklist approach maintains a list of known phishing sites to check the currently visiting website against the list. This blacklist is usually gathered from multiple data sources like spam traps or spam filters, user posts (e.g. phishtank) or verified phish compiled by third parties such as takedown vendors or financial institutions. Prakash et al. [8] used an approximate matching algorithm that divides a URL into multiple components that are matched individually against entries in the blacklist. Zhang et al. [9] proposed a system where customized blacklists are provided for the individuals who choose to contribute data to a centralized log-sharing infrastructure. This individual blacklist is generated by combining relevance ranking score and the severity score generated for each contributor. But the blacklist needs frequent updates from their sources and the exponential growth of the list demands great deal of system resources.

The heuristic-based approaches extract one or more features from a webpage to detect phishing instead of depending on any of precompiled lists. Most of these features are extracted from URL and HTML DOM (Document Object Model) of the suspicious webpage. Zhang et al. [10] proposed a content-based approach CANTINA, based on the tf–idf (term frequency and inverse document frequency) algorithm to identify top ranking keywords from the page content and meta keywords/description tags. These keywords are searched through a trusted search engine such as Google. Here, a webpage is considered legitimate if the page domain appears in the top N search results. CANTINA+ is an upgraded version of CANTINA proposed by Xiang et al. [11], where new components are included to achieve better results. Particularly, they have included ten other features along with four of the CANTINA features and one extended feature. In our approach we have used tf–idf similar to CANTINA to extract keywords from the webpage.

Another heuristic based approach exploring HTML DOM is “Phishark” developed by Prevost et al. [12]. In this research they have analyzed and studied the characteristics of phishing attack and have defined twenty heuristics to detect phishing webpages. These twenty heuristics were then checked for the effectiveness to decide as to which of these heuristics would play a major role in identifying both the phishing and the legitimate webpages. Since, these approaches do not require any pre-compiled lists, they are capable of detecting new phishing webpages by identifying anomalies in it, but legitimate sites also may have such anomalies when it is developed by novice developers. These methods fall short in detecting a phishing webpage made up of only embedded objects like images and scripts.

The other area of research focuses on detecting phishing by comparing visual and image similarities between webpages. Fu et al. [13] proposed an approach which uses the Earth Mover's Distance (EMD) to measure webpage visual similarity. In this approach they first convert the webpage into low resolution images and then use color and coordinate features to represent the image signatures. EMD is used to calculate the signature distances of the images of the webpages. They used trained EMD threshold vector for classifying a webpage as a phishing or legitimate. Medvet et al. [14] proposed an approach which identifies phishing webpages, by considering text pieces and their style, images embedded in the page and the overall visual appearance features of the webpage. Chen et al. [15] present an image based anti-phishing system, which is built on discriminative key point features in webpages. Their invariant content descriptor and the Contrast Context Histogram (CCH) compute the similarity degree between suspicious and legitimate pages. Chen et al. [16] also proposed an approach which uses gestalt theory for detecting visual similarity between two webpages. They used the concept of super-signals to treat the webpage as indivisible unite; these indivisible super-signals are compared using the algorithmic complexity theory. But these techniques may result in false positive when a legitimate page crosses the similarity threshold value and also fails to identify the targeted page.

Multifaceted approaches use any combination of techniques in computational science to detect phishing websites. Joshi et al. [17] developed the PhishGuard tool that identifies phishing websites by submitting actual credentials after the bogus credentials during the login process of a website. They have also proposed architecture for analyzing the responses from server against the submission of all those credentials to determine if the website is legitimate or a phished one. Yue and Wang [18] designed a BogusBiter tool that submits a large number of bogus credentials along with the actual credential of users to nullify the attack. A similar approach has been applied by Joshi et al. [17] but BogusBiter is triggered only when a login page is classified as a phishing page by a browser's built-in detection component.

Shahriar and Zulkernine [19] proposed a model to test trustworthiness to suspected phishing websites. In a trustworthiness testing, they check if the behavior (response) of websites matches with the known behavior of a phishing or legitimate website to decide whether a website is phishing or legitimate. The model is explained using the notion of Finite State Machine (FSM) that captures the submission of forms with random inputs and the corresponding responses to describe the website's behavior. This approach can detect advanced XSS-based attacks that many contemporary tools currently fail to detect.

A category of research focuses on experimental studies to comprehend the significance of implementing anti-phishing strategies. Bose and Leung [5] demonstrated an experimental study showing that the firms that invest in adopting advanced phishing countermeasures earn trust of the customers which in turn reflects as encouraging return in their market value. Lai et al.'s [6] study on identity theft through coping perspective creates awareness for consumers, government agencies and e-commerce industries to counteract against such threats. Chen et al. [3] proposed a method to assess the possible financial loss of phishing targets. In this method key phrase extraction technique is used to discover the reports of phishing attack on firms. To estimate the potential financial loss of firms an event study was conducted to determine the change in market value after the release of phishing attack report. These studies clearly reveal the severity of phishing attacks and requirement of an effective anti-phishing method to protect firms and consumers.

Our work is also motivated by two multifaceted approaches that detect phishing targets along with the phishing webpage. These approaches are discussed in brief below.

Wenyin et al. [20] proposed to identify legitimacy of a given suspicious webpage and discovering its phishing target by calculating and reasoning defined association relations on its Semantic Link Network (SLN). This approach first finds the given webpage's associated pages and then