

Spam Double-Funnel: Connecting Web Spammers with Advertisers

HY558 – Σοφία Λούτου

Spammers use questionable search engine optimization (SEO) techniques in order to promote their spam links into top search results. The paper focuses on one type of spam, redirection spam. It proposes a 5 layer, double-funnel model for describing end-to-end redirection spam, presents a methodology for analyzing the layers and identify prominent domains on each layer using two sets of commercial keywords – one targeting spammers and one targeting advertisers.

Popular techniques used by spammers include SEO techniques like stuffing keywords, link farms, comment spamming, click-through cloaking. Redirection spam describes web pages that redirect browsers to visit known spammer-controlled third-party domains. Many redirection spam pages use syndication where they participate in pay-per-click programs and display ads-portal pages.

The authors use the Strider Search Ranger system to analyze tens of thousands of spam links that appeared in top results across three major search engines in order to identify the major domains in each of the 5 layers and their interesting characteristics.

The automated spam detection system has the following 3 key features:

1) Web Patrol with Search Monkeys: To defend against crawler-browser cloaking techniques Search Monkeys visit each web page with a full-fledged popular browser which executes all client-side scripts and mimic the click-through as newer cloaking techniques serve spam content only to users who click through search results.

2) Follow the money through Redirection Tracking: In order to identify who is behind spam activities they use the Strider URL Tracer to intercept browser redirection traffic at the network layer to record all redirection URLs at both ads-fetching traffic and ads click-through traffic

3) Similarity-based Grouping for Identifying Large-scale Spam: Rather than analyzing all crawler-indexed pages, they focus on monitoring search results of popular queries targeted by spammers to obtain a list of URLs with high spam densities. By analyzing the similarity between the redirections from these pages they identify doorway pages

and use them to perform “backward propagation of distrust” to detect other related spam pages.

A typical syndication business

It consists of three layers: The publishers who attract traffic by providing quality content on their websites to achieve high search rankings, the advertisers who pay for displaying their ads on those websites and the syndicators who provide the advertising infrastructure to connect the two others (ex: Google AdSense program). Spammers assume the role of publishers and set up low-quality websites and use SEO techniques to attract traffic. To survive spam detection spammers use doorway pages (layer 1) and redirection domains (layer 2). To attract prudent legitimate advertisers syndicators do not want to be connected to spammers and through multiple redirections they obfuscate the connection between advertisers and spammers. The end-to end spamming business is as follows: Advertisers (layer 5) pay syndicators (layer 4) to display their ads. Syndicators buy traffic from aggregators (Layer 3) who buy traffic from web spammers to insulate syndicators and advertisers from spam pages. Spammers set up redirection domains and doorway pages. If any such URLs are promoted into top search results and clicked by users, all click-through traffic is funneled back to syndicators.

As mentioned before two benchmarks are used. The first one is for the most spammed keywords at public forums. However the primary concern of most search users and legitimate advertisers is the impact of such spam on the quality of their query results. So a benchmark based on the most-bid keywords from legitimate advertisers is also studied.

As far as the first benchmark is concerned, redirection spammers often use their targeted keywords as the anchor text of their spam links at public forums. To collect them the authors extract all the anchor text from a large number of spammed forums. Drugs and ringtones are dominating the list. The second benchmark has fewer keywords from the drugs, adult and gambling categories and more keywords from the money category, indicating the difference at the two benchmarks.

Useful results concerning the most spam-heavy sites are presented and sites that are usually used as doorways (like blogspot.com and .info sites that come with the highest ranking at this category).

As far as the 2nd layer is concerned (redirection domains) the measurements reveal that most of them (for the 2nd benchmark it's all of them) are syndication-based. Furthermore there are cases where redirection domains share the same proxy registrant and reside on the same IP block.

For the bottom three layers there are two types of analysis. Page analysis for layers 3 and 5 where target advertiser URLs are extracted as well as their associated click-through URLs from ads-portal pages without visiting the ads. Click-through analysis for layer 4, where one ad from each portal page is randomly selected and visited and all the resulting redirection traffic is recorded. This appeared to be necessary because the domain names of intermediate syndicators did not appear in the content of ads-portal pages.

As far as layer 3 is concerned (aggregators), the top-15 click-through traffic receiver domains can be grouped in two groups: 66.230.128.0-66.230.191.255 and 64.111.192.0-64.111.223.255 for both benchmarks.

For layer 5 (advertisers), on most spam ads, the click-through URLs did not contain the plaintext URLs of their target advertisers but their domain names were appeared as anchor text or in the status bar upon mouse-over. Their ranking based on the number of appearances gave only 6 overlap indicating the difference approaches of the two benchmarks.

Finally, for layer 4 (syndicators) it appears that a handful of syndicator domains have significant presence in the redirection chains involving in the search spam industry both broadly and deeply.

Other common spam are blog farms and parasite ads-portal farms and it appears that they share the same bottom half of the double-funnel with redirection spammers.

Concluding the methodology and findings of this paper are useful for search engines to strengthen their ranking algorithms against spam, for legitimate website owners to locate and remove spam doorway pages and for legitimate advertisers to identify unscrupulous syndicators who serve ads on spam pages.