World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:8, No:06, 2014

Improving Search Engine Performance by Removing Indexes to Malicious URLs

Authors : Durga Toshniwal, Lokesh Agrawal

Abstract : As the web continues to play an increasing role in information exchange, and conducting daily activities, computer users have become the target of miscreants which infects hosts with malware or adware for financial gains. Unfortunately, even a single visit to compromised web site enables the attacker to detect vulnerabilities in the user's applications and force the downloading of multitude of malware binaries. We provide an approach to effectively scan the so-called drive-by downloads on the Internet. Drive-by downloads are result of URLs that attempt to exploit their visitors and cause malware to be installed and run automatically. To scan the web for malicious pages, the first step is to use a crawler to collect URLs that live on the Internet, and then to apply fast prefiltering techniques to reduce the amount of pages that are needed to be examined by precise, but slower, analysis tools (such as honey clients or antivirus programs). Although the technique is effective, it requires a substantial amount of resources. A main reason is that the crawler encounters many pages on the web that are legitimate and needs to be filtered. In this paper, to characterize the nature of this rising threat, we present implementation of a web crawler on Python, an approach to search the web more efficiently for pages that are likely to be malicious, filtering benign pages and passing remaining pages to antivirus program for detection of malwares. Our approaches starts from an initial seed of known, malicious web pages. Using these seeds, our system generates search engines queries to identify other malicious pages that are similar to the ones in the initial seed. By doing so, it leverages the crawling infrastructure of search engines to retrieve URLs that are much more likely to be malicious than a random page on the web. The results shows that this guided approach is able to identify malicious web pages more efficiently when compared to random crawling-based approaches.

Keywords : web crawler, malwares, seeds, drive-by-downloads, security
Conference Title : ICACS 2014 : International Conference on Advanced Computing Systems
Conference Location : New York, United States
Conference Dates : June 05-06, 2014