Abstract No: SO-10 **Software Intensive Systems**

Offline analysis of web logs to identify offensive web crawlers

Nilani Algiriyage

Department of Industrial Management, Faculty of Science,
University of Kelaniya, Sri Lanka
**Email: nialnia@kln.ac.lk*

With the continuous growth and rapid advancement of web based services, the traffic generated by web servers have drastically increased. Analyzing such data, which is normally known as click stream data, could reveal a lot of information about the web visitors. These data are often stored in web server "access log files" and in other related resources. Web clients can be broadly categorized into two groups: web crawlers and human visitors. During recent past, the traffic generated by web crawlers has drastically increased. Web crawlers are programs or automated scripts that scan web pages methodically to create indexes. They traverse the hyperlink structure of the worldwide web to locate and retrieve information. Web crawler programs are alternatively known as web robots, spiders, bots and scrapers.

Web crawlers can be used by anyone seeking to collect information available on the Internet. Search engines like Google, Yahoo, MSN and Bing use web crawlers to index web pages to be used in their page ranking process. Web administrators employ crawlers for automating maintenance tasks such as checking for broken hyperlinks and validating HTML codes. Business organizations, market researchers or anyone can gather specific types of information such as e-mail address, corporate news and product prices.

A recent threat with web crawlers is some try to crawl web sites hiding their own identity and pretending to be someone else. Since Google is the widely used search engine globally and web site owners do not want to block the Googlebot, imposters try to crawl sites impersonating Googlebot. The fact is they get privileged access to web sites using the identity of "Googlebot". Googlebot impersonation can lead to spamming, information theft including business intelligence, or even application level DDoS (Distributed Denial of Service attacks). Although there were recent news items of fake Googlebots, current understanding of this problem is minimal. While it is possible to later identify (e.g. using web server access log files) these Googlebot imposters by using a reverse DNS (Domain Name System) lookup and a forward DNS lookup case by case basis, doing this real time will be much useful but challenging. We observed multiple instances of PHP remote code execution vulnerability scans by these fake Googlebots in our test data sets.

Off-line or postmortem analysis of web server access log files could give a deep understanding of traffic patterns, and especially to identify offensive web clients. Although the detection is after-the-fact, proactive strategies can be formulated based on the gathered knowledge. This research proposes a methodology to detect malicious web crawlers based on seven behavioral features including hit rate, blank referrer, hidden links, IP verification, IP blacklist checks, access of *"robots.txt"* file and the access depth. The results show that 36.23% of the crawler sessions exhibit malicious crawling patterns.

Keywords: Web-crawler, Web server access logs, Web usage mining