

Looking at the blogosphere topology through different lenses

Xiaolin Shi
Dept. of Electrical Engineering
and Computer Science
University of Michigan
Ann Arbor, MI 48109
shixl@umich.edu

Belle Tseng
NEC Labs America
10080 N. Wolfe Rd, SW3-350
Cupertino, CA 95014 USA
belle@sv.nec-labs.com

Lada A. Adamic
School of Information
University of Michigan
Ann Arbor, MI 48109
ladamic@umich.edu

Abstract

The blogosphere is a vast and dynamic complex network. Any examination of the structure of such a network is dependent on the selection of blogs sampled and the time frame of the sample. By comparing two large blog datasets, we demonstrate that samples may differ significantly in their coverage but still show consistency in their aggregate network properties. We further compare the structure of a blog dataset with and without spam blogs, which account for a majority of the links in one sample. We also show that properties such as degree distributions and clustering coefficients depend on the time frame over which the network is aggregated.

Keywords

blogosphere, topology, temporal

1. Introduction

The blogosphere serves as a medium for self expression, community formation and communication, and information diffusion and aggregation. It is a vast dynamic and growing network, with new blogs continuously emerging, millions of existing blogs creating new content, while some lay abandoned as their authors start other blogs or activities. Of particular interest are the direct citation patterns between the blogs, because they indicate interaction and information diffusion - blogs linking to posts they read on another blog while possibly writing additional content of their own. Tracking information diffusion in the blogosphere is not just an intriguing research problem, but is of interest to those tracking trends and sentiments. Several online services, such as BlogPulse and Technorati, report the most actively discussed topics in the blogosphere. A heavily blogged topic, even if it originates in the blogosphere, is likely to make its way into the mainstream media. In fact, many mainstream media sources now host blogs as an integral part of their websites, while some of the most popular blogs rival most mainstream media online outlets in popularity [4].

The rich structure of the blogosphere has proven to be fertile ground for exploring research questions from a variety of fields. Some have focused on the motivations behind blogging [7], the relationship between a person's tendency to keep blogging and their embeddedness in the online so-

cial network [18], and exploring the possibility of extending blog's interactive nature for research and commercial collaboration [6]. A few studies have specifically focused on the LiveJournal blog network and found patterns of link distribution across geography [20], factors contributing to link formation such as common interests and age [16], and even the likelihood of a blogger joining a new LiveJournal interest group if many of their blogging friends have [5]. Others, closer to our current goals, have pursued a systematic approach of analyzing the large scale network structure of the blogosphere. Kumar et al. examined the structure of the blogosphere, both in terms of the bursty nature of linking activity, the uneven distribution in the concentration of such links, and the effect of time windowing on the appearance of that distribution [15, 16, 17]. Information diffusion studies have aimed to use the link structure and other blog properties to infer the path of information flow [12, 1]. Moreover, other studies have used the link structure to solve problems such as splog detection [14] and community identification [28]. Splogs (also known as spam blogs) are blogs whose sole purpose is to direct traffic and increase the search engine rankings of particular websites.

In this paper, we have two objectives. The first is to examine how robust the features of the blogosphere are when examined through the lenses of two different samples. The second is to compare these features with previously studied Web and social network datasets, in order to understand the blogosphere network structure in the wider context of other social and technological networks. Our blog datasets stem from two sources, BlogPulse and TREC (described in more detail below), both intended for use by the research community to study different aspects of the blogosphere. We compare these two sets of blogs directly, first in terms of their coverage and overlap, then in terms of their network properties.

We find that although the datasets differ widely in size, cover different time durations, and are set months apart, their properties show remarkable consistencies. Unfortunately, a fair fraction of blogs are in fact spam blogs, automatically generated blogs created with the intention of altering search engine results and directing traffic to specific websites. These splogs account for a large fraction of the links in the datasets. Consequently, we also study the effect of splog removal on the properties of the networks. Furthermore, we examine the effect of aggregating the network over time, similar to previous work by Kumar et al. [17], and find that the degree distribution and other properties converge when the network is aggregated. Finally, we contrast the linking patterns within

and between different blog hosting sites, finding that most large blog hosting sites tend to be “exporters” of links - with many of those links going to blogs with their own domain names.

2. Datasets description

We use two datasets in our study of the blogosphere. One is the WWW2006 Weblog Workshop dataset from BlogPulse, which has 1,426,954 blog URLs in total, and 1,176,663 distinct blog-to-blog hyperlinks. This dataset covers 3 weeks of blogging activity, from July 4, 2005 to July 24, 2005. It contains hyperlinks that occur only in the blog entries themselves, and exclude blogrolls or comments. Consequently, the network is quite sparse – among the over 1.4 million blogs, only around 141,046 (10%) of them have links to other blogs in the dataset. If we only consider the blogs having at least one in-link (receiving a citation from another blog) or out-link (giving a citation to another blog) in the dataset, the average degree of this network is $\langle k \rangle = 4.924$. We omit from our analysis the additional 160,670 URLs, that were identified by BlogPulse to be blogs with at least 1 citation, but whose entries were not included in the dataset.

The TREC (Text REtrieval Conference) Blog-Track 2006 dataset is a crawl of 100,649 RSS and Atom feeds collected over 11 weeks, from December 6, 2005 to February 21, 2006. In our experiments, we removed duplicate feeds and feeds without a homepage or permalinks. We also removed over 300 Technorati tags (e.g. Technorati.com/tag/war_on_terror), which appear to be blogs, but are in fact automatically generated from tagged posts. Different from the BlogPulse data, the TREC dataset contains hyperlinks of various forms, including blogrolls, comments, trackbacks, etc. There are 198,141 blog-to-blog hyperlinks in total, and 33,385 blogs having at least one such link. However, in order to do a fair comparison of the two different blog datasets, we restrict the TREC data to only the 61,716 hyperlinks occurring within entries. There are 16,432 making or receiving at least one such link, giving us an average degree of $\langle k \rangle = 3.8$. The work of [22] describes the creation of the TREC data in more details, and reports some statistics about this dataset, such as the degree distribution.

Aside from the differences in the sizes and time spans of the two blog datasets, the nature of the two corpora and the way they are constructed are also different. The BlogPulse dataset is more like a complete snapshot, while the TREC dataset is more biased and artificially sampled. Considering all these factors, one of the main purposes of our work is to explore how these factors would affect the observations of blogosphere.

Some previous work has identified a certain fraction of splogs in these two datasets. In BlogPulse, according to the splog detection methodology presented in [14], the percentage of splogs is 7.48%. And in TREC, the percentage of splogs is about 18% [22], while after restricting the blogs to those have homepages, the percentage of splogs detected was around 7% [21].

2.1 Dataset overlap

The BlogPulse and TREC datasets are two samples of the same blogosphere, albeit of vastly different sizes, covering different time durations and about 5 months apart. We are interested in comparing them in two respects in order to assess the difficulty in obtaining a comprehensive sample of the

blogosphere. First, we compare the coverage of the two sets according to different blog hosting sites, as shown in Figure 1. In the TREC dataset, a smaller fraction of the blogs is hosted by the major blog hosting sites. The largest subset at 28% is hosted by LiveJournal, followed by 6% hosted at TypePad. In contrast in the BlogPulse dataset, a full 48% is hosted at LiveJournal, followed by 20% hosted at Xanga.

Second, we directly compare the overlap between the two blog datasets in terms of the blogs commonly crawled by both. Figure 2(a) shows that of the 16,432 blogs whose entries were included in the TREC dataset, 7,225 (or 44%) are also in the much larger BlogPulse dataset. Finally, we take this common set of blogs and compare the overlap in the undirected edges in the two datasets. Specifically, if blog A cites blog B (or vice versa) during the 3 week period covered by the BlogPulse data, we examine what percentage of the time we also observe blog A citing blog B or vice versa in the 11 week period of the TREC crawl 5 months later. Somewhat surprisingly, we find very little overlap. There were only 2823 pairs with edges in both datasets, compared to 56,387 pairs with edges in the BlogPulse data and 57,091 in the TREC data. This means that the same blogs that might be mentioning one another during one short period have only a 5% chance of doing so about half a year later. The above shows us that two relatively large datasets representing “samples” of the blogosphere actually have dramatically different coverage of blogs. Even where the two network samples overlap in nodes, we find that the connectivity, namely the links between the blogs, are likely to change substantially over time.

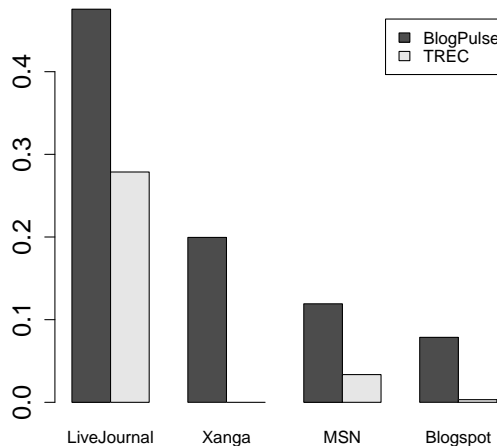


Fig. 1: Proportion of blogs at 4 large blog hosting sites over the two datasets, demonstrating that TREC is less concentrated at large hosting sites than BlogPulse.

2.2 Other network datasets

To understand the properties of the blogosphere and how they differ from other networks, we study similar features in the Web graph, which was presented in [3] and [8]. The former dataset contains 325,729 documents and 1,469,680 links taken from a 1999 crawl of the `nd.edu` domain. The latter crawl from 2000 contains 200 million web pages and 1.5 billion links.

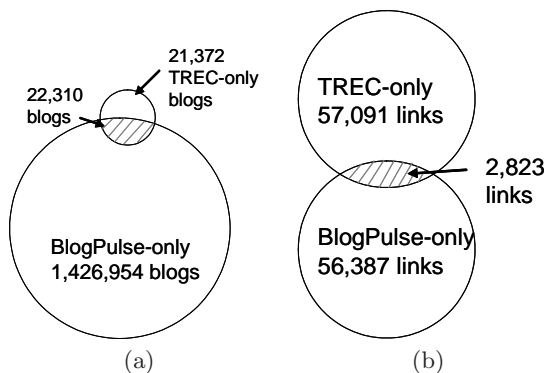


Fig. 2: *Overlap in coverage between TREC and BlogPulse: (a) overlap in crawled blogs with around 50% of the blogs covered in TREC being covered in the BlogPulse sample. (b) overlap in between-blog links in the TREC and BlogPulse datasets restricted to blogs that occur in both.*

3. Topological features and network comparisons

In this section, we study the properties and topological features of the blogosphere by analyzing the two networks constructed from the BlogPulse and TREC datasets. We first restrict our analysis to those links that are located within crawled entries and cite blogs within the data set. We then include any additional hyperlinks, such as blogrolls, comments, and trackbacks, that were included in the TREC dataset. The networks are treated as directed but unweighted graphs where we are simply taking into account whether a blog cites another blog, and not how many times it does so.

3.1 Degree distributions

In a directed graph, for a vertex v , we denote the *in-degree* $d_{in}(v)$ as the number of arcs to v and the *out-degree* $d_{out}(v)$ as the number of arcs from it. The distribution of in-degree p_{in}^k is the fraction of vertices in the graph having in-degree k and p_{out}^k is the fraction of vertices having out-degree k . If both the in-degree and out-degree of a vertex are 0, then the vertex is *isolated*.

The *average in-degree* is:

$$\langle k \rangle_{in} = \frac{1}{|V|} \sum_{v \in V} d_{in}(v) \quad (1)$$

which is a global quantity but measured locally. The *average out-degree* $\langle k \rangle_{out}$ is defined similarly.

First, we observe that the degree distributions are greatly affected by the existence of splogs. Considering all the blogs in the BlogPulse data, both in-degree and out-degree distributions have an unusually high number of blogs with degrees ranging from 10 to 500. This results in irregular shapes for the cumulative degree distributions, which represent the proportion of blogs having at least k in-links or out-links. However, after removing splogs identified by Pranam et al. [14] for the BlogPulse dataset, we replicate the result that the cumulative in-degree and out-degree distributions show smoother curves, as shown in Figure 3.

After excluding splogs from the BlogPulse data, we compare the degree distributions of the blogosphere and the Web, using the Web degree distributions measured by Broder et al. [8] for a 1999 Alta Vista crawl of 200M pages. This pre-

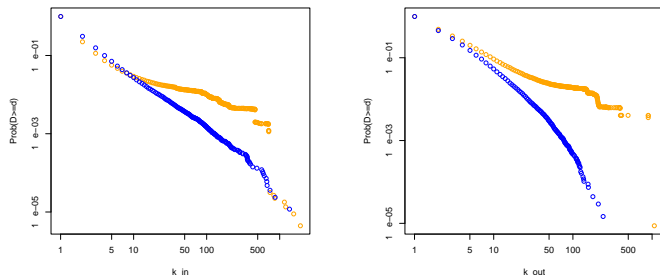


Fig. 3: *The degree distributions of the BlogPulse data with splogs (orange curves) and without splogs (blue curves).*

vious study, along with a study of the `nd.edu` domain [3, 2], and a crawl in 2001 of 200M pages by the WebBase project at Stanford [9] found that the indegree distribution of the Web is scale free with a power law exponent α of 2.1. From Figure 4, we can see that the in-degrees of the BlogPulse and TREC datasets show similar power-law distributions to the Web graph. TREC exhibits a slightly shallower slope, while the BlogPulse data a slightly steeper one. This is consistent with the previous finding that sampling a power-law network can produce networks with steeper power laws [27]. Since the BlogPulse data is of a shorter time duration than the TREC data it may be more likely to resemble a subsample of the full network. Of course it is difficult to directly compare a web page crawl which contains a single static snapshot of a page, with an aggregation over 11 weeks of an RSS feed for a blog. Although a single download of a blog would usually contain a limited number of entries (with previous ones usually moved to an archive), the RSS feed would correspond to a single long page where content is added over time, but not deleted. It is possible that this aggregation over a longer time period accounts for the similarity of slope for the TREC data compared to the Web.

The Web outdegree distribution has been found to either not follow a power law distribution, or to exhibit a steeper power law only in the tail. Broder et al. measured the tail to have a power-law exponent of 2.7, Albert and Barabasi measured it to be 2.45 [3, 2], while Donato et al. found it not to follow a power-law at all [9]. The out-degree distributions of TREC and BlogPulse, shown in Figure 5, drop off much more rapidly than the Web graph. On the one hand, this may again be due to sampling. For example, Pennock et al. [25] showed that when certain subcategories of pages are sampled, what starts out as a power-law degree distribution can exhibit sharp drop-offs. Certainly, blogs are only a subcategory of all web pages, and we are furthermore only considering links between a sampled set of blogs. But, more likely, the number of hyperlinks a blog can generate in a limited time period is bounded. This constraint is also observed in many social networks, e.g. co-authorship networks [24]. So while it is possible for one blog to gather much attention (inlinks) in a short time period, it appears less likely for a single blog to lavish as much attention on as many different blogs in the same time period. The same tends to hold true on the web, where some webpages are linked to by thousands of others, but it is much less likely for a single page to contain thousands of hyperlinks.

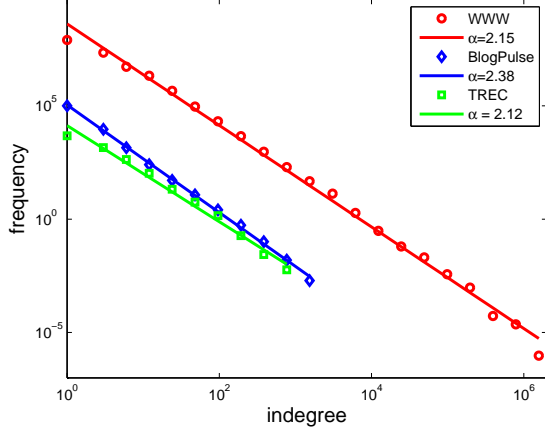


Fig. 4: In-degree distributions of Web, BlogPulse and TREC data, exponentially binned, all showing power-law structure.

Our results also concur with previous measurements of the blogosphere, which have revealed power-law distributions of in-degrees based on blogrolls and in-post citations [17, 20, 26]. Here we were interested in whether we would still observe the power-laws when considering only the in-post citations.

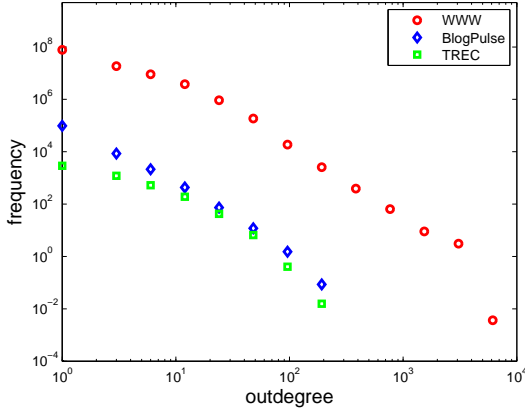


Fig. 5: Out-degree distributions of BlogPulse (blue), TREC (red) and Web (green), exponentially binned.

3.2 Small-world effect

The small world effect states that in the network, the average shortest path between every pair of reachable nodes is short compared to the total number of nodes in the network [13].

The studies of Web graph have shown that the WWW has the small-world property. Even as the number of web pages has grown exponentially, the average number of hyperlinks that need to be traversed to get from one page to any other (provided such a path exists) has remained relatively small. Albert et al. [3] give a formula to compute the average shortest paths in Web graph if the number of web pages N is known:

$$\langle l \rangle = 0.35 + 2.06 \log(N) \quad (2)$$

The estimate for the average shortest path using this formula for a graph of 200 million nodes is $\langle l \rangle = 17.45$, which is quite close to the measured value ($\langle l \rangle = 16$) found by Broder et al. [8] for a data set of 200 million web pages.

Quite similar to the Web graph, our experiments show that even when considering only entry-to-blog links, the blogosphere has the small-world property. Our two datasets, although of different time durations and only partial overlapping in blogs and links, have very consistent shortest paths considering their network sizes. For the TREC dataset (16,432 blogs), it is $\langle l \rangle = 7.12$, and for the BlogPulse dataset, which is of 143,736 blogs, it is $\langle l \rangle = 9.27$. If we let $N = 16,432$ or $N = 143,736$ in Formula 2, then the $\langle l \rangle$ s calculated for TREC and BlogPulse are 9.03 and 10.97 respectively, which are larger than what they have been in our experiments.

However, this does not necessarily mean that it is easy for information to diffuse widely in the blogosphere. This is because information diffusion is not only related to the average shortest path, but also the connectivity of the graph. Since the average shortest path is only computed between all reachable pairs, it doesn't take into account what proportion of pairs of blogs could not be reached one from the other simply by following hyperlinks. Our experiments show that only 12.37% of the pairs of blogs in TREC are reachable. For the BlogPulse data, only 6.13% of the pairs of blogs are reachable. Even when we consider the network of TREC data with all forms of hyperlinks contributing to its edges, the percentage of reachable pairs is still only 22.11%. The low percentage of reachable pairs of nodes is also true in the Web: over 75% of time there is no directed path from a random start node to a random destination node [8]. If the connectivity is low in the network, as it is for the TREC and BlogPulse data, it will yield a small average shortest path, but at the same time produce many infinite paths that are not counted.

In the following section we examine the important question of connectivity in more detail.

3.3 Connectivity

For a directed graph, there are two types of connected components: the *weakly connected component* (WCC) and the *strongly connected component* (SCC). A strongly (weakly) connected component is the maximal subgraph of a directed graph such that for every pair of vertices in the subgraph, there is an directed (undirected) path from v_x to v_y . Thus the weakly connected component is a larger subgraph than the strongly connected component.

In the two blog datasets, within a weakly connected component one can follow links within posts to reach either blog A from blog B or vice versa (but not necessarily in both directions), for each pair of blogs A and B . In practice these paths may be hard to find because the link leading to the path to the second blog could be in any one of the posts made over the 3 or 11 week period. Nevertheless, the connected components give us a sense of the connectedness of the datasets. Our experiments show that, in the TREC data, the largest weakly connected component includes 15,321 nodes, and the largest strongly connected component is of size 2,327. The sizes of largest weakly connected component and strongly connected component of BlogPulse data are 107,916 and 13,393 respectively. We have a comprehensive comparison of the connectivities of blogosphere and the Web in table 1. Similar to the Web [10], the discrepancies in the size of connected components are most likely due to the different ways the datasets

Table 1: *Connectivity comparison between the Web graph and blogosphere samples*

Network	# of nodes	Max WCC	Max SCC	Fraction of SCC in WCC
Web [3]	325,729	325,729 (100%)	53,968 (16.57%)	16.57%
Web [8]	203,549,046	186,771,290 (91.76%)	56,463,993 (27.74%)	30.23%
BlogPulse	143,736	107,916 (75.08%)	13,393 (9.32%)	12.41%
TREC	16,432	15,321 (93.24%)	2,327 (14.16%)	15.19%

are crawled, and the time periods in which the networks form. In section 4, we will further discuss the temporal features of the connectivity of blogosphere.

On the other hand, if we also consider other forms of hyperlinks in TREC, including 33,385 blogs and 198,141 blog-to-blog hyperlinks, then the resulting network has much better connectivity. The size of the largest weakly connected component is 88.93%, and the size of largest strongly connected component is 44.36%. This shows that the blogosphere is glued together by blogrolls, even if over a limited time period there is relatively little active citation.

Another interesting observation about the connectivity of the blogosphere is the following: before cleaning the TREC data, there are 363 **technorati.com** tag URLs, with 47,521 links either from or to these URLs. Our experiments show that the existence of such extremely high in-degree or out-degree nodes does not affect the overall connectivity of the blogosphere. Before removing the Technorati tag URLs and their links, the size of largest weakly connected component is 30,180 (90.40% of the whole network) and the size of largest strongly connected component is 15,176 (45.46%) - only slightly larger than the components with the tag URLs removed. This observation is similar to the one made for the Web graph by Broder et al. [8], showing that high degree nodes do not play the function of “junctions” in the connectivity of the Web.

3.4 Clustering coefficient and reciprocity

The *Clustering coefficient* is a measurement of the percentage of closed triads in a network. For every vertex v_i , its clustering is defined as:

$$C_i = \frac{\text{number of closed triads connected to } v_i}{\text{number of triples of vertices centered on } v_i} \quad (3)$$

Then the clustering coefficient for the whole graph is averaged over all vertices i .

In an Erdős-Renyi random graph (a random graph in which every pair of vertices are connected by probability p) [11] with n nodes and a constant average degree, the clustering coefficient is $O(n^{-1})$. However, in most real-world networks, the clustering coefficient is much higher - $O(1)$, reflects the prevalence of closed triads [23]; i.e. if vertex v_x is connected to vertices v_y and v_z , then the probability for v_y and v_z to be connected is higher than expected at random. For measuring the clustering coefficient in a directed graph, we ignore the directions of arcs.

The clustering coefficients of TREC 0.0617 and BlogPulse 0.0632 (including splogs in both datasets) are large compared with what they are in the corresponding Erdős-Renyi random graphs. We see that for these values of clustering coefficients, the two datasets are showing nice consistency in spite of the differences in crawling and time duration. These high values are also similar to measurements of the clustering for the Web graph ($C = 0.29$ [23]) and co-authorship networks ($C = 0.19$ [24]).

Reciprocity is another measurement that shows a significant difference between real-world networks and the Erdős-Renyi graph. The reciprocity values (how often, when A links to B, B links to A) is another measure of cohesion, reflecting mutual awareness at a minimum, and potentially online interaction and dialogue. In the datasets, we actually observe very little reciprocity: in TREC, 4.98% edges are reciprocal, and in BlogPulse 3.29% edges are reciprocal.

However, if we also consider other types of links in TREC, making the network significantly denser, then the clustering coefficient of this graph is 0.13, and the reciprocity is 20.06%, both of them are significantly larger than they are in the blogosphere merely with entry-to-blog links as its edges. A possible explanation is that people often create entry-to-blog links to cite information. Other types of links, such as comments and trackbacks are by their nature interactive (and trackbacks are by definition reciprocal). Even blog rolls may exhibit higher reciprocity, because bloggers tend to list their friends’ blogs as well as other blogs they tend to read, and friendship is often, though not always reciprocal. Therefore the low reciprocity we observe could be due to the nature of entry-to-blog links themselves and the short time window of the samples, where we simply haven’t waited long enough to observe a reciprocal link.

4. Temporal features

As we have described before, the time ranges of the two datasets are of different lengths: the BlogPulse sample covers 3 weeks, while TREC is crawled over 11 weeks. In order to explore the effects of the crawling periods on the observations of the blog datasets, we take the longer-period TREC dataset and study the properties of the subgraphs in the TREC network over 4 different time windows.

We assign a timestamp for each entry-to-blog link as the time the entry is created, where 74.72% of the entries have timestamp. Four overlapping time periods are chosen corresponding to the first 10, 20, 30 and 40 days of the TREC crawl. The 10 days capture 5,793 blogs and 8,818 entry-to-blog links with an average degree of $\langle k \rangle = 1.5$. The first 20 days capture 8,054 blogs, 16,206 links, bringing the average degree up to $\langle k \rangle = 2.0$. The first 30 days capture 9,085 blogs with 20,411 links and $\langle k \rangle = 2.2$. The last subset of 40 days contains 10,433 blogs and 27,724 links with $\langle k \rangle = 2.657$. This illustrates that as the time duration increases, the average degree also increases.

4.1 Degree distributions

We plot the degree distributions of the four time-overlapping subnetworks (10 days, 20 days, 30 days, 40 days), as well as the entire network of 11 weeks with link time stamps (denoted by “Links with TS”), and the entire network with or without link time stamps (denoted by “All links”). From the in-degree and out degree distributions in Figures 6 and 7, it is apparent that different time windows yield very similar shaped curves

for both the indegree and outdegree distributions. However, as the time periods get shorter, the curves for both in degrees and out degrees are steeper.

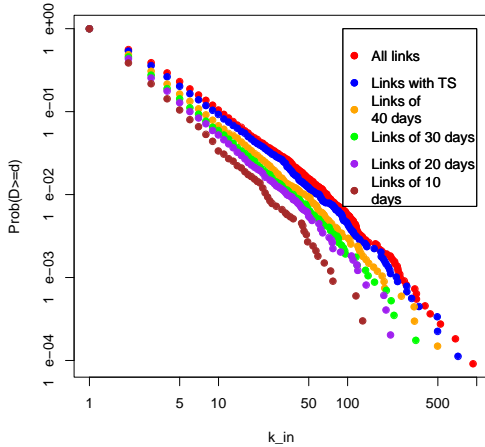


Fig. 6: Temporal changes in the in-degree distributions in TREC.

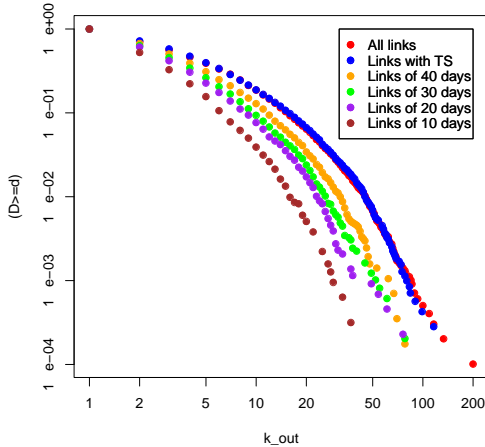


Fig. 7: Temporal changes in the out-degree distributions in TREC.

4.2 Connectivity

In section 3.3, we found that both the BlogPulse and TREC samples have large weakly connected components, but relatively small strongly connected components, with the TREC sample showing better connectivity than BlogPulse, in spite of BlogPulse having larger average degree. The dynamics in the connectivity of blog subgraphs over different time windows is shown in Table 2. As time goes on, both the size of the largest weakly connected component and the size of the largest strongly connected component grow larger, and thus

the connectivity is increasing. It can also be observed that the weakly connected component is formed earlier and grows more rapidly. In contrast, it takes a much longer period for the strongly connected component to form; however, after a certain period of time, the growth of the largest weakly connected component is relatively stable near 100% of the network, while the largest strongly connected component continues to grow.

4.3 Clustering coefficient and reciprocity

Next we examine the temporal changes in reciprocity and the clustering coefficient. Our experiments show that the values of reciprocity of the links from the first 10 days to the first 40 days are 2.88%, 3.85%, 3.84% and 4.12%. We can see that except for the shortest time period, all the other values are bigger than in the 3-weeks of BlogPulse (reciprocity of 3.29%) but smaller than in the entire 11-weeks of TREC (reciprocity of 4.98%). This indicates that reciprocity grows with time, because blogs have a longer opportunity to reciprocate. It also demonstrates that reciprocal links are still extremely sparse.

The clustering coefficients from the first 10 days to the first 40 days are 0.034, 0.043, 0.046 and 0.052. All of them are smaller than the clustering coefficient in both BlogPulse (0.0632) and TREC over the full time period (0.0617). So we know that although longer time would increase the clustering coefficient, it may depend more on the density of the sample.

4.4 Densification law

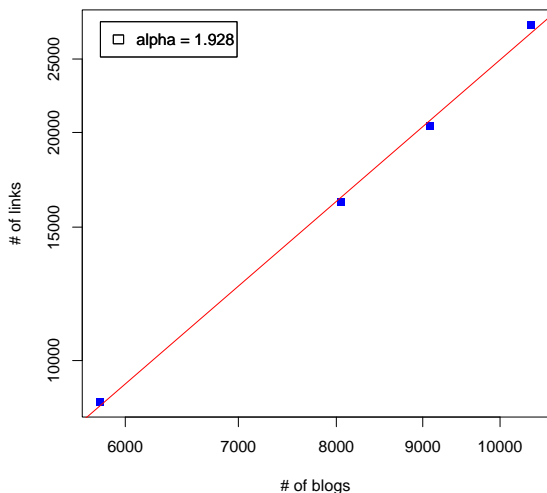
Leskovec et al. [19] described the densification law prevalent in many networks: the number of edges grows superlinearly in the number of nodes over time: $e(t) \propto n(t)^\alpha$. For example, in the Internet, there are new routers appearing and at the same time the number of connections per router is increasing, and the densification exponent is $\alpha = 1.18$. In patent networks, all the links are added from a patent at the time it is inserted into the network. The densification exponent is $\alpha = 1.66$. But in our network, probably most of the blogs already existed before the beginning of the crawl (it would be interesting to repeat the analysis with new blogs appearing). During the crawl, as more and more links are added into the network, originally isolated blogs start connecting to each other. The number of edges, shown in Figure 8 is increasing nearly quadratically with the number of nodes ($\alpha = 1.928$). This relatively large value tells us that the densification of a crawled blogosphere with a static set of blogs is a faster process than some other networks such as the Internet and patent networks.

5. Blogs in blog hosting sites

Another way of understanding the blogosphere is by analyzing it through different blog hosting sites. Currently, the four largest blog hosting sites are LiveJournal, BlogSpot, Xanga and MSN. They are also the largest four in the BlogPulse dataset, as shown in Table 3. In the table, “all links” either originate from or terminate at a blog at the specific blog hosting site; “in links” originate outside of the blog hosting site, but terminate within it; “out links” point from within the hosting site to an outside blog; “internal links” lie between blogs within the hosting site. All these links occur only within entries, and no links of other forms, such as blogrolls, comments, etc. are included. The table also lists in italic the corresponding numbers of blogs and links after remov-

Table 2: Temporal changes in the connectivity in TREC

Subset	# of nodes	Max WCC	Max SCC	Fraction of SCC in WCC
First 10 days	5,793	4,719 (81.46%)	None (0%)	0%
First 20 days	8,054	7,162 (88.92%)	349 (4.33%)	4.87%
First 30 days	9,085	8,249 (90.80%)	471 (5.18%)	5.71%
First 40 days	10,433	9,662 (92.61%)	730 (7.00%)	7.55%
All blogs	16,432	15,321 (93.24%)	2,327 (14.16%)	15.19%

**Fig. 8:** The number of edges versus number of nodes in log-log scale for blogs crawled over different time durations. It obeys the densification power law.

ing splogs according to [14]. An immediate observation we can make is that although splogs by number constitute only a small fraction of the total blogosphere, they account for a substantial proportion of the links.

Table 3: Blogs in hosting sites in the BlogPulse dataset

# Blogs	All links	Inlinks	Outlinks	Internal links
LJ 678,676 <i>676,719</i>	155,665 <i>95,161</i>	4,561 <i>2,718</i>	15,735 <i>8,731</i>	135,369(87.0%) <i>83,712(88.0%)</i>
Xanga 284,693 <i>283,952</i>	58,741 <i>8,454</i>	2,354 <i>437</i>	3,067 <i>1,534</i>	53,320(90.8%) <i>6,483(76.7%)</i>
MSN 170,108 <i>162,147</i>	58,811 <i>1,271</i>	44,180 <i>90</i>	1,528 <i>699</i>	13,103(22.3%) <i>482(37.9%)</i>
BlogSpot 112,184 <i>62,256</i>	845,093 <i>42,830</i>	34,979 <i>12,540</i>	73,730 <i>18,519</i>	736,384(87.1%) <i>11,771(27.5%)</i>

From Table 3, we also notice that for most of these large blog hosting sites, no matter whether or not splogs are removed, the internal links usually occupy the greatest portion of the all links. The percentage of internal links of MSN is relatively small (22.3%). This is because in the BlogPulse dataset, there is a large portion of links from blogs

of BlogSpot to blogs of MSN, which are mostly splogs (over 43,000 links). Due to this fact, it lowers the percentage of internal links of MSN. This is another aspect that tells us how splogs would affect our observations of blogosphere. This suggests that blogs within one hosting site are more likely to form densely connected communities, while it is less likely for blogs in different blog hosting sites to be in a community. This pattern may be a result of bloggers preferring to use the same hosting site as their friends, and different hosting sites being prevalent in different countries. In the Table 4, we can see that links connecting two different blog hosting sites are very sparse, both with and without splog links.

Table 4: Links among blog hosting sites in the BlogPulse dataset

Src & dst Blogs	LiveJournal	Xanga	MSN	BlogSpot
LJ	135,369 <i>83,712</i>	873 <i>159</i>	160 <i>10</i>	4,215 <i>1,714</i>
Xanga	1,208 <i>612</i>	53,320 <i>6,483</i>	124 <i>11</i>	659 <i>236</i>
MSN	61 <i>36</i>	179 <i>23</i>	13,103 <i>482</i>	309 <i>66</i>
BlogSpot	1,109 <i>707</i>	832 <i>151</i>	43,113 <i>17</i>	736,384 <i>11,771</i>

Another thing one observes from the Table 3 is that the numbers of out links always exceed the numbers of in links for a blog hosting site, and most of those out links point to blogs with their own domain names. Since it is easier and often free to create blogs in the blog hosting sites, these kinds of blogs are more casual and personal; in contrast, blogs with their own domain names are more likely to be maintained in a more formal and professional way. And in this sense, it is natural for the self-hosting blogs to have more in links from other blogs.

6. Conclusions and future work

For analyzing the topology of a large network such as the blogosphere, it is impossible for researchers to get all the data about it. Rather, one uses various sampling methods to gather some data, typically a small fraction of the whole network, to analyze. Thus, it is very important to examine how robust the topological features of the blogosphere are when incorporating different time durations and ways in crawling the data. Our work shows that for the two different samples of blogosphere, BlogPulse and TREC, in spite of the low overlap in their coverage and time durations of collecting, some topological features, such as the degree distributions, average shortest paths, connectivity, clustering coefficient and reciprocity are showing great consistency. Our work also shows that as the time duration of a crawl is extended, the features start to converge. This tells us that by obtaining some fairly

comprehensive samples of the blogosphere, one can start to obtain good estimates of the topological features of the whole space.

We also examined the effects of the existence of splogs in the blogosphere, and found that splogs contribute a fair fraction of the total links volume in the blogosphere, and consequently affect the degree distributions greatly. Moreover, by looking at blogs in some large hosting sites, we find that blogs within the same hosting site are more likely to be connected than blogs in different hosting sites. However, this does not mean that there are few links outside of blog hosting sites. Rather many of the links originating at large hosting sites point to blogs with their own domain names.

For understanding the topological structure of the blogosphere, we further compared the features with some other large networks, such as the Web graph and some specific social networks. We find that they share some similarities, such as in-degree distributions, the small-world effect, and overall connectivity. However, they differ in other aspects, such as the out-degree distributions and level of clustering.

As for future work, we are interested in explaining the similarities and differences of the blogosphere and other social and technology networks. What is more, we also would like to know how the information diffusion is influenced by the structure of blogosphere, and how the evolution of blogosphere is affected by the information diffusion in return.

Acknowledgments

We would like to thank Koji Hino and Yu-Ru Lin for preparing and pre-processing the TREC and BlogPulse blog datasets. We also thank Yun Chi for extracting time stamps for our temporal analysis.

References

- [1] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *WWE2006*. ACM Press, May 2004.
- [2] R. Albert, A. Barabasi, and H. Jeong. Scale-free characteristics of random networks: The topology of the world wide web. *Physica A*, 281:69–77, 2000.
- [3] R. Albert, H. Jeong, and A.-L. Barabasi. The diameter of the world wide web. *Nature*, 401:130, 1999.
- [4] C. Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion, July 2006.
- [5] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. *KDD '06*, 2006.
- [6] M. Brady. Blogging: personal participation in public knowledge-building on the web. *Participating in the knowledge society: Researchers beyond the university walls*, 2005.
- [7] M. Brady. Blogs: Motivations behind the phenomenon. *Chimera Working Paper*, (2006-17), 2006.
- [8] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Comput. Networks*, 33(1-6):309–320, 2000.
- [9] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. Large scale properties of the webgraph. *Eur. Phys. J. B*, 38(2):239–243, 2004.
- [10] D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas. Mining the inner structure of the web graph. In *WebDB*, pages 145–150, 2005.
- [11] P. Erdos and A. Renyi. On random graphs I. *Publ. Math. Debrecen*, (6):290–297, 1959.
- [12] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW '04*, pages 491–501, New York, NY, USA, 2004. ACM Press.
- [13] J. M. Kleinberg. Small-world phenomena and the dynamics of information. In *Advances in Neural Information Processing Systems (NIPS)*, page 14, 2001.
- [14] P. Kolari, A. Java, and T. Finin. Characterizing the Splogosphere. In *WWE '06*, May 2006.
- [15] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW '03*, pages 568–576, New York, NY, USA, 2003. ACM Press.
- [16] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *Commun. ACM*, 47(12):35–39, December 2004.
- [17] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *World Wide Web*, 8(2):159–178, 2005.
- [18] T. Lento, H. T. Welser, L. Gu, and M. Smith. The ties that blog: Examining the relationship between social ties and continued participation in the wallop weblogging system. In *WWE 2006*, 2006.
- [19] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05*, pages 177–187. ACM Press, 2005.
- [20] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *PNAS*, 102(33):11623–11628, August 2005.
- [21] Y.-R. Lin, W.-Y. Chen, X. Shi, R. Sia, X. Song, Y. Chi, K. Hino, S. Hari, J. Tatemura, and B. Tseng. The splog detection task and a solution based on temporal and link properties. In *TREC blog Track*, 2006.
- [22] C. Macdonald and I. Ounis. The TREC Blogs06 collection: Creating and analysing a blog test collection. Tech report (dcs), Dept of Computing Science, University of Glasgow, 2006.
- [23] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.
- [24] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *PNAS*, 101:5200–5205, 2004.
- [25] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles. Winners don't take all: Characterizing the competition for links on the web. *PNAS*, 99(8):5207–5211, 2002.
- [26] C. Shirky. Power laws, weblogs, and inequality. In *"Networks, Economics, and Culture"*. Aula, Helsinki, Finland, 2003.
- [27] M. P. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *PNAS*, 102(12):4221–4224, March 2005.
- [28] B. L. Tseng, J. Tatemura, and Y. Wu. Tomographic clustering to visualize blog communities as mountain views. In *WWE 2006*, May 2006.