Quantifying Maliciousness in Alexa Top-Ranked Domains

Paul Royal
Barracuda Labs
proyal@barracuda.com

ABSTRACT
Many people assume that it is safe to visit popular, long-lived websites. While anecdotal examples of popular website compromises contradict this expectation [1], there exist few studies that attempt to systematically quantify maliciousness in top-ranked sites. To address this gap in understanding, this paper details the design and results of long-running experiments that identify maliciousness in popular websites in a vulnerability and exploit-independent manner.

1. INTRODUCTION

At Barracuda Labs, we use a variety of research technologies to identify and study maliciousness on the web. One of these tools is an automated system that forces a web browser inside a Windows virtual machine to visit a URL to see what happens to the browser, its plugins, and the operating system. The resulting network-level actions of the virtual machine help us determine, without prior knowledge of specific exploits served to the browser or its extensions, whether a URL serves malicious content.

In February and May of 2012, we used the above-described system to examine the Alexa 25,000 most popular domains. In combination with manual whitebox analysis to confirm maliciousness, automated blackbox examination of the Alexa top 25,000 each day for our two month-long studies shows that this assumption does not always hold.

2. ESTIMATING IMPACT

After identifying a top-ranked domain as serving a drive-by download, we estimate the impact using the Alexa ranking system. While Alexa does not publish the total number of page views it uses to compute site rankings, there exists sufficient information to determine that number. As an example, Wikipedia, which represented ~0.54% of total Alexa views in February 2012, reported ~15.75 billion views for the previous month [2]. Working backwards, we can thus calculate that Alexa used an average of (15,756 * 1,000,000)/(29 * (0.5416/100)) = ~100.31 billion views each day to rank the popularity of websites.

Using the above number, we can calculate the affected views for a given site in a 24-hour period. As an example, free-tv-video-online[.]me, which via an ad network served visitors malicious content on February 13, 2012, represented ~0.0053% of the total Alexa views; that yields 5,366,895 affected views for the day. However, to estimate how many users were served exploit content, this number must be adjusted to account for the average number of views per user. Fortunately, Alexa makes this information available. Continuing with the example, free-tv-video-online[.]me has an average of 7.2 views per user per day. Thus, for this site, 5,366,895 views equates to 745,402 users served malicious content on that day.

Of course, not every user served malicious content is compromised. To estimate the number of successfully exploited users, we used several different sources, including Wikipedia's browser statistics [3]. To begin, if we examine platform and browser popularity, only about half (or 50.81%) of users (who run Windows and IE or Firefox) possess properties conducive to exploitation.

To convert the number of possibly compromised users into those probably compromised, we conservatively adjusted according to the most popular mechanism of exploitation: the Java plugin. According to Adobe [4], 73% of PC users have the Java plugin installed. According to Qualys [5], 42% of users with the Java plugin installed have versions vulnerable to exploitation. Thus, as a conservative estimate, only 42% of 73% of 50.81%, or 15.57% of users served malicious content are likely to be successfully compromised.

3. FEBRUARY AND MAY 2012 STUDIES

Using the system and procedures described in the previous sections, we performed two month-long studies: one for February 2012 and one for May 2012. During February 2012, 58 of the Alexa top 25,000 domains served drive-by downloads when visited. Applying the previously described estimations, 10,541,379 users were served malicious content and 1,642,172 were likely compromised. Details of the 58 top-ranked drive-by download sites, as well as full packet captures of each individual drive-by download session, are available for download online [6].

In May 2012, we augmented the system to re-check each drive-by download site for recurring maliciousness. Of the 39 top-ranked domains found to result in drive-by downloads in May 2012, 11 (or 28%) yielded malicious content for more than one day; the average period of maliciousness was just over 36 hours. We estimate that in total, 7,881,423 users were served malicious content and 1,227,774 were likely compromised. As with the February 2012 study, a list of the 39 sites, associated details, and network traces of each drive-by download are available for download [7].

For the May 2012 study, we also employed whitebox analysis to examine the use of ad networks and targeted software components. Of the 39 sites, 34 (or 87.1%) served malicious content (usually targeting multiple software components) that included one or more exploits for Java, which supports the widely held belief that Java is currently one of the most ubiquitous targets of drive-by download attacks. Barracuda Labs recommends that users disable Java support in the web browser and re-enable the feature only when necessary.

Finally, in addition to the investigation of targeted software, we also examined how, beginning with a visit to a popular website, malicious content was served to the browser. Given that almost all of the sites were long lived, we expected most instances of malicious content to arrive via the sites' use of ad networks, which are a frequent target of cyber criminals. However, to our surprise, malicious content originated from ad networks on only 18 (or 46.1%) of the 39 sites. The remainder were, in one form or another, the result of direct website compromise.

4. CONCLUSION

This study has provided an initial examination of maliciousness in top-ranked domains. Multiple, month-long studies were conducted to systematically evaluate the intuition that it is safe to visit popular, long-lived websites. The results indicate that even the mainstream, popular web is not a safe place – each month, millions of users are served malicious content from just tens of popular websites, and at least one million users are successfully compromised.

5. REFERENCES

[1] Barracuda Labs. PBS Website Compromised, Used to Serve Exploits. http://www.barracudalabs.com/wordpress/index.php/2009/09/16/pbs-website-compromised-used-to-serve-exploits .

[2] Domas Mituzas. Page Views for Wikipedia. http://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm .

[3] Erik Zachte. Wikimedia Traffic Analysis Report – Browsers. http://stats.wikimedia.org/archive/squid_reports/2012-01/SquidReportClients.htm .

[4] Adobe Systems Incorporated. Flash Runtime Statistics. http://www.adobe.com/products/flashruntimes/statistics.html .

[5] Neil J. Rubenking. Qualys Releases Report on Faulty Browser Plugins. http://www.pcmag.com/article2/0,2817,2380432,00.asp .

[6] Barracuda Labs. Details and Network Capture Data for Alexa Top 25K Drive-by Downloads in February 2012. https://www.dropbox.com/s/b2qf00x3t0fsdq4/alexa_ddl_feb_2012.zip .

[7] Barracuda Labs. Details and Network Capture Data for Alexa Top 25K Drive-by Downloads in May 2012. https://www.dropbox.com/s/vph70xz71docpcw/alexa_ddl_may_2012.zip .