Implicit Structure and the Dynamics of Blogspace

Eytan Adar Li Zhang
Lada A. Adamic Rajan M. Lukose

HP Information Dynamics Lab

Abstract

Weblogs link together in a complex structure through which new ideas and discourse can flow. Such a structure is ideal for the study of the propagation of information. In this paper we describe general categories of *information epidemics* and create a tool to infer and visualize the paths specific infections take through the network. This inference is based in part on a novel utilization of data describing historical, repeating patterns of infection. We conclude with a description of a new ranking algorithm, iRank, for blogs. In contrast to traditional ranking strategies, iRank acts on the implicit link structure to find those blogs that initiate these epidemics.

General Terms

Measurement, Experimentation, Algorithm

Keywords

Weblog, information flow, implicit link, ranking

1. Introduction

As new tools such as weblogs (or blogs for short) and wikis have taken hold, the WWW is being transformed. While website creation has been traditionally the domain of the technically oriented, blog software is allowing many more users to generate web content to be consumed by groups, both large and small. The ease of generating such content has caused certain functions traditionally reserved for e-mail and instant messaging to be shifted to the WWW. A significant use of blogs is as a publicly exposed, online diary describing both real-world and web based experiences. Bloggers frequently read each other's postings, and the phenomenon of listing and commenting on information found through a user's online exploration is common. These posts and comments are intended to relay the latest interesting, humorous, or thought-provoking information the user has run across. This information is added to the blog with the full realization by, or hope of, the author that it will be read by others. In this way the web, through blogs, becomes a medium for rapid transfer of *memes* [6]. A humorous example is the Giant Microbes website which sells stuffed toys in the form of common diseases.

Being the first to find or comment on the latest and newest is of clear importance to blog authors as it leads to increased readership and linking, and improving the author's status in the so-called *blogspace*. Many new services have begun to take advantage of this competition to detect network buzz, e.g. www.blogdex.net, www.technorati.com,www.blogpulse.com, as well as assign value to individual blog, e.g. www.blogshares.com. However, despite the utility of such applications for finding the latest information very little attention has been paid to its spread.

In this paper, we study the pattern and dynamics of information spreading in blogspace. We consider both the large scale aspects of spreading patterns as well as how a specific, individual link may be tracked in blog networks. Our study is enabled by the blog data that is crawled on a daily basis. In our study, we will only track link information. While memes can take many forms, those addressed by URLs are by far the simplest to track and disambiguate. For example, our system will track http://www.giantmicrobes.com instead of discussions about the Giant Microbe toys, or images copied from the source site.

With the triplets of (URL, blog, day of URL citation), we first characterize the spreading patterns of information. Our method constructs citation vectors for URLs that reflect the number of citations per day. These vectors cluster into 4 distinct types of time profiles. These profiles can be given intuitive characterizations based on the type of URLs contained in the cluster. For instance, the "Slashdot effect" [4] is clearly observed.

We then set to determine how specific epidemics travel through blog networks by inferring links between blogs. While an explicit link between blogs is a good indication of where the blog obtains information, such links are often absent on the blogs because the blogs tend to link to the source of information directly. We present a method by using various classifiers that combine features such as the similarity between blog contents, in addition to the access pattern. Our experiments show very high accuracy when tested on the blogs between which the links are known to exist. Based on the inferred links, we have created a tool to automatically generate a visual depiction of these epidemics (available at http://www.hpl.hp.com/research/idl/projects/blogs/).

Finally, we present iRank, an algorithm that ranks the blogs according to how important they are for propagating information. The ranking algorithm is based on our ideas of inferred implicit structure of blogs. We show that the ranking produced by iRank is significantly different from the ranking produced by the PageRank algorithm on the explicit link structure of blogs. While PageRank accurately identifies the authoritative blogs, iRank is more accurate in identifying those weblogs that serve as sources for information that later becomes widely linked-to. Additionally, since iRank relies on data that can change far more frequently than the static, explicit link structure that PageRank relies on, it can more accurately reflect the dynamic nature of blogspace.

The paper is organized as follows. In Section 1.2, we describe the method of classifying epidemic patterns. The method for

inferring implicit links is described in Section 3. Section 4 presents the ranking algorithm based on the implicit link structures of blogs. For this study we utilize the daily Blogpulse differential crawls for May 2003 as well as a full-text crawl from May 18th, 2003. This data set contained 37,153 blogs with 175,712 links appearing more than once.

1.1 Related work

The limited quantitative research on blogs has primarily focused on determining the size and usage of blogspace[14] as well as some explorations on dynamics [13]. Although some work on visualization of information spread [20] has been pursued no work to date has specifically addressed the various facets of blog epidemics and their applicability to ranking.

We borrow some terms from epidemiology, and our approach has been inspired by methods from that field. Notably, recent theoretical work [17] and applied research [9] have pointed at possible ways to build *epidemic trees*. While relevant these methods do not take advantage of the repeated information available to us (i.e. they will only track one strain or one epidemic).

The social networks community has also explored issues of inference and diffusion [10], most frequently using structural [1][5][7][11][12][18] and node properties [19]. As this data is mostly static, we are unaware of research using timing information for prediction.

There has been extensive study on the link structure of the web pages. Typically, the web is modeled as a graph in which nodes correspond to web pages, and edges correspond to hyper links on web pages. An approach using hyperlinks as proxies for popularity has been very successful in measuring the importance of web sites [16] and understanding the community formation of web pages [11]. While this approach is good at evaluating the sum popularity of web pages, it cannot discern how news of the webpage propagated, i.e. how other webpages that link to it found out about it. Blogs far more frequently point at the information source directly rather than to another site where they first learned about the information. Using the link structure based methods would discover popular materials discussed by bloggers but is less likely to identify the blogs that are important in information propagation.

1.2 Link structure in Blog networks

Spreading patterns can greatly inform us about general information flow in networks. However, to understand how a specific URL has spread requires a closer look at the network structure. Given any URL, we do not expect that all blogs who have mentioned it first saw it at the source. More likely, a small number of sites initially replicated the source, and their readers replicate it further. Is it possible then to determine the source of infection for any given blog, essentially performing informational contract tracing?

Of some help is the great deal of explicit structural information that can provide a starting point. Blog creators frequently provide *blogrolls* (a list of other blogs frequently read by the author) or automated *trackbacks* [21].

Furthermore, some blogs will explicitly indicate where the author first saw the URL. For example, the May 16th entry at *www.livejournal.com/users/bentleyw* reads:

"8:48a - GIANTmicrobes http://www.giantmicrobes.com/

'We make stuffed animals that look like tiny microbes—only a million times actual size! Now available: The Common Cold, The Flu, Sore Throat, and Stomach Ache.' (via Boing Boing)"

These *via* links are highly informative for the purposes of inferring how information has traveled. Interestingly, we can also use this information to address issues of missed data. In the Giant Microbes example, we see that the popular boingboing.net blog mentioned the Giant Microbes site on May 14th, but was missed by the crawler. Adding boingboing.net

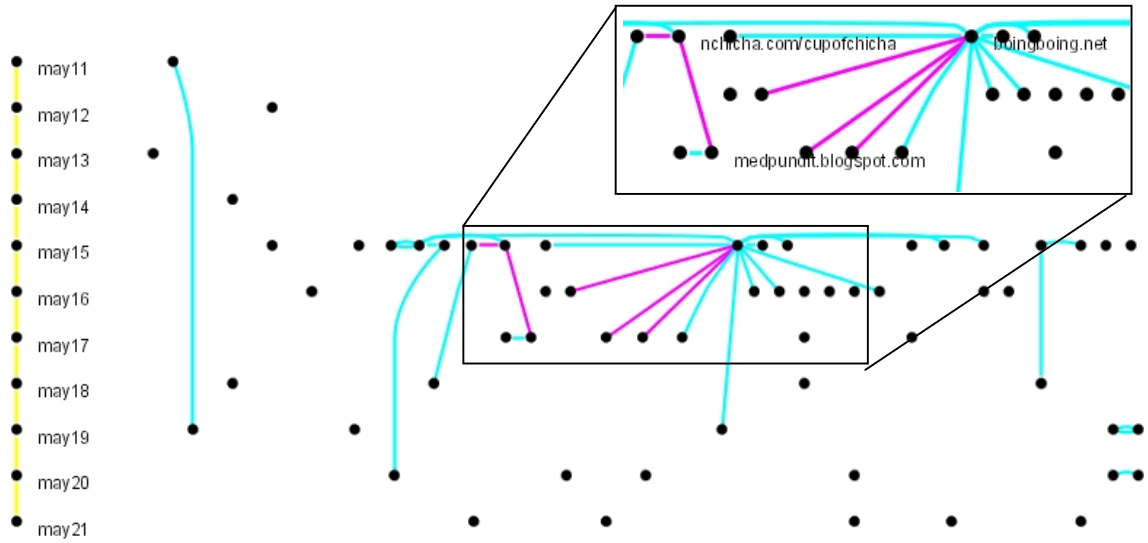

Figure 1: A visualization of all blogs linking to the Giant Microbes site. Only explicit links are shown.

into the network "explains" three additional sites that attribute the URL to it. Further, considering via links reinforces two existing links (medpundit.blogspot.com → nchicha.com/cupofchina and scrubbles.net → nchicha.com/cupofchina) making us more confident of the URL source.

Unfortunately, via links are quite rare. Among hundreds of thousands of URLs found in our data set, only 2,899 via links were found between two known blogs with an additional 2,306 links between a known blog and an uncrawled source (sometimes a blog, sometimes not). In exhausting explicit attribution (via) links as explanations we are forced to rely on inference techniques to fully describe infections.

Given the huge number of blog-blog links, one may expect that all propagation be explained by this explicit network structure. However, this is largely not the case. For URLs appearing on at least 2 blogs, 77% of blogs do not have a direct link to another blog mentioning the URL earlier. For those URLs present on at least 10 blogs, 70% are not attributable to direct links. Figure 1, is a simple example automatically generated by our system, illustrating all blogs that mention the Giant Microbes website. The lines between blogs indicate a directed, explicit, connection between the two blogs. Interpreting this figure it may be safe to assume that *medpundit.blogspot.com* saw a pointer to Giant Microbes at *nchicha.com/cupofchicha*. The bulk of blogs are free-floating and are considered "unexplained." Possible reasons for this include missing data, which may be completely un-crawled blogs or data missing from crawled blogs (e.g. blogs changing too rapidly for the crawler). Further, blogs may refer to a meme by a variant name or URL and thus will be missed. Blogs may also obtain information from non-blog sources such as mass media, email, or instant messaging. Finally, certain social dynamics in blog networks may lead to competition where blog authors may conceal where they got the information from in order to drive traffic to their own website.

2. Epidemic Profiles

URLs that are repeatedly cited within the community of bloggers are tracked by several websites (e.g. www.blogpulse.com, www.blogdex.com, www.daypop.com). The popularity measurements of these URLs act as a simple filtering and ranking mechanism, allowing users to quickly find potentially interesting URLs that many bloggers are talking about in near real-time. The propagation of such URLs also allows us to gain insight into how information may flow in general through the blog community. Using our dataset of blogs, URLs cited within them, and estimates of the time at which the citations occurred, we studied the time courses of URL citations. We clustered these citation traces in order to see if there was a small set of basic citation patterns into which most URLs fell. The procedure was able to identify four intuitively meaningful patterns, which we term "epidemic profiles".

For the analysis we selected 259 URLs from all the URLs mentioned by blogs between the period of May 2, 2003 and May 21, 2003 according to the following criteria: (i) the URL was cited at least 40 times, (ii) the URL was not on a stoplist of 198 URLs (e.g. the homepages of common news sources such as nytimes.com or blogging resources such as moveabletype.org), and (iii) the URL was not mentioned by any blogs on May 2, 2003. This tended to eliminate profiles that might have had their peak before May 2, but do not sustain interest past that point.

For each of the 259 well-cited URLs we created a vector, ordered by day, which contained the number of citations on each day. After normalizing the vectors, a standard k-means clustering algorithm was applied. The distance between vectors was measured by the Euclidean metric, and cluster centroids were defined as the arithmetic mean of the cluster member vectors.

Figure 2 shows the resulting clusters when the k-means algorithm was constrained to four clusters (k = 4). Increasing k tended to introduce redundant or lower quality clusters. These clusters were found consistently when the expectation minimization part of the clustering algorithm was repeatedly re-initialized and re-run, indicating the quality of the clusters. Figure 3 shows the centroids of the four clusters.

The cluster of Figure 2(a) contains URLs that refer to topics of sustained interest throughout the period, such as the Apple iTunes and iPod web pages and the Friendster homepage. The second cluster in Figure 2(b) contains URLs that tend to represent serious news editorial content such as opinion pieces from the New York Times. Another type of URL fitting this pattern is the Sun Microsystems Java portal page at

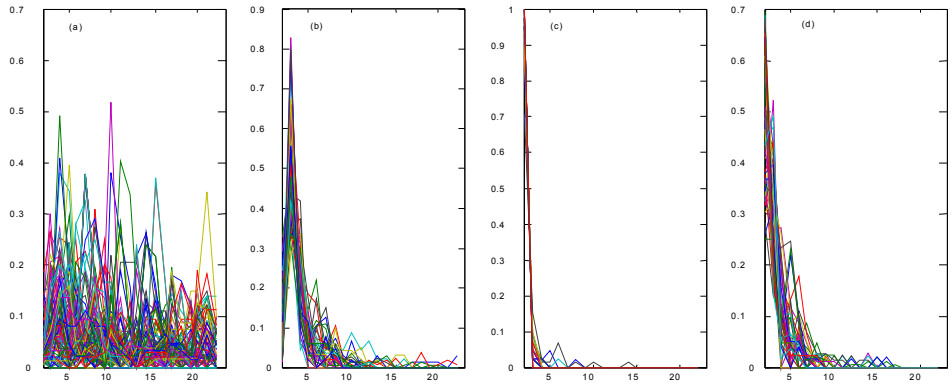

Figure 2. The four epidemic expression profiles resulting from the k-means clustering procedure: (a) 124 of 259 URL time course profiles are of the sustained interest type, (b) 51 URLs peak on day two followed by a slow decay, (c) 38 URLs peak on day one with a very fast decay, and (c) 46 URLs have a peak on day one with a slower decay.

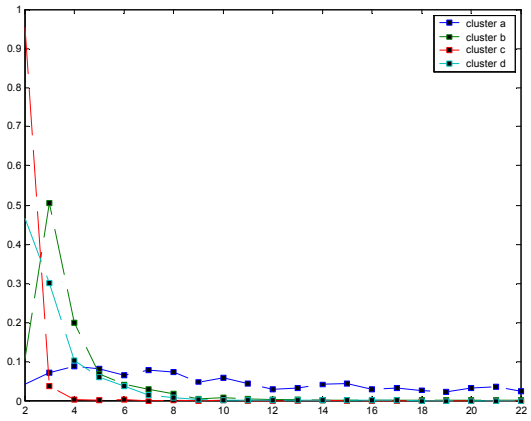

Figure 3. Centroids of the four meme expression profiles

java.sun.com/getjava. The pattern in this cluster is a rise from day one to a peak at day two followed by a slow decay. Only five of the URLs in the third cluster shown in Figure 2(c) are not Slashdot articles. These fast decaying profiles, peaking on day one, clearly represent the "Slashdot effect". Finally, the slower decaying profiles that also peak on day one, shown in Figure 2(d), appear to mostly represent less serious news content. For example, content from the culture portion of Wired News' site, is disproportionately represented. These clusters are available at the project website http://www.hpl.hp.com/research/idl/projects/blogs.

3. Inferring Infection Routes

In order study the micro-behavior of specific information it is necessary to infer infection routes. This requires the ability to predict the likelihood of two blogs linking to each other. This prediction is achieved by the use of a classifier that relies on various measures of blog similarity. These include:

- *blog_sim*: The number of other blogs two sites will have in common (i.e. a community estimate)
- *link_sim*: The number of non-blog links (i.e. URLs) shared by the two,
- *text_sim*: Textual similarity, and
- Timing and repeated history information reflecting the frequency of one blog becoming infected before another.

3.1.1 Blog and Link Similarity

Both blog_sim and link_sim were calculated as a vector-space cosine similarity measure that ranges between 0 (no overlap in URLs) to 1 (all URLs are shared). For example, the *link_sim* similarity for blogs A and B is computed as:

$$link\_sim(A,B) = \frac{\| n_A \cup n_B \|}{\sqrt{\| n_A \|} * \sqrt{\| n_B \|}}.$$

Where n_A and n_A are the sets of URLs found on blog A and B respectively. Blog similarity, *blog_sim*, is computed similarly but using just blog URLs.

Figure 4 illustrates how these scores vary for blog pairs that have bidirectional links (point at each other), unidirectional (one blog points at the other) and unlinked. From a sample set of 1000 blogs we find 1841 bidirectional pairs, 2216 unidirectional, and 1000 (randomly chosen) unlinked pairs. Clearly, the bulk of unlinked received very low similarity scores whereas the linked categories are much more distributed. Similarity distributions for the three categories of pairs were found to be distinct by the Kruskal-Wallis test with $p < 10^5$

3.1.2 Text Similarity

Textual similarity, *text_sim*, was calculated as the cosine similarity of term vectors. These term vectors were weighted using the standard TFIDF scheme [14]. This similarity measure displays similar properties as *link_sim* and *blog_sim* where linked blogs received higher scores. However, given the abundance and ambiguity of text relative to URLs, which are unique and sparser, this feature was not as discriminating.

3.1.3 Infection Timing

A unique feature available to us through differential crawling is the timing information representing when (to the day) a blogger adds a URL. The gathered timing information can give us the critical "when" of infection so that blogs can be ordered but also indicates which infection paths appear to be repeating. That is, of the URLs mentioned on Blog A how many were mentioned by Blog B a day earlier?

Table 1 summarizes our findings from this analysis. Again, we find a clear difference exists between linked and unlinked pairs. We also find that blogs pointing at each other have a 45% chance of mentioning at least one common URL, and blogs with an unreciprocated link have a 36% chance. A positive feature of this technique is that it allows us to begin to more accurately predict link directionality. For example, we note that for the bidirectional links, the timing of infection is approximately equal (the number of times A posts before B, after B, or at the same time). On the other hand, when blog A

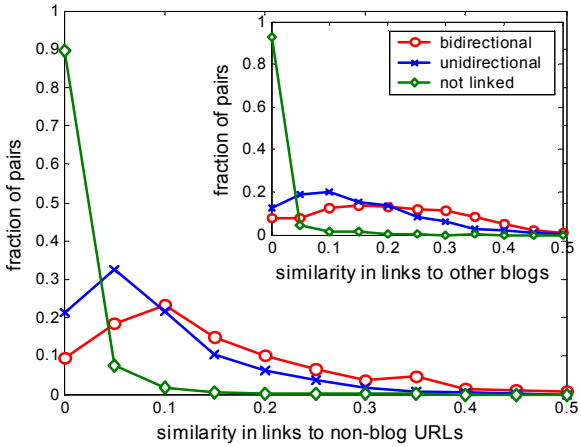

Figure 4 Similarity in links between reciprocated, unreciprocated, and non-linked blog pairs, for both non-blog URLs (main figure) and blog URLs (inset).

link type	same day	A after B	A before B
A ←→ B	17.4%%	24.5%	24.5%
A → B	10.9%	22.9%	17.0%
A,B (unlinked)	0.6%	1.5%	1.3%

Table 1 Percentage of pairs AB sharing at least one link, broken down by time ordering.

has an unreciprocated link to blog B, the distribution is biased so that A is more frequently infected after B.

Six independent features were extracted from the time ordered data: $A_{before}B/l_A$, $A_{after}B/l_A$, $A_{same\text{-}day}B/l_A$: $A_{before}B/l_B$, $A_{after}B/l_B$, $A_{same\text{-}day}B/l_B$, where: $A_{before}B$, $A_{after}B$, and $A_{same\text{-}day}B$ represent the number of links mentioned by A before, after, and on the same day as B respectively and n_A and n_B the number of links referenced by A and B over the crawl period.

3.2 Classification

We constructed two SVM models using the free LIBSVM toolkit, with a standard radial basis function. The first classifier predicted three different classes (reciprocated links, one way links, and unlinked pairs). A second classifier, which we used in the final application, distinguished simply between linked (undirected) and unlinked pairs. All classifiers were trained with 10-fold cross validation.

A quick initial experiment on three-way classification did not perform well (57% accuracy, C = 32, γ=2.0, 300 training, 1964 test samples). We believe this to be a result of the sparseness of timing data which is useful for determining link directionality.

The classifier used in the final tool was the two-way classifier trained on 3572 examples and tested on 1485 with 10-fold cross validation. With an optimal C = .03125, γ=2.0, the classifier yields an accuracy rate of 91.2458%.

Notably, a simple classifier that only takes into account whether blogs shared any had an accuracy rate of 88% links (i.e. classify as *linked* if the two blogs share one link in common, otherwise classify as *unlinked*). However, as we are frequently interested in the most likely link a classifier outputting scores or probabilities is preferable. Additionally, we are pursuing other classifiers that may better reflect information transfer rather than the simple linked/unlinked state [2].

3.3 Visualization tool

Using the various pieces of data, both explicit and inferred, we have created a web service at: http://www.hpl.hp.com/research/idl/projects/blogs/). Through the service users can enter a URL finding matches in the May crawl data. The resulting visualization allows users to quickly explore the potential routes of infection over both explicit and implicit (i.e. inferred) links.

Graph inference is achieved through the use of the classifier. For all blogs that are not connected to another blog at an earlier date the classifier proposes links. This is done by comparing the blog to all the blogs mentioning the URL on the same day or earlier. The SVM classifier then predicts which links "should" exist. The behavior of the inferred graph creation can be adjusted to specify a confidence threshold, the maximum number of inferred links, and certain display parameters.

Graphs are generated using the Graphviz tool [8], which allows for easy creation of timeline style figures. The coordinates determined by Graphviz are used to render the graph in Zoomgraph [3], a Java-based tool developed to visualize and explore graph structures. Users can use the Zoomgraph applet

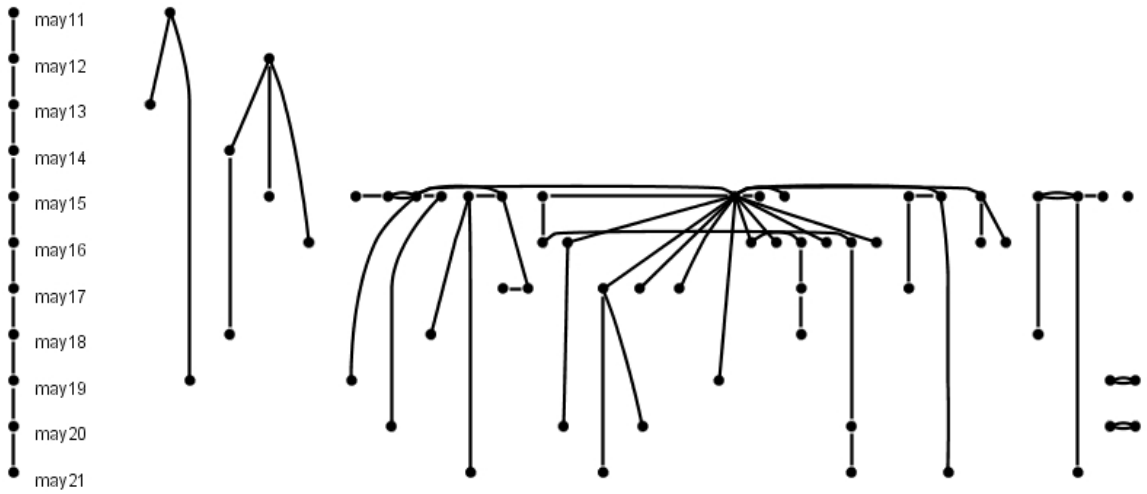

Figure 5: A visualization of all blogs linking to the Giant Microbes site with all explicit and implicit links.

to control the threshold for the display of links as well as the types of links that should be displayed (i.e. "via", explicit, or inferred). Figure 5 is a visualization of the Giant Microbes "epidemic" with implicit links.

4. iRank

In generating the implicit link structure it became apparent that these new graphs differed from the explicit structure of blogspace. These differences may alter the results of ranking algorithms that have largely depended upon the explicit link structure (e.g. www.technorati.com) or have treated blogs in much the same way as regular web pages (e.g. Google's PageRank ([16]). Since a well-read blog may obtain information from a less popular (less linked) source we may miss out on blogs that initially spread infection. To overcome this problem we utilize the timing techniques described above to infer implicit links between blogs and rank the blogs according to those implicit links.

4.1 The algorithm

For the purpose of ranking blogs we do not attempt to infer the most likely infection link but rather all possible infection routes. Thus for a given URL a directed link is constructed for any two blogs b_i and b_j where b_j cites the URL before or on the same day as b_i. As not all links are equally possible, edges are weighted in the following way.

We would like to take into account the diminishing infectiveness of information over time. That is, blogs who have more recently mentioned a URL are more likely to infect as blogs push old citations into their archives. Thus we choose weights, shown in Table 2, that decrease as the number of days between citation increases.

For each pair of blogs b_i and b_j that cite a URL u, we assign a weight, $w_{ji}^u = w(\Delta d_{ji})$, where $d_{ji} = d_j - d_i$. Specifically, if the blog b_j cites n_j URL's, the normalized link weight is assigned as

$$\frac{w_{ji}^u}{n_j \sum_k w_{jk}^u}$$

After all the URL's are considered, edges are merged to form a directed graph by adding up the weights on the same edge. The above weight assignment ensures that the total weights on the outgoing edges of a blog sum up to 1. We call this graph the *implicit information flow graph*. Once constructed, we compute the PageRank on the implicit graph, or equivalently, the stable distribution of the random walk defined by the graph. The rank obtained is an indication of the importance of a blog in terms of infectiousness. In PageRank, a website has high rating if many web sites point to it or if some important web sites point to it. Correspondingly, there are two ways for a blog to boost its iRank ratings:

- Being a popular blog that causes other blogs to discuss whatever URL they mention.
- Mentioning a URL right before it booms globally

Although in our rankings we cannot distinguish which is the case, the first relates to how effective a blog is as an information broadcaster, and the latter to how effective a blog as an information source.

4.2 Spam control

iRank is susceptible to two forms of spamming in which a website attempts to inflate its rank. Some users may duplicate a blog many times, creating artificial popular URLs. Alternatively, a user may automatically list many fresh links on the site each day in the hopes of being considered infectious.

To counter the first type of attack, we modify the algorithm to filter out the URLs that are not sufficiently cited (i.e. have not reached a certain popularity) and only consider what is left, essentially the *effective URLs*. For the data set we have, it is sufficient to set this threshold to be 10 or 20 citations to counter spamming by duplicate sites. It is still possible to spam iRank by more duplicates, which is similar to the scheme of creating a cluster of webpages to spam Google-like ranking algorithms. A more effective method to deal with such spamming is by detecting the cluster of blogs which consistently mention similar set of memes and giving less weight to those intra-cluster links.

To deal with the second type of attack, we take into account how effective a blog detects popular URLs. We simply multiply the rating by a ratio determined by the fraction of effective URLs cited by a blog. Our experiments with both real and artificial blogs show that this modification counters spamming from blogs that list new URLs indiscriminately.

4.3 iRank results

Tables 3 and 4 respectively show the top sites rated by the PageRank algorithm with escape probability 0.1 and the top sites as selected by our iRank algorithm. We observe immediately that there is no overlap in the top 20 sites and hence that the results are very different: PageRank tends to assign high ratings to the sites that contain original materials. iRank tends to assign high ratings to the sites that serve as a community portals, such as lists of popular URLs, lists of important blogs, and discussion boards. Another difference is in the way PageRank and iRank treat duplicate blogs. If a website is just a mirror of another site and the mirroring is done instantly, then those two sites will receive the same iRank rating because from their contents, we cannot distinguish them. However, they can have very different PageRank because the other websites may only have explicit links to the primary sites. Note that websites specifically aimed at finding the newest

Δd	0	1	2	3	4	5	6	7	>7
Weight, $w(\Delta d)$	2	7	6	5	4	3	2	1	0

Table 2 The weights assigned to links depending on the difference between the days blogs cite a link (Δd). The low weight assigned to $\Delta d=0$ is intended to control for the fact that crawls are done on a daily basis and ordering in this case is unclear.

Table 3. PageRank ordering of blogs

Rank	Web Site	Rating
1	boingboing.net	166
2	penny-arcade.com	166
3	caoine.org	151
4	slashdot.org	150
5	andrewsullivan.com	117
6	perversiontracker.com	114
7	crazyapplerumors.com	107
8	bloghop.com	106
9	livejournal.com	101
10	dear_raed.blogspot.com	94
11	girlsarepretty.com	85
12	fark.com	83
13	cyberlaw.stanford.edu	80
14	alwayson-network.com	76
15	oddtodd.com	74
16	instapundit.com	70
17	drudgereport.com	70
18	metafilter.com	68
19	wilwheaton.net	64
20	alternet.org	61

Table 4. iRank ordering of blogs

Rank	Web Site	Rating
1	blogosphere.us/trends.php	40
2	blogdex.media.mit.edu	32
3	blogosphere.us/trends.php?type=hours	25
4	www2.meeka.dyndns.org:81/~alyssa/	21
5	heliopod.org	21
6	indrasweb.com/blog	18
7	weinstein.org	18
8	pontobr.org	18
9	peiblog.psychoblogger.com/weblog.php	18
10	knoxgeek.com	18
11	ctdata.com	17
12	mosa.unity.ncsu.edu/brabec	17
13	timbu.org/mtblog	16
14	khader.net	15
15	burngreave.net	14
16	forum.b0rken.dk/drupal	13
17	vazdot.info	12
18	nanodot.org	12
19	opencontentlist.com	12
20	inhale.org	12

information, e.g. Blogosphere, Blogdex, appear at the top of the iRank list while receiving very low ratings in PageRank.

5. Conclusion and Future Work

In this paper we have studied the propagation of information through blogspace, both to uncover general trends and to explain specific instances of URL transmission. Through the use of cluster analysis we are able to determine general categories of popularity, ranging from sustained interest to short lived events. Digging into the routes of individual URLs we built a tool that is able to visualize and explain how information travels.

Both the availability and quantity of time resolved information is unique to blog data. Using it we were able to not only infer link structure, but also to create a novel ranking algorithm, iRank for ranking blogs. Whereas traditional ranking strategies rely primarily on explicit link structure, iRank successfully folds in implicit routes of transmission to find blogs that are at the source of information. Such, "patient zero" blogs are not always the highly connected, but are nonetheless critical in spreading information. We believe that the inference techniques we have developed can be useful in describing events in the blogspace. In the future we hope to refine these inference techniques to better predict the direction of information propagation. We have also begun exploring how graph structure, both inferred and explicit, is related to the general spread of information and may be a determining factor in how popular or infectious that information becomes.

Despite the success of the simple inference, we would like to augment the simple weighting scheme for the iRank algorithm to include the SVM to calculate the likelihood that one blog obtained information from another. With a larger data set spanning a longer period of time, we would like to apply iRank to study the dynamic change of ratings of blogs.

6. Acknowledgments

We would like to thank Natalie Glance and Intelliseek/BlogPulse for giving us access to their data without which this work would not have been possible.

7. References

[1] Adamic, L.A., O. Buyukkokten, and E. Adar, A social network caught in the web, First Monday, 8(6).

[2] Adar, E., and L. A. Adamic, "Tracking Information Epidemics in Blogspace," working paper.

[3] Adar, E., and J. R. Tyler, "Zoomgraph," working paper, last retrieved from: www.hpl.hp.com/shl/papers/zoomgraph1/index.html

[4] Adler, S., "The Slashdot Effect: An Analysis of Three Internet Publications," retrieved from: http://ssadler.phy.bnl.gov/adler/SDE/SlashDotEffect.html

[5] Butts, C., "Network Inference, Error, and Information (In)Accuracy: A Bayesian Approach," *Social Networks*, 25(2):103-140.

[6] Dawkins, R., *The Selfish Gene*, Oxford Press, 1990.

[7] Dombroski, M., P. Fischbeck, and K. Carley, "An Empirically-Based Model for Network Estimation and Prediction," Working Paper.

[8] Gansner, E. R., and S.C. North, "An open graph visualization system and its applications to software

engineering," *Software – Practice and Experience* 00(S1):1-4 (1999).

[9] Haydon, D.T., M. Chase-Topping, D.J. Shaw, L. Matthews, J.K. Friar, J. Wilesmith, and M.E.J. Woolhouse, "The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mout outbreak," *Proceedings Royal Society B*, 270:121-127.

[10] Kempe, D., J. Kleinberg, and E. Tardos, "Maximizing the Spread of Influence through a Social Network," in Proceedings of KDD '03, (Washington DC, August 2003), ACM Press.

[11] Kleinberg, J., "Authoritative sources in a hyperlinked environment," *Journal of ACM*, 46(5):604-632, 1999.

[12] Kleinberg, J. and D. Liben-Nowell, "The Link Prediction Problem for Social Networks," in Proceedings of CIKM '03 (New Orleans, LA, Novermber 2003), ACM Press.

[13] Kumar, R., J. Novak, P. Raghavan, and A. Tomkins, "On the Burst Evolutin of Blogspace," in Proceedings of WWW '03 (Budapest, May 2003), ACM Press, 568-576.

[14] Manning, C.D., and H. Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA 1999.

[15] NILTE Blog Census, http://www.blogcensus.net/

[16] Page L., Brin S., Motwani, R. and Winograd, T., "The PageRank citation ranking: Bringing order to the Web," 1998, http://citeseer.nj.nec.com/page98pagerank.html

[17] Riolo, C.S., J.S. Koopman, and S.E. Chick, "Methods and Measures for the Description of Epidemiologic Contact Networks," *Journal of Urban Health*, 78(3):446-457

[18] Skvoretz, J., T. J. Fararo, and F. Agneessens, "Advances in Biased Net Theory: Defintions, Derivations, and Estimation," submitted for publication.

[19] Taskar, B., M.F. Wong, P. Abbeel, and D. Koller, "Link Prediction in Relational Data," to appear in NIPS'03 (Vancouver, Canada, December 2003).

[20] Tscherteu, G., and C. Langreiter, "The BlogospherMap," working paper, last retrieved from: www.realitylab.at/blogospheremap/blogospheremap.PDF

[21] Trott, M., and B. Trott, "A Beginner's Guide to TrackBacks," last retrieved from: http://www.movabletype.org/trackback/beginners