VOLUME 1 NUMBER 4

TEKKA

enjoying new media • software aesthetics

Spam Will Eat Itself.

(with Mark Bernstein)

http://www.tekka.net/05/?Spam

Two years ago, Norwegian weblog researcher Jill Walker declared that links are the currency of the web (link:**http://huminf.uib.no/~jill/txt/linksandpower.html**). Today, inflation is running high in the Blogosphere. Unlike the commercial web, where links are traded like so many copper planchets, the references in blogs seemed, for a long time, to be the gold standard. Commercial blogging was uncool, and blog posts were not for sale. No weblog author would recommend a site unless they really liked it — perhaps because nobody had yet thought it worthwhile to pay off mere webloggers.

But now not only are weblogs widely read, but bloggers tend to pick up interesting topics from each other. Each additional link increases the chance that other weblogs will also link, and occasionally links and ideas spread overnight through the blogosphere, creating a tidal wave of publicity and buzz that, literally, money can't buy. Because the dominant Google search engine depends on a PageRank algorithm that treats links as a sign of respect and authority, these publicity waves have lasting consequences, even for readers who rarely seek out weblogs. To a retailer or an online casino, traffic is potential money. To a weblog writer, on the other hand, links represent status and approbation. Webloggers treat referrer logs like a clipping service, a public mirror where they can measure their popularity and hear what their rivals are saying about them

Recently, however, a raft of shady enterprises has begun to create bogus referrer logs entries or *referrer spam.* What seem, at first, to be weblogs that link to your writing turn out, instead, to be advertisements for quack medicines or gambling services. As referrer spam proliferates, it drowns out the conversation between weblogs, pollutes the gift economy of the blogosphere and pollutes Google's PageRank. Spam, which has all but killed email, now seems to have blogging by the throat.

However, unlike the floods of unsolicited and unwanted email that bog down servers and carry viruses, link spam does not impose a heavy immediate cost. Each unwanted email message consumes space on disk and requires precious time to evaluate and discard. Unless we ignore all email entirely, we have to accept the burden of coping with unwanted email, and few professionals can afford to simply discard all their email. Link spam, on the other hand, just increments a counter or creates a specious entry in a list. The cost of link spam is the cost of decontaminating a pool of information that once was unpolluted.

Link Spam

When people follow a Web link, their browser software typically sends the destination Web server a small bundle of anonymous information as a sort of calling card. The bundle usually describes the kind of browser that's being used, the operating system that controls the reader's computer, and the Web page that referred the reader to the site. This is, naturally, interesting statistical information; if, one day, I wake to discover that a horde of Linux users is reading my weblog, or that hundreds of new readers are arriving from a Scandinavian newspaper story, I'm bound to find that information intriguing.

Sometimes, of course, logs contain the record of inexplicable accident. A reader might have been catching up on the local news in Bergen or Bangkok, a friend might have stopped by and told them about your site, and they might have typed in the URL. Or maybe the friend told them about another site entirely, and they mistyped your URL. These things happen. But, if people can read Web pages, so can computers — and computers can be set up to pretend to be real readers, and to lie. Instead of sending real information, they leave false calling cards intended to trick people into viewing ads.

Similarly, computers may post specious comments on bulletin boards, extolling irrelevant businesses. Many bloggers use comments to pull the discussion sparked by their posts from the of the web onto their own pages. And while comments are rarely displayed on a blog's first page, they are still widely read by people and by search engines.

What‘s in a Link?

The Internet is a vast repository of information, and links are at once instrumental and significant — they provide access and supply context at the same time. Without the link, the Web wouldn't be the Web. In the end, too, inbound links are what distinguish an isolated file from an organic part of the world-wide Web; if you aren't connected, no one can find you. "Links have a direct value on the Web," Walker explains in her paper. They "can be seen as a pseudo-monetary unit ... and are a sign of peer-endorsement."

The real currency of the commercial internet is, of course, the click, the visit. Website popularity translates into notoriety, fame, and sometimes cash. For a Web site to be popular, it must first be discovered, and to discover it, we usually either follow a link or find it in a search engine. The more people link to your page, the more opportunities exist for people to find you. And, to complicate matters even more, the more people link to a page, the more prominent it's likely to be in search engine listings.

Early search engines simply searched for key words, and returned huge lists of pages jumbled together. As the Web grew, this became less and less satisfactory, and search engines tried to find ways to place the best, most relevant, Web pages at the top of the list. This led, in turn, to a cottage industry devoted to fooling the search engines, as individuals and businesses all sought to propel their own sites to the top of each search engine category. One way to do this, of course, is simply to make a superb site. Another route, however, is to fool the search engine; because no company can afford to pay professionals to read and evaluate everything on the Web, fooling the search engine computers is rarely difficult.

Google, at present the most popular search engine, uses an algorithm called PageRank to identify Web sites that appear to be authorities — sites to which numerous other sites link. This makes sense; if many writers all cite the same reference, that reference is more likely to be sound than another reference that is rarely discussed. It's also a fair and objective standard, since it doesn't inherently favor a wealthy corporation over an unknown student, nor does it accord more prominence to a research university than to a rural school. Crucially, PageRank requires neither trust nor private data, for it need not look at who reads a page, but only at how many paths to a page it discovers.

But the strengths of PageRank can be turned against it, too. If inbound links improve PageRank, then Web writers will seek inbound links, and inevitably some will resort to doubtful strategies. Some dispatch millions of email notes, pleading for people to link to them. Some employ hack writers to grind out vast streams of prose, simply to have a place in which they can refer to their pet projects. Some create vast numbers of Web pages that contain nothing but links — pages that are not meant to be read, only to fool the Google engine. And some, in turn, start businesses that try to sell these dubious services to others.

Why Spam Weblogs?

So why do link spammers target Weblogs? While hard data are hard to find, spot checks suggest that referrer spam concentrates on weblogs and that commercial sites are bothered less often. One explanation, of course, might be scale: a few dozen hits might represent a significant event in a personal weblog — an event that would merit investigation — while the same number of hits would go unnoticed by a large news site or a major retailer. The spam computers could generate

even more traffic — enough to be noticed even by big corporate sites — but that much traffic would swamp the smaller servers that personal sites favor. But the most important reason for spamming weblogs is simply the close relationship between weblogs and Google.

The first wave of coordinated referrer spam came in summer 2003, as ordinary blogs suddenly seemed to receive floods of visitors from porn sites, online casinos, and fly-by-night phone operators. Curious weblog writers visited the advertisers, generating traffic. Features like comments and trackback offered new opportunities for spammers to sneak links into unsuspecting weblogs. Those links, in turn, were seen by Google and improved the spammer's rank in search engine listings.

At first, it seemed that link spam undermined the entire value system of the blogosphere. In practice, though, the impact has been softened by the media literacy of weblog readers and writers. As link spam proliferated, everyone learnt to ignore apparent links from dubious sources. Some began to check their logs less frequently. Some took measures to penalize attempts to plant bogus links, blocking the computers of known spammers or deleting their links whenever they appeared. New technologies require commenters to respond to email, making it harder for spam computers to pretend to be real people. The wave subsided quickly, without much damage.

The second wave of referrer spam came in fall 2003 and was rather more perfidious than its predecessor: the faked links came from pages that looked like weblogs. One could find one's own blog listed in their blogrolls, and they had real contents, albeit somewhat dull and mechanical. Almost all of the links on these sites, and especially those in the would-be "blogrolls", linked back on themselves. The only working links went to Casino or sex sites and were tucked away at the bottom of the page. Adam Gessaman, author of the weblog **idly.org**, offers a plausible explanation: this set of synthetic weblogs was a front, intended only to create an apparent community of weblogs that would fool Google into thinking that a great many individuals had developed a sudden passion for discussing a particular casino or sex site.

Spam is for Those Who Deserve it.

In a recent letter to the editor of the German magazine *Internet Professionell*, a reader insinuated that spam is for those who deserve it. When some people choose to click the links in spam, obviously this decision must stem from a consumer desire for spam.

Comment spam poses a subtler problem. Some spammers, for example, employ computers to post millions of irrelevant links in the midst of any open discussion, and this is clearly pernicious. But some real individuals do much the same thing: we all know people who talk too much about themselves, who are inclined to bring their enthusiasms or their businesses into any discussion, who will extol their new line of insurance or their latest religious experience to strangers at a party.

Drawing the line in a public discussion between spam robots and those who are simply annoying people may be difficult.

Thus far, weblog writers have met the challenge of spam. Bloggers warn each other, search out the owners of the fake weblog sites and block (or filter) their referrers. They have used the tool they know best: information-sharing. Perhaps the responsible use of complex communication technology needs to be taught at an early age, like the need to look left and right before crossing the street. Or perhaps kids will assimilate strategies for recognizing link spam, just as they learn to recognize an unfashionable brand of sneaker — or resist the lure of' email messages that ask for URGENT ASSISTANCE or that offer to enlarge body parts the recipients do not even possess.

Sponsored by Eastgate

Tekka, 134 Main Street, Watertown MA 02472 USA. email: **editor@tekka.net** **info@tekka.net**