Hillstone NETWORKS

Hillstone intelligent Next-Generation Firewall White Paper: A Hybrid Approach to Detect Malicious Web Crawlers

What is a web crawler?

A web crawler (also called web spider, web robot) is typically a script or computer program that browses the targeted website in an orderly and automated manner. It is an important method for collecting information on the Internet and is a critical component of search engine technology. Most popular search engines, such as GoogleBot and BaiduSpider, use underlying web crawlers to get the latest data on the internet.

All web crawlers take up internet bandwidth. But not all web crawlers are benign. A well behaved web crawler usually identifies itself and balances the crawling frequencies and contents and thus the bandwidth consumption. On the other hand, an ill-behaved or malicious web crawler can consume large amounts of bandwidth and cause disruptions, especially to companies that rely on web traffic or content for their business.

For companies that rely on their website and online content to conduct business, if a web crawler is created by a hacker or unauthorized users and used on bots, it can be used to steal data and information from businesses with the possibility of staging DDOS attacks towards targeted websites.

How to effectively detect malicious web crawlers has become a critical topic in today's cyber threat defense sector.

Web Crawler Characteristics

Since malicious or ill-behaved web crawlers are primarily scripting programming that runs on bot machines, they typically have the following behavior with some variants:

- High HTTP request rate and typically done in parallel.
- Large amount of URL visits in terms of total number of URLs as well as the number of directories
- More requests for specific file types versus others; for example, more requests for .html, .pdf files, and fewer for .jpeg, .PHP files, etc.
- Scarce use of HTTP POST method since the main purpose is to download information from the website versus uploading.
- Potentially more HTTP HEAD methods used (compared with normal browsing) since a crawler often needs to determine the types of files before it tries to crawl it.
- Potentially higher numbers of smaller sized files among the HTTP GET method returns. This is because, very often, a crawler needs to maximize results of its crawling within a minimal amount of time and therefore skip those large sized files and go for smaller ones.
- In case some URLs being crawled need further authentication, the HTTP requests from the crawlers will be directed to those authenticating pages, resulting in 3XX or 4XX of HTTP request return codes.

Common Web Crawler Detection Methods

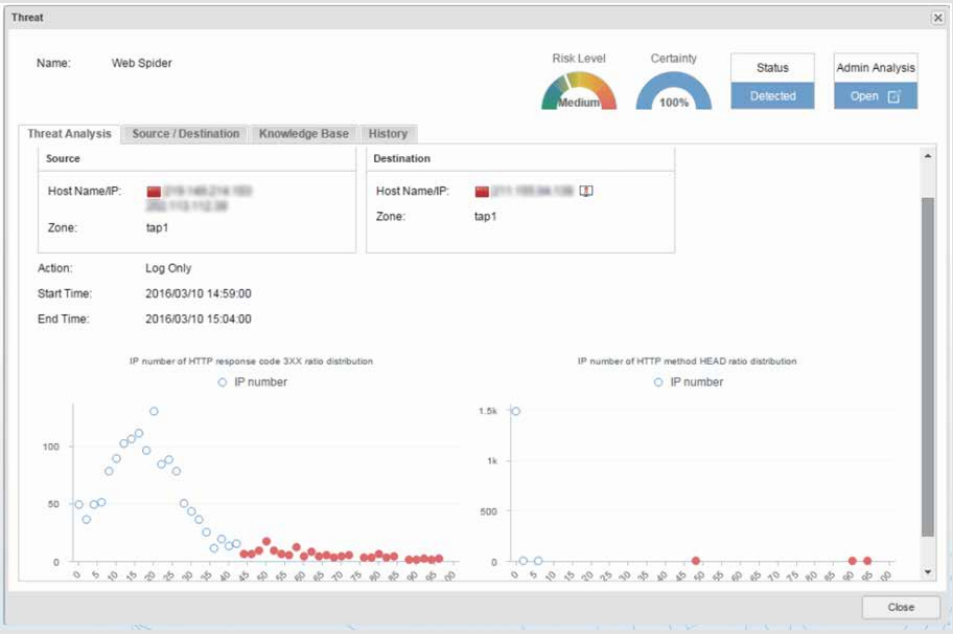
Commonly used methods such as proper configuration in robots.txt files on server, whitelisting user-agent, among others, can detect and block some low level malicious crawlers. Advanced and sophisticated web crawlers are still difficult to detect because they can hide behind legitimate ones. Additionally, IT departments can invest time and resources to collect and analyze network traffic logging reports to surface hidden traces of web crawlers.

Take for example, the below snapshot of an actual logging data from a content hosting company. IT staff can identify the most visited IP addresses after sorting the log data; after filtering out those on the whitelists, the most visited and suspicious IP addresses can be further examined and action can be taken if they are determined to not belong to known and benign lists.

Hillstone' s Hybrid Approach to Detecting Suspicious Web Crawlers

Using logging data analysis to identify suspicious or malicious web crawlers, however effective, is a labor intensive and sustaining effort and often consumes a lot of time and resources for IT departments.
Detection methods that are solely based on statistics from logging data can often generate false positive alerts, for example, they can't distinguish a DOS attack from a crawler. Furthermore, this method can be ineffective in detecting slow moving web crawlers. This is because there is usually a vast amount of log data collected at any given point of time, and log data can only be stored for specific periods of time, and as time passes, slow moving crawlers usually lose all traces.
Hillstone Networks has adopted a hybrid approach that uses not only statistical logging data analysis but more importantly, focuses on behavioral modeling to detect suspicious web crawlers. This has proven to be effective in detecting sophisticated, malign crawlers as well as slow crawlers that are prone to losing trace.
In this hybrid approach, a set of pre-defined L3-L7

behavioral features monitor and collect data at the data plane, which is then fed into several behavioral models using machine learning algorithms that learn and profile these behavioral features periodically.
In tandem, network and application level traffic logging data collected over specific periods of time, are also processed, sorted, filtered and analyzed.
Built on the predictive results of the behavioral modeling and statistical analysis from the logging data, a set of correlation rules are defined to correlate the corresponding results from different detecting modules. They are used to identify those IP addresses that are "abnormal" compared with the IP addresses that have normal web accessing and browsing behavior. The final result is a classified threat event that is saved into the threat event database.
The solution also offers a user interface for network and IT staff with clear and accurate visibility of suspicious web crawler activity along with corresponding IP addresses and other forensic data so that they can take proper action to mitigate these actions.
The following are two examples that illustrate suspicious web crawler activity and detection using behavioral modeling and analysis:

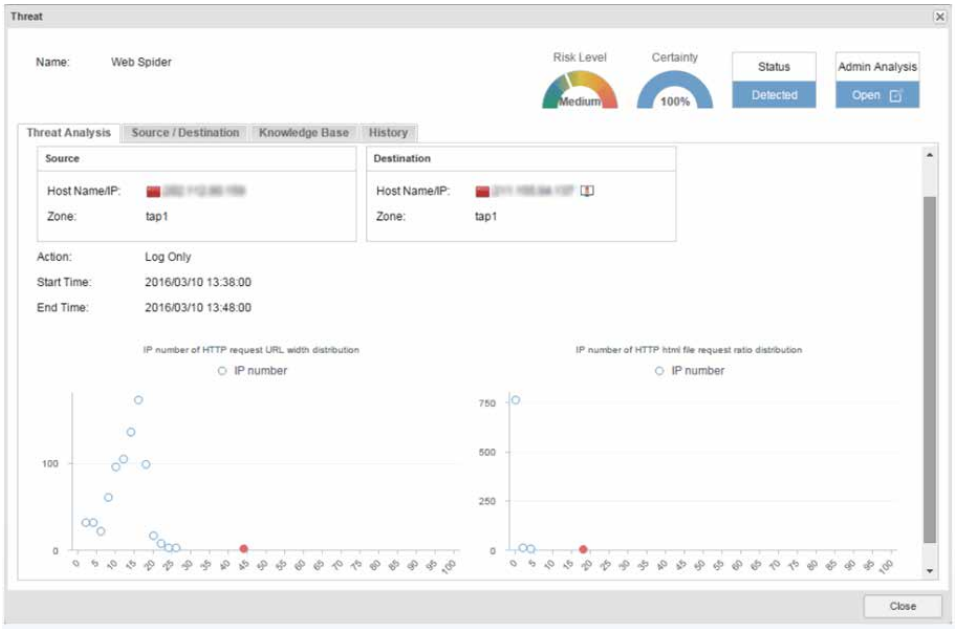


In the above example, you can note the following:

- On the left graph, the red dots represent the abnormality of HTTP requests with 3XX return code. You will notice that some IP addresses have 65% of 3XX return codes; other IP addresses have 100% of 3XX return codes.
- On the right graph, the red dots indicate the abnormality of those having URL width (i.e. directories visited) requests within a learning cycle. Some IP addresses have a significantly higher number of URL directories visited over others in one learning cycle.

Hillstone's behavioral model feature analyses these abnormal IP addresses (those depicted by the red dots) and correlates those IP addresses that fit these two behavioral abnormality rules. It is easy to identify potentially suspicious IPs that might be conducting malicious web crawling. In this case, the IP address 219.149.214.103 is a suspicious candidate.

Another example is shown below:



In this example, we note the following:

- On the right graph, the red dots indicate the abnormality of those having URL width (i.e. directories visited) requests within a learning cycle. Some IP addresses have significantly higher number of URL directory visits than others in one learning cycle.
- On the left graph, the red dots indicate those IP addresses that have abnormal (higher) number of HTML files request compared with other IP addresses that are monitored.

The Hilstone behavioral model features will then perform and analysis of the abnormal IP addresses (those depicted by the red dots), and correlate those IP addresses that fit these two behavioral abnormality rules. It is easy to identify the potentially suspicious IPs that might be conducting malicious web crawling. In this case, the IP address 202.112.90.159 is such a suspicious candidate.

Conclusions

Using manual and static analysis on logging data (based on most visited IP addresses) can be labor-intensive and incur higher cost and more overhead; but more importantly, can be often ineffective if it mistakenly misses slow crawlers with lower IP address numbers in the logging data. Hillstone's hybrid solution uses a proprietary self-learning behavioral modeling mechanism that is more effective in detecting these slow crawlers. It also provides statistical analysis to automatically detect sophisticated and suspicious web crawlers as well as rich and actionable forensic evidence to the administrator.

How to contact Hillstone Networks
AMERICAS: +1-800-889-9860
APAC: +86-400-828-6655
EUROPE: +420-910-117-735
EMAIL: inquiry@Hillstonenet.com