International Journal of Engineering & Technology, 7 (4.31) (2018) 7-14

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Building Standard Offline Anti-phishing Dataset for Benchmarking

Kang Leng Chiew[1]*, Ee Hung Chang[2], Choon Lin Tan[3], Johari Abdullah[4], Kelvin Sheng[5] Chek Yong[6]

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
**Corresponding author E-mail: klchiew@unimas.my*

Abstract

Anti-phishing research is one of the active research fields in information security. Due to the lack of a publicly accessible standard test dataset, most of the researchers are using their own dataset for the experiment. This makes the benchmarking across different anti-phishing techniques become challenging and inefficient. In this paper, we propose and construct a large-scale standard offline dataset that is downloadable, universal and comprehensive. In designing the dataset creation approach, major anti-phishing techniques from the literature have been thoroughly considered to identify their unique requirements. The findings of this requirement study have concluded several influencing factors that will enhance the dataset quality, which includes: the type of raw elements, source of the sample, sample size, website category, category distribution, language of the website and the support for feature extraction. These influencing factors are the core to the proposed dataset construction approach, which produced a collection of 30,000 samples of phishing and legitimate webpages with a distribution of 50 percent of each type. Thus, this dataset is useful and compatible for a wide range of anti-phishing researches in conducting the benchmarking as well as beneficial for a research to conduct a rapid proof of concept experiment. With the rapid development of anti-phishing research to counter the fast evolution of phishing attacks, the need of such dataset cannot be overemphasised. The complete dataset is available for download at http://www.fcsit.unimas.my/research/legit-phish-set.

Keywords: *Anti-phishing; Dataset for benchmarking; Features; Legitimate and phishing webpages*

1. Introduction

The advancement of information technology has provided many benefits to our life as we are able to handle many daily works by using the Internet services. For example, instead of going to the respective service counter, people nowadays are able to pay their bills at any place they feel convenient and with an Internet connection. However, the extension of this convenience has also come along with some immoral activities that are known as the online crimes. Online criminals always gained their illegal profits from their targets through the vulnerability of the Internet service, and one of the common online crimes is called online phishing.

Online phishing is a security threat which combines social engineering and website spoofing techniques to deceive users into revealing their confidential information [1, 2]. Typically, phisher will try to harvest online users credential such as username, passwords and credit card detail by masquerading as a trustworthy entity on the Internet [3, 4, 5]. To prevent the users from becoming the victims of phishing attacks, many software vendors, research institute and companies have released various anti-phishing techniques [6].

There are many survey publications related to the phishing attacks and anti-phishing techniques. However, according to the best of our knowledge, survey publication on the anti-phishing dataset is still unavailable. The discussions on the correlation of anti-phishing dataset and the experimental results are still inadequate and not profound. Furthermore, there is a lack of a widely recognised standard offline dataset which available for the research community to utilise. A complete package of downloadable anti-phishing datasets is also limited. This situation has caused difficulty to create a consistent condition for fair and rapid benchmarking.

This paper aims to fulfil the gap by providing a standard offline dataset for the anti-phishing research community. Although in a different field, similar works on constructing a standard dataset can be found in [7, 8, 9]. We will look into multiple types of anti-phishing approaches from the past, review on their approaches and datasets, and identify the related anti-phishing features. This insight will ensure the dataset built later to be at optimum flexibility and adaptable by the research community. The discussion on “how a good anti-phishing offline dataset should be designed” will be included in this paper by highlighting some of the factors that may influence the accuracy of experimental results.

The remainder of the paper is structured as follows. In the next section, we will review some of the past anti-phishing works in Section 2 and discuss the dataset used in those works in Section 3. We will later discuss on the factors that may contribute to a good design of an anti-phishing offline dataset in Section 4. In Section 5, we will discuss the construction of the offline dataset in detail based on the review from previous sections. The paper concludes in Section 6.

2. Anti-Phishing Approaches and Features

There are varieties of anti-phishing techniques available in the literature and broadly they can be divided into list-based and heuristic-based approaches. Each has its own effective features to be utilised for the phishing detection. In this section, we will group and highlight the major features from the past research works according to their approaches.