Towards Spam Detection at Ping Servers [*]

Pranam Kolari, Tim Finin, Akshay Java, Anupam Joshi
Department of Computer Science and Engineering
University of Maryland, Baltimore County
Baltimore, MD 21250, USA
{kolari1, finin, aks1, joshi}@umbc.edu

Abstract

Spam blogs, or splogs feature plagiarized or auto-generated content. They create link farms to promote affiliates, and are motivated by the profitability of hosting ads. Splogs infiltrate the blogosphere at ping servers, systems that aggregate blog update pings. Over the past year, our work has focused on detecting and eliminating splogs. As techniques used by spammers have evolved, we have learned how splog signatures are tied to tools that create them, that they are beginning to be a problem across languages, and that they require a much quicker assessment. Though we continue to address these specific challenges, we discuss our larger goal in this work, of developing a scalable meta-ping filter that detects and eliminates update pings from splogs. This will considerably reduce computational requirements and manual efforts at downstream services (search engines) and involve the community in detecting spam blogs.

1. Motivation

The Blogosphere is growing and so is the Splogosphere. Even though estimates vary, spam ranges from as high as 75% [5] at ping servers, to around 20% [9] at blog search engines, with similar high values at web search engines [2]. Multiple factors have led to the use of blogs by spammers. Just to name a few, this includes (i) the higher relevance web search engines associate with blogs (ii) an infrastructure of ping servers that enables quick notification to search engines, and (iii) the availability of third party services that host blogs for free. Though the primary targets of spam blogs are web search engines, they present severe computational overhead and skew analysis at systems that index and analyze (only) blogs.

To protect their index from spam, search engines (both web and blog) have fairly sophisticated filters in place. Before indexing content, they filter updated blogs using blacklists, regular expressions, or models based on local content of blogs. Splogs that escape these pre-indexing filters are subsequently eliminated using semi-automatic or more complex detection techniques based on the link graph. The quality of these search engines (indexing services) is judged by how well and early they eliminate splogs, both pre-indexing and post-indexing.

Splog detection is however highly fragmented across these systems. This has several drawbacks that include [3] (i) repetitive detection efforts, (ii) unnecessary network and computational overheads, and (iii) high barrier of entry for new indexing services. We hence propose infrastructure insertion on the blogosphere, one that eliminates splogs pre-indexing; by developing an effective intermediary server that also encourages community participation. We call it a *meta-ping* server. Such a system can either work together with a ping server, or independently, feeding either blacklists or filtered blog updates to search engines.

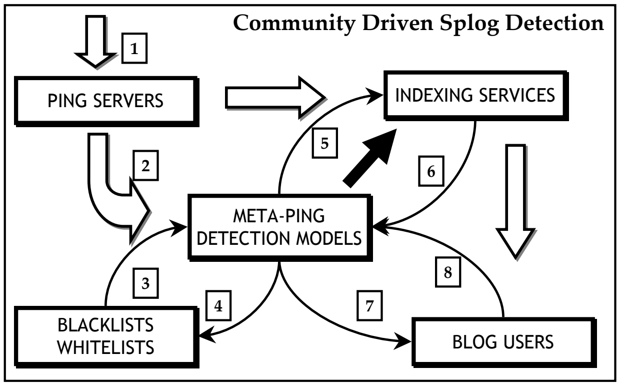

Fig. 1: *The overall architecture for a meta-ping server. Pings (1) from existing ping servers are aggregated (2), followed by splog detection using models, and blacklists (3). Either blacklists or filtered update pings are made available to indexing services (5). Indexing servers also perform their own spam detection. APIs provided for integration and feedback from user blog tools (7), (8), and possibly indexing services (6).*

A meta-ping server of this kind should satisfy certain basic properties including the following. (i) **Scalable**. Ping servers have scaled well[1] because of their simple and almost stateless operation across pings. The proposed solution should operate as independently across pings as possible. (ii) **Fast**. Indexing services are judged by how quickly they can index new blog content. The solution should hence introduce the minimal amount of delay, by minimizing the number of web fetches required to identify splogs. (iii) **Adaptive**. Just as with any form of spam, eliminating splogs is an adversarial classification problem between filters and spammers. The solution should hence be adaptive, built on detection models that use new training sets and features, gathered by effective community contribution. (iv) **Effective**. Since a meta-ping server is a pre-indexing solution, when detecting splogs it should feature very high precision, with good recall. In addition, detection of splogs should be early in their lifecycle,

[*] Partial support was provided by an IBM CAS Fellowship and by NSF awards NSF-ITR-IIS00326460 and NSF-ITR-IDM-0219649.

[1] http://weblogs.com now accepts 200K pings per hour

measured using a combination of number of posts, and time-stamp of last post.

Our overall architecture is shown in figure 1. By encouraging the involvement of users and blog search engines, and by publishing blacklists, a collective community-effort driven better tackles spam. To enable user participation, we are developing tools and plug-ins for popular blog publishing software. In what follows we discuss the feasibility of our broad goal, by identifying existing useful splog detection techniques that fit into the framework.

2. Approach

An update ping received by a ping-server[2] consists of *URL* of the updated blog, its *name*, its *time-stamp*, and could also feature URL of the *RSS* feed.

The detection modules of the meta-ping server that taps into these pings will work in the following steps, publishing intermediary results, and accruing more evidence until sufficient confidence on a classification is attained, (i) URL based filtering, (ii) blacklist based filtering, (iii) blog home-page based filtering, and (iv) feed based filtering. Steps (iii) and (iv) are language dependent, and models are to be employed, based on their language independent nature. Researchers, and practitioners working on detecting splogs have addressed each of these steps separately and shown their feasibility. We believe combining these approaches using a unifying theme of a meta-ping server can be very useful for the blogosphere, as a whole. We briefly discuss each of these steps.

2.1 URL Based Filtering

URL filtering is characterized by requiring no web fetches making it very fast. Informally, combining multiple heuristics can be effective: the length of URL (spam ping URLs are typically long), non-alphabetic characters (hyphens, forward-slashes), and in some cases completely blacklisting URLs from the *info* domain[3] have worked in the past. A more formal analysis made by Salvetti et al [8] introduces new techniques for URL segmentation and detection based on tokens in the URL. The resulting precision was 93.3% with a recall of 50.9%. In addition *name* field and how it correlates to the URL can also be used to support classification, though this is yet to be addressed.

2.2 Blacklist Based Filtering

Developing a catalog of spam domains, IP addresses, as well as one for authentic blogs also forms a core component of the meta-ping server. The approach of using blacklisted IP addresses [1] has been found effective previously. This requires only a DNS query to find IP address associated with the URL. Additionally lists of authentic blogs can also be used. Many such lists are also available online. The proposed meta-ping server will incorporate all of this available background information, and evolve it.

2.3 Blog Home-Page Based Filtering

The third step in filtering is based on analyzing contents of the blog-homepage. Though this step involves a web-fetch per URL, it offers many advantages. (i) It enables the identification of the ping source as a blog or non-blog. Ping servers are easy inlets to indices of search engines. Spammers exploit this to unjustifiably notify the existence of non-blog pages. Our prior work shows that non-blogs can be easily identified through models that are language independent [4]. (ii) It enables identifying the blog's language. This facilitates the use of language independent or language dependent models. (iii) It enables the use of existing splog detection models [6] (87% precision and recall) over blog home-pages. This model is fairly accurate since it captures recently made posts with link-rolls and blog-rolls.

Our tools are in use by our academic and industrial collaborators. We have drawn on these experiences and are addressing certain specific concerns. One such step is tackling feed based spam detection and complementing it with blog home-page based detection, which we discuss next.

2.4 Syndication Feed Based Filtering

In addition to developing models over text and hyperlinks in a blog, temporal correlations across posts can also be effective. This requires the use of structured content, available through syndication feeds, and involves two web-fetches per URL. Detection approaches are now beginning to explore correlations across posts [7] in a language independent manner. We believe this will complement the previous step well, reinforcing evidence on the authenticity of a blog. The structured nature of feeds also enables evaluating effectiveness temporally i.e. how early is splog detection possible, in terms of number of posts and lifetime of the blog.

We have bootstrapped the entire process by analyzing detection approaches that currently exist, and how they fit into the overall system. We have also generated new training sets using which models can be built for classification. We plan to make part of our system public in late January.

References

[1] S. Critchley. Suss out spam networks (in comments), 2006. [Online; `http://spamhuntress.com/2005/12/31/suss-out-spam-networks/#comment-2229`].

[2] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[3] H. Kaushansky. Marketer, Beware: The Threat of Blog Spam (Splogs) to WOM Marketing and Market Insight, 2006. [Online; `http://www.womma.org/research/studies/howard_kaushans.htm`].

[4] P. Kolari, T. Finin, and A. Joshi. SVMs for the Blogosphere: Blog Identification and Splog Detection. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[5] P. Kolari, A. Java, and T. Finin. Characterizing the splogosphere. In *WWW 2006, 3rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[6] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi. Detecting Spam Blogs: A Machine Learning Approach. 2006. Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006).

[7] Y.-R. Lin, W.-Y. Chen, X. Shi, R. Sia, X. Song, Y. Chi, K. Hino, H. Sundaram, J. Tatemura, and B. Tseng. The Splog Detection Task and a Solution Based on Temporal and Link Properties. In *TREC Blog Track*, 2006.

[8] F. Salvetti and N. Nicolov. Weblog Classification for Fast Splog Filtering: A URL Language Model Segmentation Approach. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 137–140, New York City, USA, June 2006. Association for Computational Linguistics.

[9] Umbria. Spam in the blogosphere, 2005. [Online; `http://www.umbrialistens.com/consumer/showWhitePaper`].

[2] http://weblogs.com/api.html

[3] http://memeta.umbc.edu/ping/info/