Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 129 (2018) 224–230

Procedia Computer Science

www.elsevier.com/locate/procedia

2017 International Conference on Identification, Information and Knowledge in the Internet of Things

Detecting Phishing Websites via Aggregation Analysis of Page Layouts

Jian Mao[a,*], Jingdong Bian[a], Wenqian Tian[a,b], Shishi Zhu[a], Tao Wei[c], Aili Li[d], Zhenkai Liang[e]

[a]*School of Electronic and Information Engineering, Beihang University, Beijing 100183, China*
[b]*Shanghai Key Laboratory of Integrated Administration Technologies for Information Security, Shanghai 200240, China*
[c]*Baidu USA LLC., Sunnyvale, CA 94089, USA*
[d]*Information Technology Service Center, China National Petroleum Corporation, Beijing 100007, China*
[e]*School of Computing, National University of Singapore, Singapore 117417, Singapore*

Abstract

Phishing websites are typical starting points of online social engineering attacks, including many recent online scams. The attackers develop web pages mimicking legitimate websites, and send the malicious URLs to victims to lure them to input their sensitive information. Existing phishing defense mechanisms are not sufficient to detect with new phishing attacks. In this paper, we aim to improve phishing detection techniques using machine learning techniques. In particular, we propose a learning-based aggregation analysis mechanism to decide page layout similarity, which is used to detect phishing pages. Our experiment results shows that our approach is accurate and effective in detecting phishing pages.

Selection and peer-review under responsibility of the scientific committee of the 2017 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2017).

Keywords: Anti-phishing; Web security; Machine learning; Aggregation analysis

1. Introduction

Phishing websites are typical entrance points of online social engineering attacks, including many recent online scams. In such attacks, the attackers develop web pages mimicking legitimate websites, and send the malicious URLs to victims via spam emails, instant messages, or social network communications. Their goal is to deceive the victims to input their sensitive information (e.g., bank accounts, social security number, etc.). Though phishing attacks do not require advanced technical knowledge and these attack techniques are becoming familiar to users, they are still causing major financial damages. These attacks also negatively influence users' trust toward the web services greatly.

* Jian Mao. Tel.: +86-10-8231-7212; Fax: +86-10-8231-9474.
E-mail address: maojian@buaa.edu.cn

Selection and peer-review under responsibility of the scientific committee of the 2017 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2017).
10.1016/j.procs.2018.03.053

According to the report from anti-phishing working group (APWG), there are 1,220,523 phishing attacks reported in 2016, which is a 65% increase over 2015 [3].

Several types of anti-phishing solutions have been developed. The traditional URL-based antiphishing solutions [18, 9, 20, 12, 11] are limited by the timeliness of malicious URL database update. The solutions based on page contents [26, 17] heavily rely on the context or image processing techniques, which cause high performance overhead. As the phishing pages usually maintain similar page layouts to their target websites, the similarity of page layouts has been demonstrated as an important metric to detect phishing pages [14]. However, these metrics are derived from human experiences, and thus may not be comprehensive to detect new attacks. How to comprehensively evaluate the pages' similarity remains a great challenge.

In this paper, we explore learning techniques to address this problem. Our solution is based on the aggregation analysis mechanism to automatically generate rules to determine layout similarity of web pages and then detect phishing pages. It first trains a similarity classifier using page layout features, then uses the classifier to detect phishing pages. Our evaluation used more than 2,900 phishing web pages from *phishtank.com*. It shows that our approach is effective in creating classifiers and detecting phishing pages via page layout similarity.

In summary, we made the following contributions in this paper:

- We propose a learning-based mechanism to evaluate the similarity of web page layouts and identify phishing pages.
- We define the rules to extract and create effective page layout features and develop a phishing page classifier based on two typical learning algorithms, supporting vector machine and decision tree.
- We prototyped our approach and evaluated it with real-world web page samples from *phishtank.com* and *alexa.com*.

Paper Organization. The rest of this paper is organized as follows. Section 2 introduce the background of our work and gives an overview of our approach. Section 3 presents our main algorithm. Section 4 presents the evaluation results. We discuss closely related work in Section 5 and conclude the paper in Section 6.

2. Overview

In this section, we introduce the problem faced by the layout-based similarity detection and give an overview of our solution.

2.1. Learning-based Layout Similarity Detection

Cascading Style Sheets (CSS) is the commonly used visual layout definition of web pages. Widely support by browsers, CSS rules specify how different classes of web page components should appear, for example, the font type and color of the body of a page.

In our previous work [14], we have demonstrated that CSS-based page layouts can be used as the basis to detect phishing pages. However, the metric used is mainly based on human experiences, and may not comprehensively represent all the statistical similarity properties between page layouts of phishing pages and legitimate pages. Especially, the *threshold*, a critical parameter of that approach is selected based on the similarity score distribution of the collected samples. As a result, its accuracy heavily relies on the completeness of the sample collection and attackers may craft new phishing pages to bypass the detection.

Our goal is to develop methods that can detect the similarity among two page layouts by comprehensively "considering" layout features. Machine learning mechanisms are typically used in such situations, where they are used to infer similarity models according to the statistical properties retrieved from the training samples.

In this work, we integrate and analyze a few of potential learning algorithms. Support Vector Machine (SVM) [23] is a widely used classification algorithm due to its good performance. The basic idea of SVM is to maximize the margin between two classes closest points and find an optimal separating hyperplane between them. Decision Tree (DT) learning [22] is one of the predictive modelling algorithms. It takes a decision tree as the predictive model and