Modeling the Spread of Influence on the Blogosphere

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

ABSTRACT

Blogs have become a means by which new ideas and information spreads rapidly on the web. They often discuss the latest trends and echo with reactions to different events in the world. The collective wisdom present on the blogosphere is invaluable for market researchers and companies launching new products. In this paper, we evaluate the effectiveness of some of the influence models on the blogosphere. We also examine the robustness of different heuristics in the presence of splogs or spam blogs. Experiments show PageRank based heuristics could be used to select an influential set of bloggers such that we could maximize the spread of information on the blogosphere.

1. INTRODUCTION

Forster Research [5] projects a sustained growth in advertising and marketing spending from around $14.7 billion in 2005, to about $26 billion by 2010 (in US alone). According to this survey, 64% of the advertisers are interested in advertising on blogs. Advertisers have realized the potential of blogs in influencing buying decisions of their target audience. Often when a buyer is interested in purchasing a product, blogs offer free and frank customer reviews. As Robert Scooble puts it "blogging is one huge word-of-mouth engine" [22]. Many startups have pitched pre-launch and beta version of their products to various bloggers with the hope that their reviews would create a buzz on the blogosphere and bring more attention to their company. Microsoft recently launched the Origami tablet PC amongst speculations that the hype was generated by a targeted campaign to a selected set of bloggers. Companies like Wal-mart are now trying to win back public opinion on their corporate policies by providing bloggers with exclusive news and inviting them for visits [2]. Even the US Department of Defense is now reaching out to hundreds of bloggers with more information about its counter-terrorism initiatives [9].

The blosphere provides a unique resource for studying information flow. Blogs are more structured than personal social networks, hence it is possible to mine them to identify influential nodes. There has been great interest in this topic and a number of models have be proposed that model information flow and influence in such social networks.

In this work we present the results of applying influence models proposed by Kempe et al. in the blogosphere and show how these techniques can automatically predict a set of influential blogs which are likely to be able to spread an idea most effectively. We also experiment with simpler heuristics such as PageRank which could be used to efficiently approximate the greedy algorithm for this problem. In this work we also highlight the importance of splog removal and discuss some of its implications on influence models.

The paper is organized as follows: Section 1 gives background of the related research in this area and section 2.1 discusses the basic diffusion model as adapted from the related work. Section 2.3 provide some results of different heuristics explored and section 3.4 talks about the broader implications on the blogosphere. Finally we conclude in section 4 and describe some of the future work in this area.

2. BACKGROUND AND RELATED WORK

Research in the area of information propagation was inspired by a large body of work in disease and epidemic propagation. As described in [8] this model applies well in the blogosphere where a blogger may have a certain level of interest in a topic and is thus *susceptible* to talking about it. By discussing the topic he/she may *infect* others and over time might *recover*. The authors use this approach in characterizing individuals into various phases of a topic in which they are more likely to become *infected*. They model individual propagation and use an expectation maximization algorithm to predict the likelihood of a blogger linking to another blogger. They also study the different types of topics present in the dataset and describe an approach to categorize topics into subtopics. Certain topics are more *infectious* than others and spread through the social network of bloggers. Automatically predicting such topics and developing models to accurately identify the propagation patterns on the blogosphere is the main focus of this work.

Adar et al. [1] have proposed the use of URL citations to infer the dynamics of *information epidemics* in the blogspace. They also show that the PageRank algorithm finds authoritative blogs. A variation, called iRank, is described to rank blogs based on their *informativeness*. In this scheme, each directed edge is assigned a weight $W_{ij} = w(\Delta d_{ij})$ where Δd refers to the time difference between the blogs citing a URL and $w(\Delta)$ is the weight function that gives importance to URL citations which are closer in time. The edge weights are then normalized and PageRank computation follows. This weighted graph is called the *implicit information flow graph*. iRank makes use of the temporal nature of blogs by differentially weighing each citation in the graph by the time difference between when the blog mentions a URL and

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

how soon it is referenced by other blogs. In our work we only use the PageRank computation and will investigate and compare iRank in future studies.

Since bloggers are constantly keeping abreast of the latest news and often talk about new trends before they peak, recent research has focussed on extracting opinions and identifying buzz from blogs [6]. Gruhl et al. [7] have found strong correlation between spikes in blog mentions to amazon sales ranks of certain books. More recently, Lloyd et al. [19] found similar trends for named entities in blog mentions and RSS news feeds.

Blogs are often topical in nature and their link structures constantly evolve as new topics emerge. Ravi et al. [16] study the word burst models [13] and community structure on the blogosphere [17]. They find a sustained and rapid increase in the size of the strongly connected component on the blogosphere and explain that the community structure is due to the tendency of the bloggers to topically interlink with posts on other blogs. Although, in our study we do not explicitly identify communities and subgraphs, models for information propagation might benefit from being able to find a set of individuals to target in a particular community. Conversely, if a highly influential node is selected within a community, the heuristics used should not select other nodes which would possibly be within the range of influence of the already selected node.

Our work is based on the problem posed by [21] and influence models proposed by [11, 10]. These models aim to mathematically simulate the spread of information in social networks. Kempe et al. proposed an approximation of the NP-Hard problem of identifying a set of influential nodes to target so that we can maximize the number of nodes that are activated or influenced. We use the basic *Linear Threshold Model* as proposed by Kempe et al. The influence model is described in Section 2.1. While the original models were validated on citation graphs, which are much smaller, we apply these algorithms on graphs derived from links between blogs. The citation graphs tend to be much cleaner and some of the techniques proposed do not apply well in the presence of splogs. We also discuss the applicability of simpler, PageRank-based heuristics for influence models on the blogosphere and the web in general.

2.1 Weblog Dataset

The dataset released by Intelliseek/Blogpulse[1] for the 2006 Weblogging Ecosystems Workshop consists of posts from about 1.3 million unique blogs. The data spans over 20 days during the time period in which there were terrorist attacks in London. This time frame witnessed significant activity in the blogosphere with a number of posts pertaining to this subject. The link graph that we extracted from this dataset consists of 1.2 million links among 300K blogs. However it was also observed that Livejournal [2] sites tend to be highly interlinked and hence for the purpose of the influence models presented in the following sections, we do not consider blogs from these sites for inclusion in the initial activation set. However, we do not discard the blogs from the link graph.

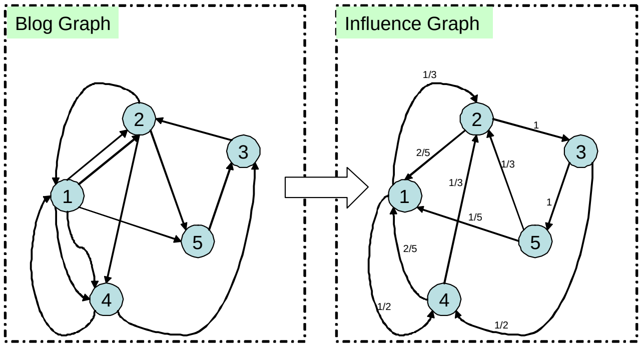

Figure 1: This diagram shows the conversion of a blog graph into an influence graph. A link from u to v indicates that u is influenced by v. The edges in the influence graph are reverse of the blog graph to indicate this influence. Multiple edges indicate stronger influence and are weighed higher

2.2 Influence Model

Bloggers generally tend to follow mainstream media, the Web, and also blog posts from people who may share similar interests. When an interesting *meme* emerges on some site, a blogger may choose to share it with his audience. Additionally, he may provide more insights and *trackback* to other sources of similar information. Other readers may comment on this post and thereby contribute to the conversation. Such an interactive process leads to the flow of information from one blogger to another. In approximating this interaction, we consider the presence of a link from site u to site v as evidence that the site u is *influenced by* site v. We consider only outlinks from posts and do not use comment links or trackback links for building the blog graphs. we take a rather simplistic view in the influence models and convert the *blog graph* to a directed *influence graph*. Figure 1 shows a hypothetical blog graph and its corresponding influence graph. An influence graph is a weighted, directed graph with edge weights indicating how much influence a particular source node has on its destination. Starting with the influence graph we aim to identify a set of nodes to target a piece of information such that it causes a large number of bloggers to be influenced by the idea.

Different influence models have been proposed [10, 11]. The two general categories are *Linear Threshold Model* and *Cascade Model*. We describe some of these below:
In the basic *Linear Threshold Model* each node has a certain threshold for adopting an idea or being influenced. The node becomes activated if the sum of the weights of the active neighbors exceeds this threshold. Thus if node v has threshold θ_v and edge weight b_{wv} such that neighbor w influenced v, then v becomes active only if

$$\sum_{w\ active\ neighbors\ of\ v} b_{wv} \geq \theta_v$$

and

$$\sum b_{wv} \leq 1$$

Another model is the *Independent Cascade Model* in which a node gets a single chance to activate each of its neighboring nodes and it succeeds with a probability P_{vw} which is independent of the history.

[1] http://www.blogpulse.com
[2] http://livejournal.com

As described in the above model, we rank each directed edge between u, v in the *Influence Graph* such that the presence of multiple directed edges provides additional evidence that node *node u influences node v.* If $C_{u,v}$ is the number of paralled directed edges from u to v the edge weight

$$W_{u,v} = \frac{C_{u,v}}{d_v}$$

where d_v is the indegree of node v in the influence graph.

Since computing the optimal value of expected size of the influenced set, $\sigma(A)$, remains an open question, the algorithm runs the influence propagation model for pseudo-random threshold values and computes the approximate size of $\sigma(A)$.

In selecting the order of activation of nodes, the simplest ranking scheme is one using the number of inlinks (which corresponds to the outlinks in the influence graph). This represents how many other nodes can be influenced by activating the selected node. We also explored PageRank [20] as a heuristic in selecting the target set. The PageRank algorithm works by simulating a random walk by a user who follows a link with probability q and jumps to a random section of the web graph with probability of $(1\text{-}q)$. If $C(a)$ is a set of outlinks for page a and $P_i..P_n$ are the pages pointing to a. Then

$$PR(a) = q + (1-q)\sum_{i=1}^{n} PR(p_i)/C(p_i)$$

One advantage of the pagerank algorithm is that it is relatively less expensive to compute and has fast convergence.

Another ranking scheme used is HITS [12] which assigns a hub and authority score. A good hub is one that points to a number of authoritative sources, while a good authority is one that is pointed to by many hubs. $H(p)$ = hub value of the page p and $A(p)$= authority value of a page p.

$$Authority(v) = \sum_{v \in S, v \rightarrow p} Hub(p)$$

And

$$Hub(p) = \sum_{v \in S, p \rightarrow v} Authority(v)$$

Finally we compare these heuristics with the greedy hill climbing algorithm. In the greedy approach nodes are incrementally added to the initial activation set without backtracking. At each time step, the influence model is run and a node is selected to be added to the initial target set. The node is selected such that adding it to the target set would provide the largest locally optimal increase in the size of the influenced node set.

Other methods such as "distance centrality" based heuristic are also widely used in many studies. This however could not be applied to the blog dataset since computing the centrality scores over large graphs is expensive without partitioning or identifying subgraphs.

2.3 Splog Detection

Splogs are spam blogs that consist mostly of automatically generated content with the malicious intent of hosting advertisements of redirecting traffic to a particular website. Often these unscrupulous techniques are used as means of search engine optimizations and aim to boost the search engine rank of some of the sites. Splogs, if not eliminated, can cause the influence models to be less effective at predicting good target sets.

We have developed splog detection models [14][15] trained using Support Vector Machines [4]. These models are based on logistic regression and can hence predict probabilities of class membership. 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 word n-grams to detect splogs, with an accuracy of around 90%. Our working splog detection system employs a multi-step approach to detect splogs, filtering out and eliminating URLs as it goes through these steps: (i) from known blacklists, (ii) language detection[3], (iii) blog identification, and (iv) splog detection. Steps (iii) and (iv) can currently be employed only on blogs in the English language. Blogs that pass through all four filters are tagged as authentic English blogs with a high confidence.

Using our splog model on the weblogs in the dataset, we identified 32,514 (2%) as splogs and 71,173 (5%) no longer accessible. The splog detection is based on analyzing the entire home-page of a blog. Since July 2005, when this dataset was generated, a number of blogs have become non-existent. A blog can become non-existent if it was explicitly deleted by its owner, or it it was pulled down by the blog host if it was identified as a splog. Since in both these cases, the blog does not contribute to the influence propagation, we ignore these and for simplicity we assume that failed URLs are also splogs.

3. EXPERIMENTS

The following section describes some of the experiments and results. The experiments were performed using the influence models described in the previous sections. Here we investigate the effect splogs have on various ranking schemes and evaluate the performance of various node selection heuristics in the presence of splogs and after splog elimination. Finally, we compare the Technorati ranks of the different target sets generated to estimate their quality.

3.1 Influence of Splogs

Firstly we investigate the influence of splogs on the different ranking schemes. Splogs are identified using supervised learning with Support Vector Machines as described in [14]. Figure 2 shows the distribution of the splogs in the top results as ranked by PageRank, HITS and the Indegree heuristic. From this graph we can observe that the indegree and HITS based heuristics were subject to being easily spammed. As described in [23] the HITS algorithm is susceptible to spamming and this is particularly true in the presence of Tightly Knit Communities (TKC). This phenomenon was originally proposed by [3] and also studied by [18]. Other variations of the HITS algorithm attempt to solve this problem, however it was interesting to note that this issue, though it may have already been seen in the case of web spam, is also true for the blog domain. However, PageRank was found to be relatively resilient to splogs.

To further investigate this, we analyze the link types. Table 1 shows the frequency of different types of links in the graph. From this, we can observe that splogs generally tend to link to other splogs. Thus indicating that there is a possibility of a community structure in splogs. However

[3] Provided by James Mayfield

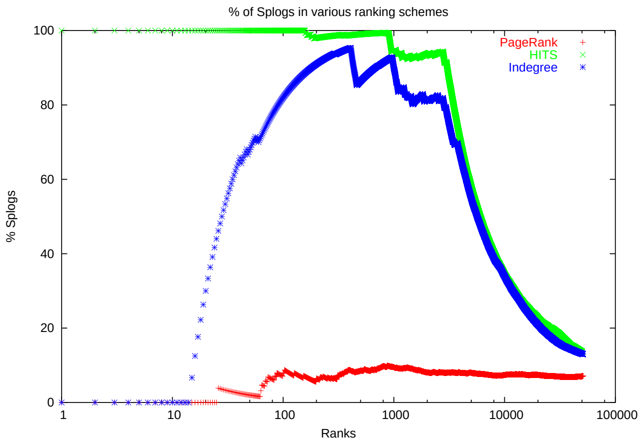

Figure 2: ***The graph shows the percentage Splogs in PageRank, HITS, Indegree. The above graph shows what percentage of the top N results in each of the ranking schemes are splogs. The topmost curve corresponds to HITS, the next curve shows indegree and the lowest curve shows PageRank.***

	splogs+failed	**blogs**
splogs+failed	704,451	81,846
blogs	33,108	452,305

Table 1: ***Table shows the link types from the link graph of 301700 nodes. 35.5% were from a blog to a blog while 55.4% were from splog+failed to splog+failed.***

splogs also link to legitimate blogs, which most often are high ranked or popular pages. Although less often, there are links from blogs to splog pages. One explanation for these is that it could be due to comment spam, however this would require further investigation. As seen in table 1, a large fraction of the link graph consists of links from splogs+failed to splogs+failed URLs.

3.2 Node selection heuristics

We run the influence models using different heuristics such as PageRank, indegre and greedy algorithm. In PageRank and indegree the nodes are added to the initial target set in the order of their rank. Once a node is selected, the influence model is run and nodes are activated depending on their threshold. The number of nodes influenced is estimated over multiple iterations. In greedy algorithm nodes are incrementally added to the target set if they locally maximize the size fo the influenced node set. Figure 3 shows that the indegree heuristic (which corresponds to the outdegree in the influence graph) at first seems to perform well, but fails to influence more nodes as the initial set grows. Looking further at the results of the top ranks, as shown in table 2, suggests that the spread of influence plataues due to the presence of a community of splogs, quite likely from a tightly knit community, around the top ranks. Eliminating such spam using the algorithm previously described results in the 4. Thus, by eliminating splogs, the top results obtained from the indegree heuristics almost approximated PageRank. This was also due to the fact that about 70% of the blogs as ranked by PageRank and indegree match after splog elimination.

However, it was found that the pagerank and greedy heuristics seem to perform almost the same even after the elimination of roughly 103687 nodes which correspond to splogs (including failed URLs).

3.3 PageRank vs Greedy Heuristic

The Greedy heuristic of node selection performs better than both PageRank or indegree. However one of the disadvantages of the greedy approach is that it is computationally quite expensive. PageRank on the other hand is an iterative algorithm that converges to the principal eigenvector of the adjacency matrix. While it is faster to compute, it requires knowledge of the structure of links which might emerge only after the blogpost has been read and linked to by other blogs over a period of time.

As observed in table 1, splogs often link to blogs and hence when a blog is activated it may happen that the influence propagation could lead to the activation of some splogs (which may be a part of a community). These splogs may in turn activate other splogs. Due to this reason it can be observed that after the elimination of splogs from the link graph, the number of activated nodes at target set size of 100 is actually slightly lower than the number of activated nodes when we considered the link graph with splogs in it.

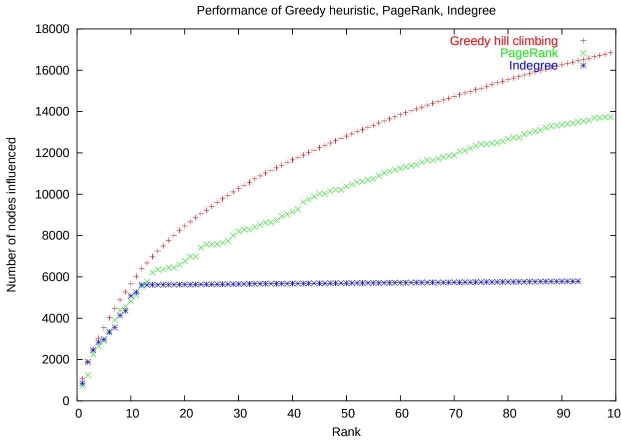

Figure 3: ***The graph shows the performance of greedy heuristic vs PageRank heuristic vs indegree. The above results show the average influence after 10 iterations of each heuristic.***

3.4 Comparison with Technorati Ranks

For a sample run of each of the heuristics, we use the Technorati[4] API to find the ranks of the blogs that were selected in the initial activation set. Figure 5 shows the top 100 blogs selected by each of the heuristics sorted by their Technorati ranks. The lower Technorati ranks indicate that the blog is more popular (as per Technorati's index). The results suggest that PageRank is a good approximation to discover the top ranked blogs. The greedy algorithm can select blogs

[4] http://www.technorati.com/

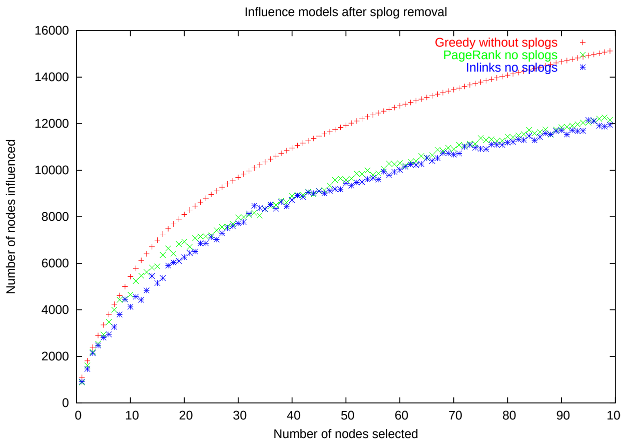

Figure 4: ***The graph shows the performance of the indegree and pagerank heuristic after splog elimination. The above results show the average influence after 10 iterations of each heuristic.***

that are somewhat less popular but have recently influenced highly popular or well connected blogs. Even though the dataset is a relatively small sample of the blogosphere's link structure, it found that from the top 100 popular blogs, as listed on Technorati, 65 were present. Of these, 57% were found to be in the top 100 by the PageRank heuristic, while 70% of them were found in the top 200 ranks. Thus, PageRank based heuristics would also perform well in identifying the A list bloggers.

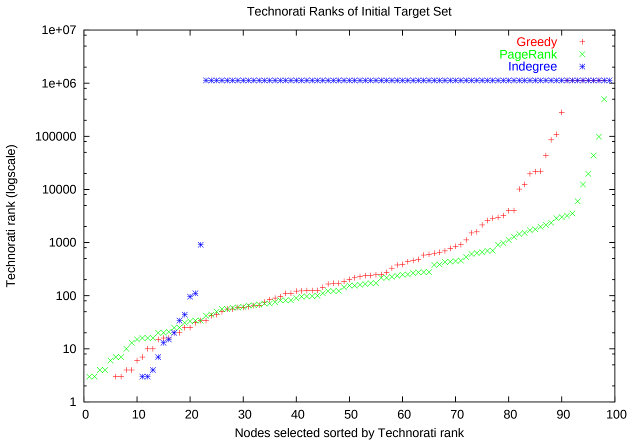

Figure 5: ***The graph shows the nodes selected by various heuristics sorted in the order of their Technorati ranks. Lower technorati ranks indicate that the blog is more popular.***

4. DISCUSSION

There are several issues that we will investigate in future work. These include refining the influence model with respect to topics, enhancing it by tracking a blog's existing sentiment toward a topic, and taking into account the dynamics of the Blogosphere. We briefly discuss each issue in turn.

Rank	url	inlinks
1	http://www.livejournal.com/users/pics	3072
2	http://www.boingboing.net	2191
3	http://www.dailykos.com	2017
4	http://www.engadget.com	1942
5	http://profiles.blogdrive.com	1526
6	http://michellemalkin.com	1242
7	http://www.opinionjournal.com	1232
8	http://instapundit.com	1187
9	http://slashdot.org	1124
10	http://www.powerlineblog.com	909
11	http://www.huffingtonpost.com/theblog	905
12	http://corner.nationalreview.com	853
13	http://www.talkingpointsmemo.com	733
14	http://www.captainsquartersblog.com/mt	728
15	**http://espn-presents2003-world-seriesofpoker.blogspot.com**	711
16	**http://3-world-series-of-poker-online-3.blogspot.com**	711
17	**http://worldseries-of-poker-network-tv-show.blogspot.com**	711
18	**http://wsop2003.blogspot.com**	711
19	**http://wsop-bracelet1.blogspot.com**	711
20	**http://worldseries-poker.blogspot.com**	711
21	**http://worldseries-of-poker-official.blogspot.com**	711
22	**http://worldseries-of-poker-wsop.blogspot.com**	711
23	**http://world-series-of-poker-nocd-patch66.blogspot.com**	711
24	**http://4-world-series-of-poker-past-winners.blogspot.com**	711
25	**http://7-wsop-games-7.blogspot.com**	711

Table 2: ***Table shows the list of blog homepages ranked by number of inlinks. Highlighted links show the splogs in the top ranked list and indicate a tightly knit community of splogs.***

It is clear that a blog's influence is often a function of its topic. *Engadget*'s influence is centered in the domain of consumer electronics and *Daily Kos* in politics. A post in the former is not likely to be very effective at influencing opinions on political issues even though *Engadget* is one of the most popular blogs on the Web. For topics of interest to a narrow community of people, a relatively low ranking blog might be very influential. The blog *Lambda the Ultimate*, for example, is thought to be very influential with respect to programming languages. Categorizing blogs by topics can be done in a variety of ways including using a reference topic hierarchy (e.g., Yahoo's), an ontology (e.g., Wordnet) or a combination of user defined and assigned tags. It is also possible to develop a model that predicts a persons interest, e.g., expressing the probability that a person will be interested in topic X if they are interested in topic Y).

A second issue involves introducing notions of the current

Rank	PageRank	Greedy
1	*http://www.boingboing.net*	*http://www.engadget.com*
2	*http://www.dailykos.com*	*http://www.boingboing.net*
3	*http://www.engadget.com*	*http://www.dailykos.com*
4	*http://slashdot.org*	*http://postsecret.blogspot.com*
5	*http://www.opinionjournal.com*	*http://slashdot.org*
6	*http://postsecret.blogspot.com*	*http://www.albinoblacksheep.com*
7	http://www.huffingtonpost.com/theblog	*http://www.opinionjournal.com*
8	*http://www.albinoblacksheep.com*	http://profiles.blogdrive.com
9	*http://blogs.guardian.co.uk/news*	http://godlessmom.blogspot.com
10	http://www.the-leaky-cauldron.org	*http://thinkprogress.org*
11	*http://radio.weblogs.com/0001011*	*http://www.huffingtonpost.com/theblog*
12	*http://instapundit.com*	*http://corner.nationalreview.com*
13	http://michellemalkin.com	*http://radio.weblogs.com/0001011*
14	*http://www.gizmodo.com*	*http://www.gizmodo.com*
15	*http://fifthnail.blogspot.com*	http://www.brain-stream.com
16	http://www.powerlineblog.com	http://www.appleinsider.com
17	http://www.washingtonmonthly.com	*http://www.metafilter.com*
18	http://www.crooksandliars.com	*http://www.lifehacker.com*
19	*http://www.gawker.com*	http://www.threadwatch.org
20	*http://corner.nationalreview.com*	http://bagofchips.net
21	*http://www.metafilter.com*	*http://blogs.guardian.co.uk/news*
22	*http://www.andrewsullivan.com*	*http://instapundit.com*
23	*http://thinkprogress.org*	http://googleblog.blogspot.com
24	http://www.talkingpointsmemo.com	*http://fifthnail.blogspot.com*
25	http://americablog.blogspot.com	*http://www.andrewsullivan.com*
26	http://billmon.org	http://www.netzeitung.de/deutschland
27	http://www.latimes.com/news/opinion/commentary	*http://www.gawker.com*
28	*http://www.warrenellis.com*	*http://www.defamer.com*
29	*http://www.defamer.com*	http://spaces.msn.com/members/l1n
30	http://www.penny-arcade.com	http://taopoker.blogspot.com
31	*http://www.journalfen.net/community/fandom_wank*	*http://www.dpreview.com*
32	http://atrios.blogspot.com	http://yglesias.tpmcafe.com
33	*http://www.overheardinnewyork.com*	http://prawfsblawg.blogs.com/prawfsblawg
34	http://www.dooce.com	http://www.lewrockwell.com
35	*http://www.lifehacker.com*	http://spaces.msn.com/members/5-06-20_17.53/
36	http://www.makezine.com/blog	*http://www.overheardinnewyork.com*
37	http://www.littlegreenfootballs.com/weblog	http://www.djournal.com/pages
38	http://www.micropersuasion.com	http://bronski.net
39	http://www.blogathon.org	*http://www.warrenellis.com*
40	http://www.tpmcafe.com	*http://www.londonist.com*
41	http://volokh.com	http://www.respectcoalition.org
42	*http://www.londonist.com*	http://notes.torrez.org
43	http://www.captainsquartersblog.com/mt	http://www.codeproject.com
44	http://subvic.blogspot.com	*http://www.sctnomination.com/blog*
45	http://www.juancole.com	http://joi.ito.com
46	http://www.thejackol.com	*http://www.journalfen.net/community/fandom_wank*
47	*http://www.dpreview.com*	http://oghc.blogspot.com
48	http://bitchphd.blogspot.com	http://www.kotaku.com
49	http://xiaxue.blogspot.com	http://www.koreus.com
50	http://www.betanews.com	http://www.sixapart.com/movabletype/beta

Table 3: *Table shows an example of top 50 target set blogs selected using pagerank and greedy heuristics. Italics indicate that the url is common in both the heuristics.*

beliefs, opinions and sentiments of blogs and other Web sites with respect to a given topic, concept or object. This can be important in two ways. The first is to modify the influence graph to associate a sentiment to each link. Many blogs might link to blog B expressing strong negative opinions about topic T and few blogs might link to B with positive opinions on T. In such cases, it's unlikely that information on topic T posted on B will have a positive impact. The second way that this can be important is to avoid "preaching to the choir". Getting *iloveipods.com* to post positive information about ipods may not have a big effect. Although the site influences a large community, most of its members are already ipod fanatics.

Finally, the Blogosphere is quite dynamic, with blogs coming into and going out of existence regularly and their popularity rising and falling, sometimes rapidly. Current heuristics generally tend to be biased towards the already popular blogs. It is quite possible that if an obscure site with a low PageRank was to put an interesting meme it would suddenly be popular amongst most bloggers. Developing influence models that could take this temporal nature into consideration could identify blogs that are upcoming and rising in popularity.

5. CONCLUSIONS

In this paper we have presented an analysis of influence models on a large scale and a real world blog graph. We have also shown how splogs effect some of the heuristics such as indegree, while others such as greedy and pagerank perform well even in presence of splogs. We suggest pagerank as an inexpensive approximation to the greedy heuristic in selecting the initial target set for activation.

6. ACKNOWLEDGEMENTS

We would like to thank Intelliseek for providing the weblog dataset and Technorati for the blog information API.

7. REFERENCES

[1] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, New York, NY, USA, May 2004.

[2] M. Barbaro. Wal-mart enlists bloggers in p.r. campaign. `http://www.nytimes.com/2006/03/07/technology/07blog.html`.

[3] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *SIGIR*, pages 104–111, 1998.

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

[5] T. Elkin. Just an online minute... online forecast. `http://publications.mediapost.com/index.cfm?fuseaction=Articles.showArticle&art_aid=29803`.

[6] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *KDD*, pages 419–428, 2005.

[7] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *KDD*, pages 78–87, 2005.

[8] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW*, pages 491–501, 2004.

[9] U. P. International. Pentagon reaches out to bloggers. `http://www.upi.com/SecurityTerrorism/view.php?StoryID=20060306-020621-7264r`.

[10] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[11] D. Kempe, J. M. Kleinberg, and É. Tardos. Influential nodes in a diffusion model for social networks. In *ICALP*, pages 1127–1138, 2005.

[12] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[13] J. M. Kleinberg. Bursty and hierarchical structure in streams. *Data Min. Knowl. Discov.*, 7(4):373–397, 2003.

[14] 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.

[15] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi. Detecting spam blogs: A machine learning approach. 2006. submitted to AAAI 2006 - AI on the Web.

[16] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW*, pages 568–576, 2003.

[17] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *Computer Networks*, 31(11-16):1481–1493, 1999.

[18] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. *Computer Networks*, 33(1-6):387–401, 2000.

[19] L. Lloyd, P. Kaulgud, and S. Skiena. Newspapers vs blogs: Who gets the scoop? In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006. to appear.

[20] 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.

[21] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, pages 61–70, 2002.

[22] R. Scooble and S. Israel. *Naked Conversations: How Blogs are changing the way businesses talk to their customers.* 2006.

[23] B. Wu and B. D. Davison. Identifying link farm spam pages. In *WWW (Special interest tracks and posters)*, pages 820–829, 2005.