Characterizing the Splogosphere*

Pranam Kolari and Akshay Java and Tim Finin
University of Maryland Baltimore County
1000 Hilltop Circle
Baltimore, MD, USA
{kolari1, aks1, finin}@cs.umbc.edu

ABSTRACT

Weblogs or blogs collectively constitute the *Blogosphere*, forming an influential and interesting subset on the Web. As with most Internet-enabled applications, the ease of content creation and distribution makes the blogosphere spam prone. Spam blogs or splogs are blogs hosting spam posts, created using machine generated or hijacked content for the sole purpose of hosting ads or raising the PageRank of target sites. These splogs make up the *splogosphere*, and are now inundating blog search engines and update ping servers. In this work we characterize splogs by comparing them against authentic blogs. Our analysis is based on a dataset made publicly available by BlogPulse, and employs a machine learning model that detects splogs with an accuracy of 90%. To round off this analysis and to better understand splogs, we also present our study of a popular blog update ping server, and show how they are overwhelmed by pings sent by splogs. This overall study will facilitate finding effective new techniques to detect and weed out splogs from the blogosphere.

1. INTRODUCTION

Weblogs or blogs are web sites consisting of dated entries typically listed in reverse chronological order on a single page. Based on the nature of these entries, blogs are considered to be one of personal journals, market or product commentaries, or just filters that discuss current affairs reported elsewhere, participating in an online dialogue. While traditional search engines continue to discover and index blogs, the Blogosphere has produced custom blog search and analysis engines, systems that employ specialized information retrieval techniques. As the Blogosphere continues to grow, several capabilities have become critical for blog search engines. The first is the ability to recognize blog sites, understand their structure, identify constituent parts and extract relevant metadata. A second is to robustly detect and eliminate spam blogs (splogs).

Splogs are generated with two often overlapping motives. The first is the creation of fake blogs, containing gibberish or hijacked content from other blogs and news sources with the sole purpose of hosting profitable context based advertisements. The second, and better understood form, is to create false blogs, that realize a link farm [19] intended to unjustifiably increase the ranking of affiliated sites. The urgency in culling out splogs has become all the more important in the last year, evident from the frequent discussions and reports [18, 3, 17, 16] on this issue.

The general problem of spam is not new to Internet based applications. The ease of content creation (and plagiarism) and distribution has made the Internet a haven for spammers. While past research has shown that spam can be controlled on the WWW [10, 6, 9] and electronic mail applications [8, 20], spam in blogs is not all that well studied. Blogs and the blogosphere in general have certain unique characteristics: (i) they are freely hosted by blog hosts, (ii) they provide content syndication for distribution, (iii) they support remote web service APIs for publishing, and (iv) provide update ping servers to notify search engines. We first present a characterization of *splogs vs. blogs* to highlight the discriminating features. We then formally report on our previous analysis of blog ping servers [11], which makes some rather disturbing conclusions on spam faced by update ping servers. This overall analysis will aid in better understanding the domain, and to develop useful new splog detection techniques.

The rest of the paper is organized as follows. Section 2 reports on our methodology and provides background to this work. Section 3 reports on our characterization of splogs vs. blogs. In section 4 we detail our analysis of a blog ping server. Finally we discuss the implication of our results in Section 5, which can be useful in developing new techniques for splog detection.

2. BACKGROUND

In this section we provide background to this work by introducing the BlogPulse dataset, blog update ping servers and the pings dataset, and our working splog detection system.

2.1 BlogPulse Dataset

BlogPulse[1], a popular blog search and mining system, recently released a dataset spanning a period of 21 days in July of 2005. This dataset consists of around 1.3 million blogs including additional metadata about them. To enable better understanding of our results, we base a large part of our analysis on this dataset. The relative frequency of various blog hosts in this dataset is shown in Figure 1.

*This work is supported by NSF Awards NSF-ITR-IIS-0326460 and NSF-ITR-IDM-0219649

WWW2006, May 22–26, 2006, Edinburgh, UK.

[1]http://blogpulse.com/

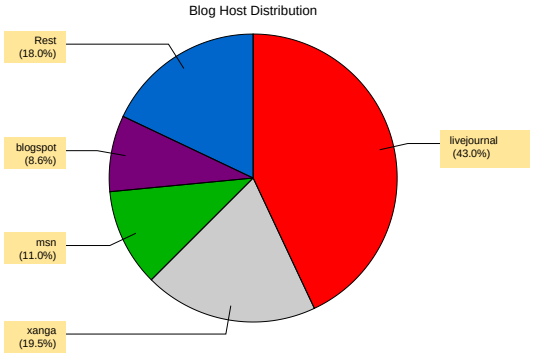

Figure 1: Blog host distribution in the BlogPulse dataset.

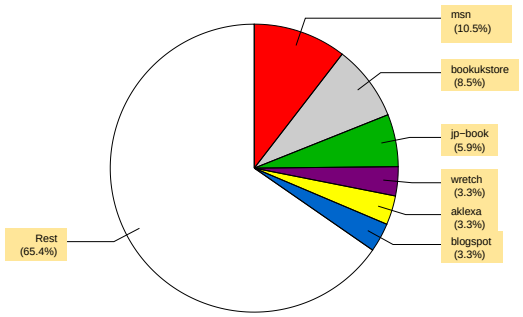

Figure 2: Host distribution of pings received by an Update Ping Server.

2.2 Blog Ping Servers

Blogs notify update ping servers when new posts are made; these servers then route such pings to systems that index and analyze blog content. Independent from the BlogPulse dataset we also analyzed pings received by a popular blog update ping server[2] that makes these pings public. We analyzed around 15 million pings over a period of 20 days from November 20, 2005 to December 11, 2005 to check how many of these are spings, i.e. from splogs. The relative frequency of hosts pinging the update server is shown in Figure 2.

2.3 Splog Detection

We have developed splog detection models [12][13] trained using Support Vector Machines [7]. These models are based on logistic regression and can hence predict probabilities of class membership. We currently have two separate models, one that identifies blogs on the Web, and the other that detects splogs amongst the identified blogs. These models are constructed on blog home pages and can potentially make use of a combination of various features including textual content, anchor text, urls and n-gram words to identify blogs and detect splogs. The *F1* measure for blog identification is around 97% and that for splog detection is around 90%.

Our working splog detection system employs a multi-step approach to detect splogs, filtering out and eliminating URLs as it goes through steps: (i) from known blacklists, (ii) in languages other than English[3], (iii) are identified as non-blogs i.e a web-page in general, and (iv) are detected as splogs. Steps (iii) and (iv) can currently be employed only on blogs in the English language. So we allow only English blogs to pass through the language detection module that employs the libsvm[2] toolkit for SVMs. Blogs that pass through all four filters are tagged as authentic English blogs with a confidence provided by associated logistic regression based probability.

[2]http://weblogs.com/

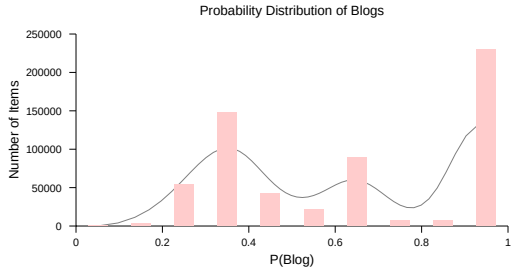

Figure 3: Probability Distribution of Blogs.

3. BLOGOSPHERE VS. SPLOGOSPHERE

Our detection modules are based on analyzing the complete structure of blog home-pages and not just individual posts. Such an approach captures interesting features common to multiple posts on a blog home-page and also uses other information like blogrolls and non-post out-links before making a splog judgement. To adhere to this requirement, we extracted blog home page URLs from the BlogPulse dataset, and re-fetched the complete home-pages to analyze their content. It turns out that many of these home-pages (possibly splogs) are now non-existent either because they were detected and eliminated by blog hosts or pulled down by spammers as they were no longer useful. The number of failed blogs was as high as around 200K. Since we are not in a position to ascertain the true nature of these failed URLs with a high confidence we dropped them from consideration.

Of the remaining blog home pages we noticed that live-journal had an insignificant percentage of spam blogs[4]. Given that live-journal forms a large fraction of authentic blogs in the dataset we eliminated all blogs from this domain and worked with blogs from other domains and self-hosted blogs. The primary reason was to eliminate the characteristics of live-journal blogs biasing our results.

After filtering out the above mentioned blogs, and blogs that are not in English we ended up with around 500K blogs. The probability distribution provided by our blog identification module is shown in Figure 3, and the distribution of splogs returned by the splog detection module is shown in Figure 4. Each bar on the x-axis represents a probability range and values on the y-axis represent the number of pages (blogs) that are within this range.

[3]Provided by James Mayfield

[4]This need not necessarily hold for blogs created as of March 2006

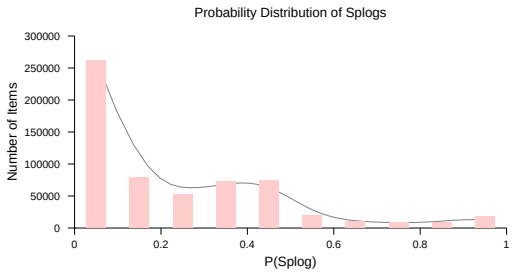

Figure 4: Probability Distribution of Splogs.

Typically, we use results from blog identification to feed into splog detection. However we ignored probability distribution of blogs and made an assumption that all blogs in the BlogPulse dataset are truly blogs. We then used the following thresholds from the splog detection module to create subsets of authentic blogs and splogs used in our characterization.

$$X \in AuthenticBlog,\ if\ P(X = Splog/Features(X)) < 0.25$$

$$X \in Splog,\ if\ P(X = Splog/Features(X)) > 0.8$$

In these two created subsets, the cardinality of the splog subset was around 27K. We uniformly sampled for 27K authentic blogs to have two subsets of the same cardinality. In what follows, our comparative characterization is based on 27K splogs and 27K blogs.

3.1 Frequency of Words

We first analyzed the distribution of certain discriminating terms in both blogs and splogs. Since our splog detection module is built using only local features, it is these discriminating features that were employed by our system. We created a ranking of features based on weights assigned to the features by the SVM model. This list consists of 200 word features common to blogs and 200 word features common to splogs. The word features common to blogs included pronouns like "I", "We", "'My" and words from anchor text to popular websites like flickr, Technorati etc, which were all less common in splogs. Splogs generally feature high paying adsense[5] keywords.

The occurrence based distribution of terms common in blogs and splogs for these top features is shown in Figure 5. The first half on the x-axis depicts the top blog features and the second half depicts the top splog features. The y-axis represents the difference between the number of blogs in which the feature occurred to the number of splogs in which the same feature occurs. Clearly, the top blog features occur more frequently in blogs than splogs and vice-versa. Similar patterns can be observed in a comparison using 2-gram words and 3-gram words [5], and models based on such local knowledge give detection F1 estimates of close to 90%.

3.2 Link Structure

Splogs that escape existing filters engage in creating link-farms to increase the importance of pages in the farm, scores computed using PageRank[15]. The distribution of inlinks

[5]http://google.com/adsense

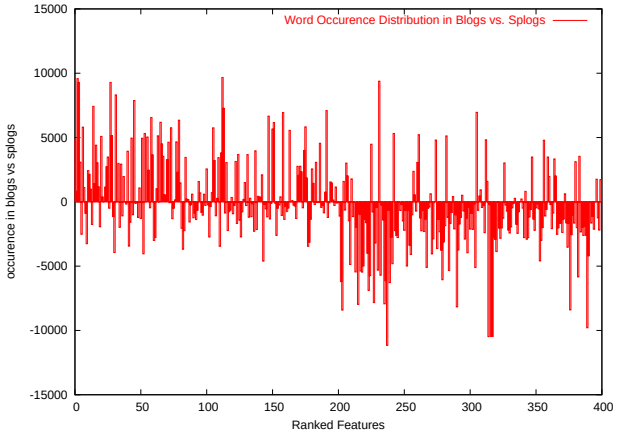

Figure 5: Distribution of top discriminating word-features in blogs and splogs.

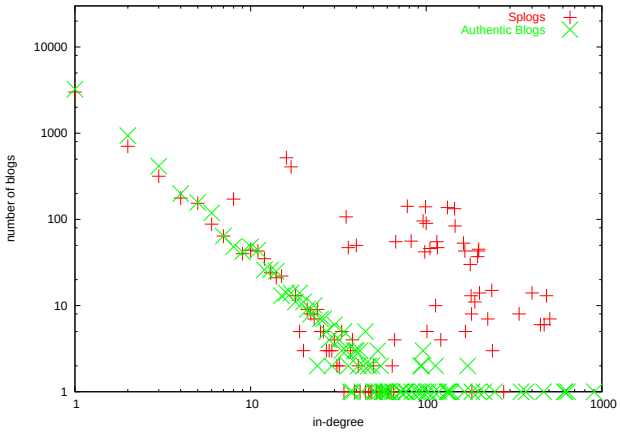

Figure 6: In-degree distribution of authentic blogs subscribe to a power-law.

for splogs and authentic blogs is shown in figure 6, where the link graph was obtained from the weblogs dataset. Blogs show a power-law that is typical to the Web in general[1], whereas splogs deviate from this norm. We also followed this up by checking for outlink distribution of splogs and blogs. Figure 7 shows this distribution, with blogs complying with the power-law as opposed to splogs which does not adhere to it.

Since post time-stamps in the BlogPulse dataset are not normalized across blogs, we do not make an analysis of post time stamps here. Any such analysis will be similar to that put forward in our next section on spings. In addition to these characteristics, we also noticed certain patterns in other aspects of splogs. For instance, from the tagging perspective most of the splogs are tagged as "un-categorized". However all these discriminating features are incorporated in the word characterization discussed earlier, which incorporates all of the text (including anchor-text) on a blog.

Based on these results, and a related analysis [13], we make the following observations:

- Given the current nature of splogs, their detection is

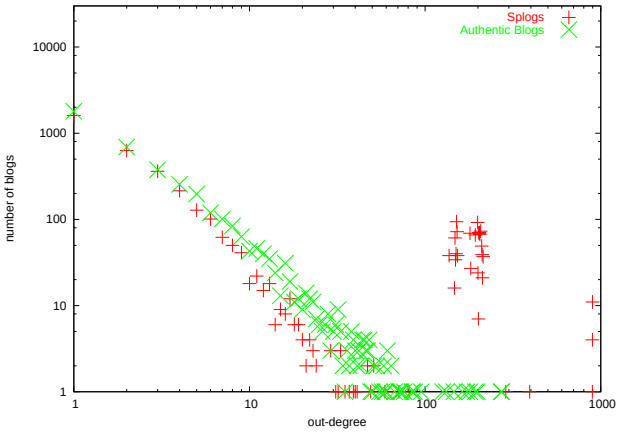

Figure 7: Out-degree distribution of authentic blogs subscribe to a power-law.

quite effective through the use of only local features. Word model of blogs based on local features create an interesting "authentic blog genre" that separate them from splogs.

- If splogs do happen to escape filters and then indulge in the creation of link-farms, many of them can be detected using spam detection algorithms on the web graph [20]. However, this approach taken alone has two disadvantages. First, it allows splogs to thrive in blog hosts and search engines for a longer period of time, and second, it fails to detect splogs which are not part of abnormal link sub-structures.

4. SPLOGS AND PING SERVERS

Ping Servers define standard interfaces that can be used by blogs to notify new (or updated) posts. Information about the blog home-page and blog title[6] typically accompany these pings. Additional information like syndication feed location can also be specified, but is less common. Other than restrictions on their frequency, no other restriction is usually placed on pings. Driven by this restriction-free nature, and the improved search engine exposure (both blog search and web search) ping servers provide, splogs overwhelm ping servers.

Ping Servers are faced with two kinds of spam - (i) pings from non-blogs, and (ii) pings from splogs, both of which are referred to as spings. We used a similar approach to the one we used for splog detection in the BlogPulse dataset. However to scale up to the number of pings that have to be processed, we used simpler techniques and made some exceptions. We used URL based heuristics for blog identification and did not pass pings from the info domain through our filters. However for all other pings, we fetched the home-pages of pings to make a splog judgment. We also identified pings from different languages to work with splogs in the English language. Additionally, unlike the thresholds used on the BlogPulse dataset, we used less stricter thresholds.

$$X \in Blog,\ if\ P(X = Splog/Features(X)) < 0.5$$

[6]http://www.weblogs.com/api.html

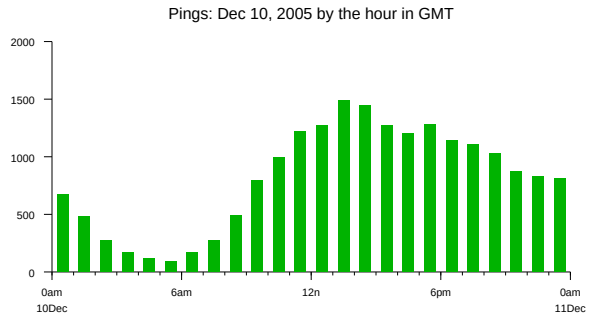

Figure 8: Ping Time Series of Italian Blogs on a single day.

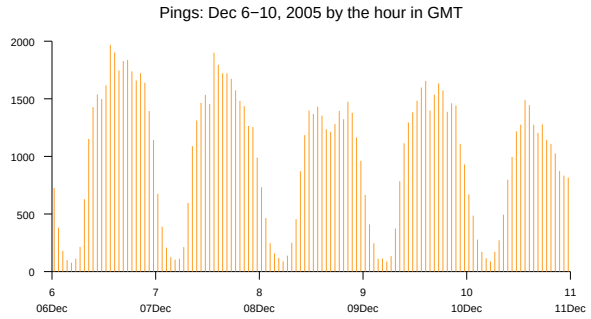

Figure 9: Ping Time Series of Italian Blogs over five days.

$$X \in Splog,\ if\ P(X = Splog/Features(X)) >= 0.5$$

Figure 8 shows the ping distribution from blogs (around 50K) in Italian. All times are in GMT, and each bar accounts for total pings in an hour. Similarly figure 9 shows these pings distributed over five days, with each line accounting for an hour of pings. These distributions make it quite evident that blogs written in Italian language show an interesting posting pattern, higher during the day and peaking during mid-day. We observed similar patterns with many other languages[7] that are restricted to specific geographic locations, and time zones. Though our splog detection system is currently not capable of splog detection in these other languages, these charts do show that blogs in non-english languages are less prone to splogs.

Figure 10 shows the ping distribution from authentic blogs on a single day and figure 11 shows it across five days. Unlike ping distribution of blogs in Italian, blogs in English do not show well formed peaks. We attribute this to English being commonly used across multiple geographical locations/time-zones. However, pings from English blogs are relatively higher during the day-time in US time-zones, where blog adoption is relatively higher.

[7]See http://memeta.umbc.edu

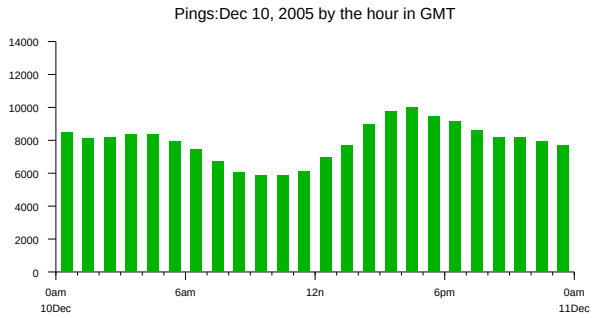

Figure 10: Ping Time Series of Blogs on a single day.

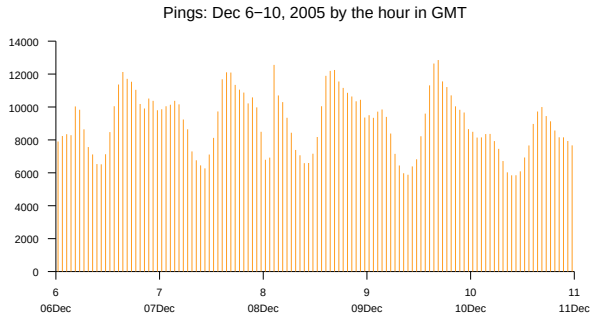

Figure 11: Ping Time Series of Blogs over five days.

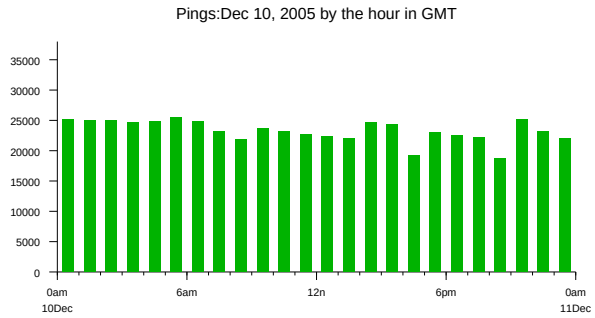

Figure 12: Ping Time Series of Splogs on a single day.

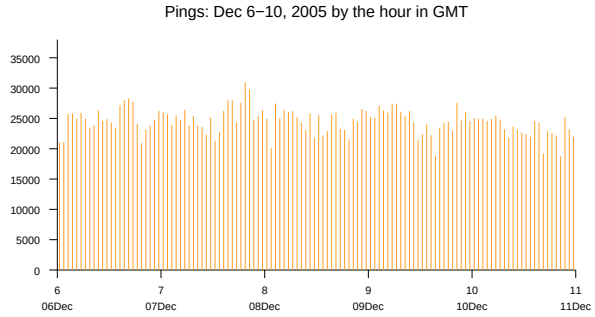

Figure 13: Ping Time Series of Splogs over a five day period.

In comparison with pings from authentic blogs in English, Figure 12 shows the ping distribution from splogs on a single day, and figure 13 shows it across five days. Two characteristics make this interesting. First, splog pings do not show any patterns that are associated with typical blog posting times. Second, the number of spings are approximately three times the number of authentic pings suggesting that around 75% of pings from English Blogs are from splogs.

As mentioned earlier, to make our splog detection system scale up with pings, we did not pass pings from info domains through our filters, other than tagging these pings for later analysis. Figure 14 shows the ping distribution from the info domain across five days. Clearly, there is no pattern in the posting time-series; we also observed a sudden increase in pings from this domain around Dec 11, 2005 without any evident explanation. This continued for the next five days beyond which we stopped monitoring ping servers. We believe that info domains are highly sploggy as well.

Finally, Figure 15 shows the nature of URLs (as encoded in the home-page field) pinging weblogs.com and the percentage of all the pinging URLs they constitute over the entire 20 day period. This graph makes even more disturbing conclusions, the number of splogs constitute around 56% of all pinging URLs (blog home-pages) in English whereas those from authentic English blogs is only around 7%. This implies that around 88% of all pinging URL's in English are splogs. Based on our analysis of ping servers, we make the following observations:

- Even though splogs constitute around 88% of all pinging URLs, they account for only 75% of all pings. This is attributed to the fact that many splog pings are one-time pings. The same URL is not used in subsequent pings. Such pings specify arbitrary pages as blog home-pages even though they have no relationship with blogs or the blogosphere.

- Many of the URLs are from non-existent blogs, i.e., they constitute failed URLs. They constitute what could be termed as zombie pings, spings that exist even though the splog (or page) they represent is non-existent (or is already eliminated) in the blogosphere.

- Most of the popular web search engines give particular importance to the URL tokens of page. In addition to checking if page content matches a particular query they also check if URL text has similarities. Splogs exploit this ranking criteria by hosting blogs in the *info* domain, where domain registrations are less expensive and easily available, as opposed to those in the *com* domain.

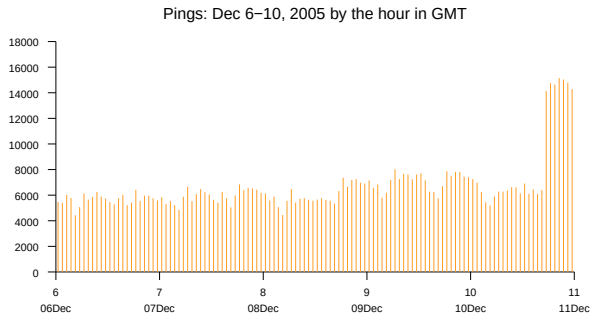

Figure 14: Ping Time Series of .info blogs over a five day period.

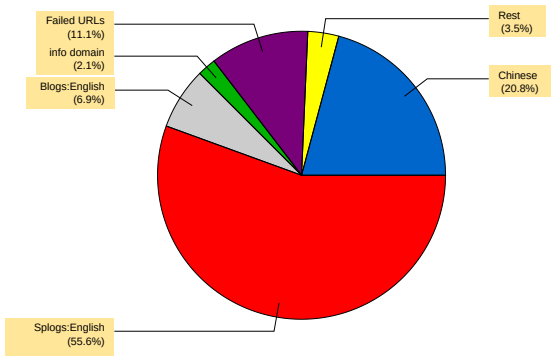

Figure 15: Distribution of URLs that ping the Update Ping Server

5. DISCUSSION

Based on our overall analysis we have the following useful observations that can be helpful for combating splogs. Splog detection can be employed at various points during its lifecycle:

1. **At Update Ping Servers.** Eliminating splogs at the blog ping servers is arguably the most effective approach. This will avoid computational overheads at the downstream services and systems. Splog detection here is challenging, since a judgment cannot be made with a high confidence until sufficient posts from the blog is observed. However, these ping servers should eliminate the overwhelming non-blog pings using blog identification techniques.

2. **Before Indexing Content.** Blog Search Engines can detect splogs fairly early during its life-cycle by analyzing content of blog homepages. Our own detection system provides an accuracy of 90% using features only local to the page. This is further evident in the keyword characterization of blogs and splogs.

3. **After Indexing Content.** Even if splogs escape filters in step one and step two, they can be detected later in the life-cycle if they indulge in the creation of link-farms. Such an approach can also make use of clues provided in step one and two to further enable splog judgment.

Our current system employs only local knowledge, and we have been quite successful in employing techniques at stages one and two above. We are currently working towards algorithms [13] that will be effective at all stages of a blog life-cycle. These algorithms draw from link-based [14] and adversarial classification [4] approaches.

6. REFERENCES

[1] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.

[2] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[3] M. Cuban. A splog here, a splog there, pretty soon it ads up and we all lose, 2005. [Online; accessed 22-December-2005; `http://www.blogmaverick.com/entry/1234000870054492/`].

[4] N. N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma. Adversarial classification. In *KDD*, pages 99–108, 2004.

[5] M. Darnashek. Gauging similarity with n-grams: language independent categorization of text. *Science*, 267:838–848, 1995.

[6] I. Drost and T. Scheffer. Thwarting the nigritude ultramarine: Learning to identify link spam. In *ECML*, pages 96–107, 2005.

[7] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In *NIPS*, pages 155–161, 1996.

[8] H. Drucker, D. Wu, and V. Vapnik. Support vector machines for Spam categorization. *IEEE-NN*, 10(5):1048–1054, 1999.

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

[10] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Databases*, pages 576–587. Morgan Kaufmann, 2004.

[11] P. Kolari. Welcome to the splogosphere: 75% of new pings are spings(splogs), 2005. [Online; accessed 22-December-2005; `http://ebiquity.umbc.edu/blogger/?p=429`].

[12] 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. to appear.

[13] 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).

[14] Q. Lu and L. Getoor. Link-based classification. In *ICML*, pages 496–503, 2003.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web.

Technical report, Stanford Digital Library Technologies Project, 1998.

[16] C. Pirillo. Google: Kill blogspot already!!!, 2005. [Online; http://chris.pirillo.com/blog/_archives/2005/10/16/1302867.html].

[17] S. Rubel. Blog content theft, 2005. [Online; http://www.micropersuasion.com/2005/12/blog_content_th.html].

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

[19] B. Wu and B. D. Davison. Identifying link farm spam pages. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 820–829, New York, 2005. ACM Press.

[20] L. Zhang, J. Zhu, and T. Yao. An evaluation of statistical spam filtering techniques. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(4):243–269, 2004.