Available online at www.sciencedirect.com

ScienceDirect

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

Computers & Security

Characterization and classification of malicious Web traffic

Katerina Goseva-Popstojanova [a,*,1], Goce Anastasovski [e,1], Ana Dimitrijevikj [d,1], Risto Pantev [b,1], Brandon Miller [c,1]

[a] Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA
[b] Microsoft, Redmond, WA, USA
[c] KeyLogic Systems, Morgantown, WV, USA
[d] Matrix IT Global Services, Macedonia
[e] Alarm.com, Vienna, VA, USA

ARTICLE INFO

Article history:
Received 25 August 2013
Received in revised form
14 December 2013
Accepted 25 January 2014

Keywords:
Web security
Empirical study
Malicious web sessions
Vulnerability scans
Attacks
Statistical inference
Classification

ABSTRACT

Web systems commonly face unique set of vulnerabilities and security threats due to their high exposure, access by browsers, and integration with databases. This study is focused on characterization and classification of malicious cyber activities aimed at Web systems. The empirical analysis is based on three datasets, each in duration of four to five months, collected by high-interaction honeypots which ran fully functional three-tier Web systems. We first explore the types and prevalence of malicious scans and attacks to Web systems, and the extent to which these malicious activities differ in different periods of time or on Web servers running different services. In addition to descriptive statistical analysis, we include an inferential statistical analysis of the malicious session attributes, such as duration, number of requests and bytes transferred in a session. Then, we use supervised machine learning methods to classify attacker activities to two classes: vulnerability scans and attacks. Our main observations include the following: (1) Some characteristics of the malicious Web traffic were invariant across different servers and time periods, such as for example the dominant use of the search-based strategy for attacking the servers and the heavy-tailed behavior of session attributes. (2) On the other side, servers running different services experienced almost complementary profiles of vulnerability scan and attack types. (3) Supervised learning methods efficiently distinguished attack sessions from vulnerability scan sessions, with high probability of detection and very low probability of false alarms. (4) Decision tree based methods J48 and PART performed better than SVM across all datasets. (5) Attacks differed from vulnerability scans only in a small number of session attributes; depending on the dataset, classification of malicious activities can be performed using from four to six features without significantly affecting learners' performance compared to when all 43 features were used.

* *Corresponding author*. Tel.: +1 3042939691.
E-mail address: Katerina.Goseva@mail.wvu.edu (K. Goseva-Popstojanova).
[1] This work was done while all the authors were affiliated with West Virginia University.
0167-4048/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cose.2014.01.006

1. Introduction

Many business and everyday activities heavily rely on Web applications. These applications have many vulnerabilities and typically are targeted by a large number of cyber attacks due to their high exposure, access by browsers, and integration with databases. The 2012 Cost of Cyber Crime Study conducted by the Ponemon Institute reported that the average annualized cost of cyber crime for the 56 organizations included in the study was 8.9 million dollars per year (CyberCrime, 2012). The most costly cyber crimes were caused by denial of service, malicious insiders and Web-based attacks. SANS Institute Annual update of the top 20 security risks (SANS, 2007) stated that almost half of the vulnerabilities discovered in 2007 were Web application vulnerabilities. Another study recently conducted by the WhiteHat Security (WhiteHat, 2012), which was based on assessment of around 7000 Web sites, reported that the average number of serious vulnerabilities found per Web site in 2011 was 79. When it comes to cyber attacks, the Computer Security Institute reported that 92% of respondents to a survey experienced more than ten Web site incidents (Gordon et al., 2005). The cyber attacks have short-term impacts on day-to-day activities of end users, businesses, and governments (e.g., losses due to fraudulent activities, unavailability of computer resources) and long-term impacts (e.g., loss of intellectual property, national security breaches) (Choo, 2011).

The constant introduction of new technologies makes the problem of securing Web systems even more challenging. For example, Web 2.0 technologies enhance information sharing, collaboration, and functionality of the Web, but due to users ability to create content they also provide attackers with a broad range of new vulnerabilities to exploit. These trends clearly illustrate the need for better understanding of malicious cyber activities based on both qualitative and quantitative analysis, which will allow better protection, detection, and service recovery.

To be of practical value, analysis of malicious activities have to account for emerging technologies that typically introduce new types of vulnerabilities. However, there is an evident lack of publicly available, good quality, recent data on cybersecurity threats and malicious attacker activities. Therefore, significant amount of intrusion detection research work in the past was based on publicly available, but outdated datasets, such as the KDD Cup 1999 dataset (KDD, 1999) derived from the DARPA Intrusion Detection Evaluation Project (DARPA, 1999). Even more, most of research work on intrusion detection was focused on development of data mining techniques aimed at constructing a "black-box" that classifies the network traffic on malicious and non-malicious, rather than on discovery of the nature of malicious activities (Julisch, 2002).

Facing the lack of publicly available, recent data on malicious attacker activities, we decided to develop and deploy high-interaction honeypots as a means to collect such data. These honeypots were legitimate servers, which were used to collect information on attacker activities. They were configured in a three-tier architecture (consisting of a front-end Web server, application server, and a back-end database) and had meaningful functionality. Furthermore, they ran standard off-the shelf operating system and applications which followed typical security guidelines and did not include user accounts with nil or weak passwords.

Our experimental setup was based on a sound design, which limits the threats to validity. We developed and deployed honeypots running two different sets of services: one running Web 2.0 applications (i.e., blog and wiki) and another running widely used Web-based database administration software (i.e., phpMyAdmin). The work presented in this paper is based on three datasets collected by these honeypots, which allowed us to compare malicious activities aimed at systems with same configuration during different periods in time (i.e., Web 2.0 I and Web 2.0 II), as well as malicious activities aimed at different system configurations during same period of time (i.e., Web 2.0 II and WebDBAdmin). Each dataset is in duration of four to five months and consists of two sets of data collected by a pair of identical honeypots, one advertised and the other unadvertised. This way we were able to distinguish between malicious activities that used search-based strategy (based on search engines and crawlers) and those that use IP-based strategy (when an attacker scans or attacks an IP addresses without previous involvement of search engines and crawlers).

Using these three datasets we conducted in-depth empirical analysis of attacker activities classified as different types of vulnerability scans and attacks. It should be noted that our datasets represent dynamic information on attacker activities, unlike data extracted from vulnerability databases (e.g., (NVD, 2013)) that are focused on static information related to description of known vulnerabilities and the ways they may be exploited. Correspondingly, we study and model dynamic attacker behaviors aimed at scanning and attacking Web systems, which is different than modeling the discovery process of vulnerabilities present in software applications (see for example the work by Woo et al. (2011)).

In the context of this paper, a Web session is considered as an *attack* session if the attacker attempts to exploit a vulnerability in at least one request in that session. If all requests in the session were used to check for vulnerabilities then the session is considered as *vulnerability scan*. Specifically, we addressed the following research questions related to the characterization of the malicious cyber activities:

RQ1: What types of vulnerability scans and attacks are launched on Web systems? (Sections 4.1 and 4.2)

RQ2: What are the statistical characteristics of malicious Web sessions? (Section 4.3)

RQ3: Are the types and distributions of the vulnerability scans and attacks invariant (1) over time (i.e., for the same system configuration during different periods of time) and (2) across systems running different services? (Sections 4.2 and 4.3)

This paper also addresses the problem of automatic classification of malicious Web sessions to two classes: vulnerability scans and attacks. Both attacks and vulnerability scans are malicious activities. Being able to automatically classify them is important because actual attacks are much more critical events than vulnerability scans. It should be noted that our goal was not to identify whether attacks were preceded by vulnerability scans. Rather, our goal was to distinguish