Available online at www.sciencedirect.com

ScienceDirect

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

Computers & Security

Analyzing the ecosystem of malicious URL redirection through longitudinal observation from honeypots

Mitsuaki Akiyama [a,*], Takeshi Yagi [a], Takeshi Yada [a], Tatsuya Mori [b], Youki Kadobayashi [c]

[a] NTT Secure Platform Laboratories, Tokyo, Japan
[b] Waseda University, Tokyo, Japan
[c] Nara Institute of Science and Technology, Ikoma, Nara, Japan

ARTICLE INFO

Article history:
Available online 11 January 2017

Keywords:
Honeypot
Compromised website
URL redirection
Drive-by download
Domain generation algorithm

ABSTRACT

Today, websites are exposed to various threats that exploit their vulnerabilities. A compromised website will be used as a stepping-stone and will serve attackers' evil purposes. For instance, URL redirection mechanisms have been widely used as a means to perform web-based attacks covertly; i.e., an attacker injects a redirect code into a compromised website so that a victim who visits the site will be automatically navigated to a malware distribution site. Although many defense operations against malicious websites have been developed, we still encounter many active malicious websites today. As we will show in the paper, we infer that the reason is associated with the evolution of the *ecosystem of malicious redirection*.

Given this background, we aim to understand the evolution of the ecosystem through long-term measurement. To this end, we developed a honeypot-based monitoring system, which specializes in monitoring the behavior of URL redirections. We deployed the monitoring system across four years and collected more than 100K malicious redirect URLs, which were extracted from 776 distinct websites. Our chief findings can be summarized as follows: (1) Click-fraud has become another motivation for attackers to employ URL redirection, (2) The use of web-based domain generation algorithms (DGAs) has become popular as a means to increase the entropy of redirect URLs to thwart URL blacklisting, and (3) Both domain-flux and IP-flux are concurrently used for deploying the intermediate sites of redirect chains to ensure robustness of redirection.

Based on the results, we also present practical countermeasures against malicious URL redirections. Security/network operators can leverage useful information obtained from the honeypot-based monitoring system. For instance, they can disrupt infrastructures of web-based attack by taking down domain names extracted from the monitoring system. They can also collect web advertising/tracking IDs, which can be used to identify the criminals behind attacks.

[*] *Corresponding author.*
E-mail address: akiyama.mitsuaki@lab.ntt.co.jp (M. Akiyama).

http://dx.doi.org/10.1016/j.cose.2017.01.003

1. Introduction

Attack campaigns targeting websites, e.g., Beladen, Gumblar, and Nine-ball, successfully affected tens of thousands of websites in 2009 (Websense Security Labs, 2009). In addition to this mass compromising, we faced serious server-side vulnerabilities such as Heartbleed, ShellShock, and Poodle in 2014 (Durumeric et al., 2014). The above threats exploit websites to expose many of them to the risk of data tampering. If such data tampering is applied to a website, it can generate web-based attacks for accomplishing attacker's *purposes*. One such purpose of inflicting serious damage to victims is to compel a compromised website to serve as a stepping-stone of various web-based attacks, e.g., drive-by download exploits. Once a victim visits such a compromised website, he/she will be redirected to another website, which exploits the vulnerabilities of web browsers to automatically download and install malware. URL redirection is used as a fundamental *mechanism* to broadly collect web accesses across websites. In addition, injecting redirect codes into various compromised websites is one *strategy* to surreptitiously conduct attacks.

As many defense operations against malicious websites have been developed (Grier et al., 2012; Moshchuk et al., 2006; Provos et al., 2008), the web-based attack, which consists of the aforementioned purposes, mechanisms, and strategies, must also have substantially evolved. We infer the reason we still encounter a large number of active malicious websites used for web-based attacks today may originate from the evolution of its core mechanism – *malicious URL redirection*.

On the basis of the discussion above, our goal in this study was to characterize the *ecosystem of malicious URL redirection*, which plays a vital role in the attack vector. While many new mechanisms have been incorporated into attacks, one interesting observation we found through the long-term study of the ecosystem was that URL redirection has always been used since the emergence of attacks. Such an *invariant* should play a key role in controlling the success of attacks.

With this background in mind, we pose the following research questions:

RQ1: *What are the key characteristics of URL redirection mechanisms?*
RQ2: *Have their purposes been changed over time?*

To answer the research questions, we developed a honeypot-based monitoring system, which specializes in monitoring the behavior of URL redirections. We deployed the monitoring system across four years, resulting in the collection of more than 100K malicious redirect URLs extracted from 776 websites compromised by fraudulent accesses with stolen credentials. We conducted an in-depth analysis of collected URL redirections.

Our key findings can be summarized as follows:

- Although the main purpose of URL redirection caused by redirect code injections has been drive-by download-based malware infection, click-fraud has become a new purpose in addition to malware infection in recent years.
- The use of *domain generation algorithm* (DGAs), which were originally used for communication of a bot and a command and control (C&C) server, has become popular as a means to increase the entropy of redirect URLs to thwart URL blacklisting.
- Both *domain-flux* and *IP-flux* should be concurrently used for deploying the intermediate sites of redirect chains to ensure robustness of redirection.

We unveiled the above change through four years of measurement. The insight obtained from analyzing the observed ecosystem helps network and security operators disrupt attack campaigns in appropriate points and layers corresponding to the distinct purpose, mechanism, and strategy. For this purpose, it is essential to reveal the purposes of attacks, mechanisms of redirection, and strategies to conduct attacks to gain security knowledge in a timely manner as well as protocol-level measurements such as passive DNS and web proxy. To the best of our knowledge, there have been no studies on the analysis of malicious websites introduced from the longitudinal viewpoint, i.e., even though some studies used a huge volume of datasets, these datasets were obtained by just one-time inspection or repeated inspections during a short period. Our study is a pioneer example of honeypot-based measurement to observe the ecosystem.

Based on the obtained knowledge, we also investigated the operational difference between types of DGAs and effectiveness of conventional methods, and discussed practical mitigation strategies against the ecosystem of malicious URL redirections: countering DGAs, discovering unknown compromised websites, and disabling web advertising and tracking IDs of attackers.

The rest of the paper is organized as follows. In Section 2, we detail our monitoring system and give a summary of the data we collected. In Section 3, we analyze the URL redirection mechanism in detail. In Section 4, we present an analyzed complex URL redirection structure, which uses DGAs. In Section 5, we discuss countermeasures to malicious URL redirection and domain-flux techniques. Then we evaluate the generality and impact of our observation in Section 6. We introduce related work in Section 7 and conclude this paper in Section 8.

2. Extracting URL redirection

2.1. *Definition of URL redirection*

Redirection refers to automatically replacing access destinations, and it is generally controlled over an HTTP protocol on the web. In addition to this conventional method, other methods for automatically accessing external web content, e.g., *iframe* tag, have been often used, particularly for web-based attacks. In this paper, we originally define that URL redirection additionally includes automatically occurring web access to URLs corresponding to an initial accessed URL and assume that URL redirection methods are tag redirections (`iframe`, `script`, `meta`, etc.), script redirections (JavaScript `location`'s methods), and HTTP redirection (HTTP-3xx status code).