



WHITE PAPER



SPIDERS AND BOTS

Preventing Click Fraud

September 7, 2007

Version 2.0



1 Spiders and Bots

A spider is a program that reads individual pages within a website in order to create entries for a search engine index. It is also known as a crawler or a bot. SiteCatalyst does not track most spiders since they do not execute JavaScript or make image requests. If the search engine spider has been modified to make image requests, the corresponding traffic that the spider has created on the website will be part of the traffic displayed within SiteCatalyst.

Click fraud is the practice of using technology (most commonly “bots”) or humans to artificially inflate traffic data to defraud advertisers and web sites that provide venues for advertisers. The main goal of those engaging in click fraud activities is to click on advertiser’s links on search engines or web sites in order to maliciously drive CPC and other advertising costs higher for companies (many times competitors) and to falsely increase traffic that results in low conversion rates. Clients who use Omniture’s SiteCatalyst product can also be indirectly affected due to the increased fraudulent traffic and clicks that result in more server calls and associated CPM costs. Furthermore, the abnormal traffic and clicks caused by click fraud may skew the rest of the site’s SiteCatalyst data.

There are a variety of crawlers that will “spider” a site and create web log entries. Many of these are filtered by popular log analyzer products. New crawlers are being created all the time, and the filters require constant updates. Many crawlers are homegrown and others are disguised to appear like a browser to harvest emails. SiteCatalyst’s data collection technology eliminates the spiders’ traffic because spiders and bots normally do not execute JavaScript nor make image requests.

The SiteCatalyst data collection method is more accurate than older, web log based solutions for various reasons. First, SiteCatalyst overcomes both proxy and browser caching of pages (which web logs do not measure). Second, SiteCatalyst measures only pages loaded within a web browser and not automated traffic such as spiders and bots. Third, SiteCatalyst uses cookies to measure visitors, whereas most log file-based solutions rely upon IP addresses; the visitors are, therefore, more accurately measured.

Much of the click fraud activity from bots or spiders will be filtered from the SiteCatalyst reports. Omniture does not proactively filter bots or spiders (including those bots intending to inflict fraudulent click-through traffic) from any client’s data or reports. However, most bots and spiders do not load images or execute JavaScript, and would not create page views in the first place.

Most spiders do not load images or execute JavaScript, and would therefore not create pageviews in the first place. However, on rare occasions, some bots and spiders request images, and may request the <noscript> image from a client’s web page, which results in an image request to SiteCatalyst. This can only occur if the customer is using “split” code that includes the <noscript> image. Otherwise, only bots and/or spiders that execute JavaScript will be included in SiteCatalyst data. It is very rare that a bot or spider would execute JavaScript (no common ones are known), but it is possible. Additionally, not every spider wishes to be detected. As a result, many spiders and bots no doubt mimic the User Agent of a “standard” browser.

For those sites that are affected by click fraud, SiteCatalyst can in some cases be used to identify the click fraud activity. Some of the SiteCatalyst reports that can be useful in identifying click fraud activity include the Page View Report, Referrer Reports, Campaign Reports, and others. Unexpected and/or unexplained traffic spikes could result from click fraud and may warrant further investigation.

If click fraud activity is identified through the SiteCatalyst reports (or by other means), clients can use the standard IP exclusion functionality to filter the click fraud activity from their reports. The IP exclusion functionality incurs no additional charge, but clients are only able to exclude up five IP addresses/ranges for free. Clients can also request that Omniture implement a custom VISTA rule that filters click fraud traffic based on either IP address or User Agent string (fees may apply). The customer will need to supply a list of User Agents or IP addresses to filter. Using VISTA to filter click fraud activity prevents the data from showing in the SiteCatalyst reports, but does not prevent the activity from being included as a billable server call since the data is still processed by the Omniture data collection servers before the VISTA rule filters the data.

1.1 Examples of Spiders and Bots

There are many different types of spiders and bots, including the following.

- Search engines (Google, Inktomi, etc.)
- Other "indexers" such as shopping sites (mysimon.com, etc.)
- Bots that automate processes
- Information portals and processing services
- Multimedia databases (singingfish, etc.)
- Educational Institutions/Projects (Carnegie Mellon CS Dept, etc.)

1.2 Listings of Spiders and Bots

The Internet contains listings of spiders and bots used online. Online listings are located in the following URLs.

<http://www.psychedelix.com/agents.html>

<http://www.dwoz.com/default.asp?Pr=58>

<http://www.spiderhunter.com/>

http://net-promoter.com/robots-txt/spider_list



CALL 1.877.722.7088
1.801.722.0139

www.omniture.com
info@omniture.com

550 East Timpanogos Circle
Orem, Utah 84097

