What's with all those spam ping-bots?

devblogs.microsoft.com/oldnewthing/20080218-01

February 18, 2008

Raymond Chen

Last December, some people started to get annoyed by the pingback-bots, and others were confused by them. What's the deal with those pingback-bots? It's all about fooling the search engines in order to make money, taking advantage of friendly policies at domain registrars to make it less costly an undertaking. Step one: Register a bunch of domains with a domain registrar that includes a money-back guarantee. Step two: Set up fake blogs on each of those sites, with different keywords. Step three: Use a script to search the blogosphere for articles that contain keywords that match your site. (There appears to be a single script that 90% of the spam blogs use, since they all look exactly the same, and have the same bugs!) Step four: Create a bogus blog entry for each one that say something like "Hey, here's something interesting I found on the Internet" and then reprints the article in question. (You may notice that many of these sites mis-attribute the authorship; some of them even claim to have written the article themselves!) Step five: Host ads on the site. Step six: Just before the money-back guarantee period expires, look at each of your fake blogs to see which ones have made money from the ads and which ones haven't. Cancel the domain registrations of the ones that didn't make money. Most of these sites are in existence for only a few days, so trying to stop each individual site is a waste of effort; the site is going away soon anyway. The way to get the attention of the spammers is to hit them in the pocketbook. Go to the site and look at the ads. if they're using Google Ads, look for violations of the terms of service, such as having more than three sets of ads on a single page or hosting ads from other companies on the same page. Even if you can't find anything wrong, click the "Ads by Google" link. From the Google Ads page, click "Send Google your thoughts on the site or the ads you just saw," then "Also report a violation," and then say that you had a problem with "the website," and then say that "The site violates AdSense policies in other ways." Here is where you can write "Hosted more than three ad blocks" or "Also hosts ads from competing vendor." But always write "Contains no original content." The theory here is that once Google has determined that the site is violating AdSense policies, they will shut down the account, preventing them from getting any more money, which was the whole point of their scam in the first place. Now, I don't hold out much hope that this will work, since I've reported sites and found that even weeks later, the site is still up, happily serving up Google ads and pocketing the click-throughs. But maybe it's because they don't act until there is some critical mass of complaints. (I can find no way of reporting violations to the Yahoo Publisher Network.)

Another category of these types of sites is just people who reprint blog articles (usually erroneously attributed) in order to improve the search engine ranking of the non-spam part of the site. Now, you may notice also that there is a "The site is hosting/distributing *my* copyrighted content" checkbox. That box is useless to me because I am not the copyright owner of the content of this blog. The content of this blog is owned by Microsoft Corporation, If you check that box, Google demands that you file a formal DMCA complain, and I'm pretty sure our legal department is busy with plenty of more important things than chasing down people who rip off the content of some random employee's blog in order to generate ad revenue. Normally you don't see the spam pingbacks because I tend to delete them pretty quickly. If you're really clever, you might use the fact that the spam pingbacks linger for days at a time to determine that I'm out of the office. **Sidebar**: Here are some examples of spambots. Feel free to report them to the ad vendor, if they are hosting ads. And as I already noted above, some of these sites may already be down.

- Geek Lectures. (Since this entry was originally written, it appears that the site has taken down all its scraped content.)
- Noticias Externas. This one is particularly annoying since its scraped content sometimes rank higher in Google than my originals!
- D'Technology.
- The Coder Blogs.
- Techy News Blog which freely admits to using blogdigger to decide which articles to scrape.
- One a day vitamin info: trying to drive traffic to their medical devices Web site.
- Fiber Optic Christmas Decorations: This was a classic fly-by-night spam blog. It was gone after Christmas.
- Music News.
- Tattooed Marketer.
- Boxing, Famous Quotes, Famous Birthdays, Actors and Actresses, and more get added every day. This site is relentless; every few hours, another wave of bogus trackbacks comes streaming in. I actually succeeded in getting the ad vendor to stop hosting ads on this spam site. A rare victory. (Sometimes the harvester script goes haywire and you get a messed up page like this one.)
- BioSensorAB, which started up recently but didn't waste any time in spamming viciously.

Update: The victory over 247blogging was short-lived. Within a month, they moved to a new ad company whose terms of service have no problem with sites with no original content. One annoying consequence of all these content-scraping sites is that they end up ranking higher in Google than me, and I'm the one who wrote the article in the first place! For example, a Google search for *Joshua Roman groupies* on 17 February 2008 doesn't even show my blog article; instead, the top hits are

1. A site which scraped my entry.

2. Another page from the same site as #1 which also scraped my entry.
3. A different site which scraped my entry.
4. An article from this Web site but not the one that says *Joshua Roman groupies* in the title.
5. Another misfire from this Web site.
6. A third site which scraped my entry.
7. A fourth site which scraped my entry.
8. A fifth site which scraped my entry.
9. An unrelated hit.
10. Another unrelated hit.

So there you go. The top ten search results contain five sites that scraped my entry and no links to the original! On the other hand, Live Search is not fooled and finds the right article as the top search result. Yahoo ranks my article as #1 and #3 (go figure), which is nice, but all but one of the remaining hits are for scrapers.

A Google search for *bands of Valentine minstrels* is even worse. The first three hits are sites which scraped my article and there are *no hits at all to this Web site* in the top 100 search results, although nine scrapers rank in the top 100. Again, Live Search is not fooled and finds my article as its #1 hit. Yahoo also ranks my article at #1 although a scraper sneaks in at #2.

Raymond Chen

Follow