

Leveraging contextual information to explore posting and linking behaviors of bloggers

Sofus A. Macskassy

Fetch Technologies

841 Apollo Street, Suite 400

El Segundo, CA 90245

Email: sofmac@fetch.com

Telephone: (310) 414-9849

Abstract—The last decade has seen an explosion in blogging and the blogosphere is continuing to grow, having a large global reach and many vibrant communities. Researchers have been pouring over blog data with the goal of finding communities, tracking what people are saying, finding influencers, and using many social network analytic tools to analyze the underlying social networks embedded within the blogosphere. One of the key technical problems with analyzing large social networks such as those embedded in the blogosphere is that there are many links between individuals and we often do not know the context or meaning of those links. This is problematic because it makes it difficult if not impossible to tease out the true communities, their behavior, how information flows, and who the central players are (if any). This paper seeks to further our understanding of how to analyze large blog networks and what they can tell us. We analyze 1.24M blogs posted by 298K bloggers over a period of three weeks. These bloggers span private blog sites through large blog-sites such as livejournal and blogspot. We first characterize the behavior of bloggers, validating some (but not all) common beliefs about how often bloggers post, how long their posts are, who they link to and how much reciprocity there is in links. We then take a look at bloggers from the larger blog sites to understand whether and how they differ in terms of these metrics. Finally, we extend our analysis to focus on contextual links: what is the textual content of the blog which had a link. We identify topics from the textual content of all the blog posts and use these to tag links based on the topics that were discussed in the blog.

I. INTRODUCTION

The past decade has seen a dramatic increase in online social activities, many of which are publicly observable. This trend, which continues to grow and is likely to continue growing, provides a rich environment in which to explore social behaviors at a scale not possible before. This online social network data is large, often contains contextual information such as text or semantically typed relations, is temporal in nature and is also extremely noisy. All of these provides an extremely rich set of data but also introduces new interesting analytic problems for analyzing these large social networks. However, as the data has become increasingly available over recent years, researchers in computer science, physics, social sciences, and more have started to analyze these networks to identify network characteristics, communities, behaviors, influencers, etc. (see, e.g., [1]–[6]).

There are many types of online social network data available, which span everything from social networking sites such as facebook, myspace and LinkedIn to message boards to publishing sites such as digg and flickr, to public email archives such as the Enron email archive to the blogosphere.

All of these types of data have their own idiosyncrasies and each of these have seen their share of attention from researchers in various fields. In this paper we will focus on the blogosphere, where the data consists of blogs and links between blogs. What makes this data particularly interesting is that it contains rich data in the form of the text and it contains explicit relations between bloggers. Further, there are millions of people blogging, generating gigabytes of data on a daily basis, providing a continuous source of an increasingly complex and rich social network. However, such a large network is also difficult to analyze to gain any insight into the population and their behaviors.

The focus of this paper is to analyze this large social network to understand the behavior of the active bloggers in terms of how often they post, how diverse their topics are, how they link to each other, and the context of such links. In addition to exploring the online behavior of bloggers, we also explore the social network generated by the links between blogs. If we consider only the large social network generated by the links without leveraging the context of the links, we end up with a mostly useless network that provides very little information on communities, influencers, information flow, or a plethora of other analytic questions we can ask of the social network. By explicitly taking context into account we can extract out cleaner and more informative networks and relations such that we can focus our attention on the parts of the network that would be of interest.

The key contribution of this work is to apply text mining on blogs and combine it with social network analysis in order to improve the efficacy and understanding of the behaviors of bloggers. Our approach in this paper will be to use recent text mining techniques for topic detection in large textual corpora, and then “categorizing” blogs based on the topics they most relate to. From this, we can tag links between blogs based on the topics that the categories of the blogs. This results in a large contextually tagged network where we can focus on the small communities that result by only following relations that are categorized with the same topic.

The remainder of the paper is structured as follows: we next describe the data preparation approach used in this paper, including gathering and cleaning the data, text-mining for topic-detection and then how we combine text and links. We then describe the analytics we will perform on the data, including behavioral analysis, comparative analysis between blog-sites, and our network analysis. This is followed by the

analytic study of 1.24 million blogs and 298,000 bloggers, exploring first their online behaviors and then the effect of using contextual information to identify communities. We end with concluding remarks.

II. DATA PREPARATION

Handling real world data is not easy or straight forward and we here describe some of the steps needed in order to gather and clean data for analysis. While these are often not described, we feel that they are an important step in any analysis and that the data preparation must be described such that the full process is understood and reproducible.

In the real world, data is not often clean nor in a proper format for analysis. In our particular setting, we are dealing with blog data which consists of a large set of blogs from a set of bloggers. Each blog is formatted in html and contains html markup (paragraphs, bold-faced text, etc) as well as multimedia (pointers for pictures and movies) and hyperlinks to other web-pages or blogs. Underlying all this data is an emerging social network between these bloggers and the topics they write about.

However, this raw format is not easily ingested by most analytic tools or algorithms, and care must be taken to get the data into a proper format. Specifically, we are interested in the high-level behaviors of bloggers and how they link to each other, and so we will create a semantically rich graph out of the raw blogs.

We do this in four steps: we first gather and clean up the blog data into a format that is appropriate for text and link mining, we then analyze the text to identify general topics within the set of blogs, and we then finally tag blogs and links between blogs with these topics. We next describe each of these steps in some detail.

A. Blog Gathering and Cleaning

We first need to gather blog data. While there are many snapshots of data available on the net, we decided to gather our own data set from a variety of sources. We use a commercial system—the FetchBlogs system from Fetch Technologies¹—to gather blog posts from millions of bloggers. As blog-posts reside within a web-page, they must first be extracted appropriately. For example, a rather simple page such as the one shown in Figure 1 shows the top-part of a blog, but also all the extra information on the page itself: title, other sections (e.g., “about me” and “I’ve read”), extra information on the right (e.g., “bilingual comments”, “tags”). Not shown in the figure is the plethora of other exogenous information later on the page, including a calendar, a link to archives, other blog postings by the same blogger, etc. None of this information is relevant to the single blog post shown in the figure, and care must be taken to extract only the blog-post and links within the blog-post itself.

While comments and HTML markup (bold-faced text, images, ...) are clearly indicative of interest and meaning, they are also difficult to work with and so we will ignore these. For the purposes of the study in this paper, only the text and the links to other blogs are of interest.

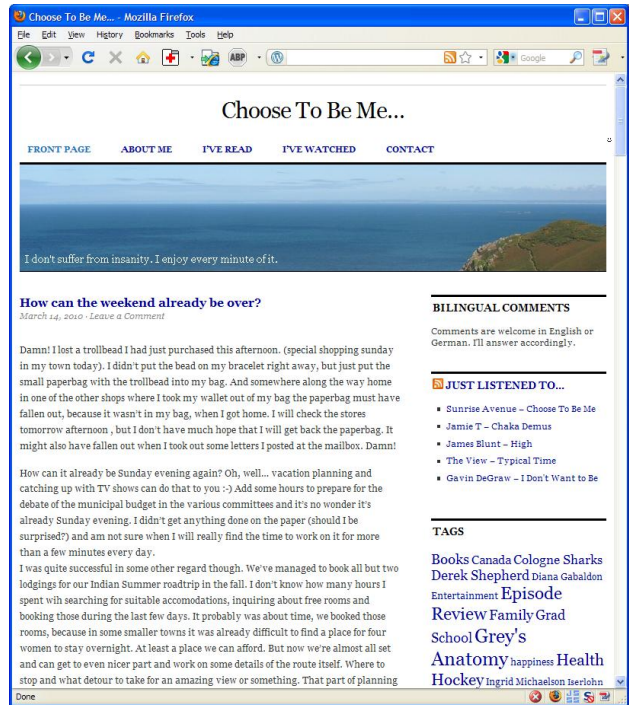


Fig. 1. Example blog from wordpress.com.

Finally, in order to create a social network, we need to refine the links such that we can identify which bloggers link to whom. We do this by analyzing the url itself to identify the blogger. For example, links to livejournal blogs contain the blogger userid as part of the URL (<http://user.livejournal.com>).

The working data we are left with is a set of blog-posts (text only) and the links from the blogger of the blog-post to other bloggers (and non-blog web-pages as well, although we do not use those in this paper). The posts also contain extra meta-information such as the blog userID, the date of the post and the length of the post.

B. Text-Mining

When dealing with millions of blog posts, we get a very large text corpora that may not be all that informative in its raw format and for our purpose. In particular, we are interested in meta-level behaviors such as what topics are being posted as well as the textual context of links between bloggers. To this end, we categorize each blog post into one or more topics that the blog post covers.

We do this by using an unsupervised topic-detection and clustering algorithm. Specifically, we use a method known as *Latent Dirichlet Allocation* (LDA) [7], which models a text corpora as a set of topics. LDA is a generative probabilistic model of a corpus, represented as a graphical model in Figure 2. The generative process is as follows: for each blog, m , choose a topic distribution θ ($\theta \sim \text{Dir}(\alpha)$). For each word w in the blog, first use θ to choose a topic z and then using β distribution of words, conditioned on z , choose a word. Given a corpus D of blogs, the modeling parameters for LDA are the four white circles: α , β , θ and z , since we are given all the

¹<http://www.fetch.com>

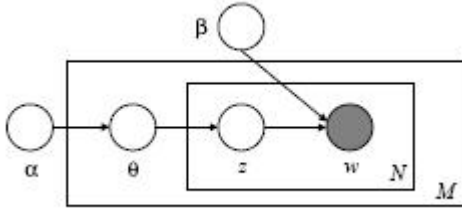


Fig. 2. Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer box represents the M blogs, and the inner box represents the N words for each blog.

w 's in the corpus. We refer the reader to [7] for more details.²

Given a number of topics, LDA will find the best parameters to fit the corpora into that number of topics. However, one large problem of LDA is estimating the number of topics. We do not address that problem here and instead view the number of topics as a way to make the data more granular to fit the requirements of our study. For our purposes, we found 1000 to be at a good granularity.³

LDA generates a language model over all observed words. However, our blog data contains over 2.5 million unique words, which makes this a very large modeling problem. We therefore first reduce the dimensionality of the problem by applying stemming [9] and stoplisting. Stemming transforms words into their stem (e.g. “running” and “ran” become “run”). Stoplisting removes words that are not normally informative such as pronouns and articles (e.g., “I” and “the”). We further remove words that only occur infrequently (less than in 100 blogs) or very frequently (in more than 10% of the blogs). Doing this, we end up with a set of 34K words.

The result of running LDA on our blog data is that each blog is “fitted” to each topic—i.e., each blog belongs to each topic with some likelihood (many of which are near-zero). Each topic is automatically found using LDA and is represented as the set of words most likely to occur for that topic. We tag a blog with the 5 most likely topics that it covers, keeping track of the likelihoods because we use them later.

C. Tagging Links

In order to create a semantically rich social network, we tag links between bloggers based on the context of the links. Specifically, we do not consider all links as equal because there is some reason behind links being created and if we can infer this reason or context, then we can better use the link to understand how bloggers group together in meaningful ways. Further, it has been observed that as graphs become larger and more connected, identifying small high-fidelity groups becomes increasingly difficult [10]. Our approach to handling the problem of large connected graphs is to tag links based on

²We chose to use a standard version of LDA because there are efficient implementations of it available which can handle relatively large corpora. We here use MALLET [8], an open-source language modeling toolkit, because it is specifically designed to handle large corpora efficiently. Other public versions of LDA could not handle the size of our data set.

³We tested with smaller number of topics down to 100 and found qualitatively similar results as those reported here.

the topics of the blog that is the source of the link.⁴

Specifically, we use the tags of a blog created by text-mining to tag the links. Some blogs are very topic-specific and only one topics has a high likelihood whereas other blogs are related to more topics. We want to handle links from each of these types of blogs differently. Specifically, we prune the topics (max of 5 from our text mining) down to only the topics that a blog is most closely associated with. We do this as follows, for each blog b :

- 1) $t_b = \{\}$
- 2) $t'_b = \{t_1, \dots, t_5\}$ where $p(t_i) > p(t_j), \forall i < j$, and $p(t_i)$ is the likelihood that blog b covers topic t_i .
- 3) $i = 1$
- 4) $t_b = t_b \cup \{t_i\}$
- 5) $i \leftarrow i + 1$
- 6) if $i \leq 5$ and $p(t_i) > 0.9 * p(t_{i-1})$, repeat from step 4

The result is that t_b is the set of topics most closely aligned to blog b . Each link in the social network is generated by a blog b , where the source node of the link is the blog-userID of b and the destination node is the blog-userID of the blog-post being pointed to. The tag for this link is t_b .

The result of this step is therefore a semantically rich social network, where links are directed and tagged with the topics of the blogs that generated the links.

III. ANALYTICS

The data which we will analyze has two forms: the blogs and their meta-information (such as blog user ID, topics, date, length of post), and the social network generated by the linking between blogs. Each of these two forms of data enables us to analyze the behaviors of the bloggers in a variety of dimensions.

We perform four analytic studies:

- 1) **Posting behavior:** How often do bloggers blog, how many topics do they cover, etc.
- 2) **Topic behaviors:** Are topics short- or long-lived and do they tend to have spikes.
- 3) **Linking behavior:** How often bloggers links, to whom they link, the propensity to link to one's own blogs, etc.
- 4) **Community-detection:** Which groups are formed by the links and how does the size and fidelity of components and groups change if we only consider topic-specific links instead of all links.

Because the data we collect contain bloggers from a variety of large blog-sites such as livejournal and wordpress, we will be analyzing not only at the micro-level of individual bloggers, but also at the higher-level across blog-sites to see if there are fundamental differences in behaviors of bloggers across sites.

A. Posting Behavior

Our first behavioral analysis of bloggers is based on an aggregate of bloggers individual posting behaviors. Firstly, we are interested in identifying whether there are any patterns in the frequency and the days bloggers tend to post.

⁴We realize that we could also look at the content of the destination blog. However, those blogs were not part of our data set and so we were unable to do that particular analysis.

We are specifically interested in tracking the following behaviors:

- 1) When do bloggers post? Are they more likely to post on a Monday or during the weekend? We track this by counting up the number of bloggers that post on a particular day and then plot a timeline (x -axis), where the y -axis is the ratio of all bloggers that blogged on a particular day. We would expect to see dips on the weekend, where fewer bloggers are active.
- 2) How often do bloggers post? Do they tend to post infrequently (once a month, once a week?) or do they tend to blog every day? We expect to see that most bloggers are infrequent while a few bloggers post every day. This would suggest, as has been often publicized, that a few bloggers are responsible for the majority of blogs. We explore by plotting on the x -axis the number of days a blogger is active and on the y -axis how many bloggers were active for that many days.
- 3) Do bloggers tend to write large or small blogs? Again, we expect a variety of bloggers, some tending to write larger or smaller blogs. This may offset to some extent the prolificness of bloggers if the infrequent bloggers tend to write larger blogs, then perhaps the frequent bloggers will have less of a majority in overall volume. We plot, for each blogger, the number of posts versus the average post size, which will tell us how those two interact.
- 4) Do bloggers tend to stick to a few topics or are they more likely to write about multiple topics? If bloggers tend to stick to a few topics, this would suggest that there are “experts” in the general blogosphere outside of the more professional bloggers. We plot, for each blogger, the number of posts versus the number of topics covered. If the plot shows any points around only a few topics, even for prolific bloggers, then we can say that there are a few experts or consistent bloggers in our period. Given that we are looking at the general population and using general topics, we would expect bloggers to cover a wide variety.

B. Topics

Our second behavioral analysis focuses on the “behavior” of topics based on our text-mining results. Treating topics as first-order objects, we here ask three questions:

- 1) How long do topics last? Do they tend to be short-lived or are they more persistent? Often very specific topics tend to be short-lived, but is this also true for more general topics such as those found by LDA? We identify topic-behavior by plotting how many topics had posts for 1, ... k dates
- 2) How popular are topics? As with bloggers, are there topics that tend to dominate the blogosphere or are the 1000 general topics all equally popular? We would expect there to be more popular topics amongst the 1000 topics found, and if so, then it would suggest that LDA could be used to find topics that are probably not part of the topic-segmentation normally used. We address this question by plotting, for each of the 1000 topics

found, how many posts each topic received, sorted by the number of posts the topics receive. If we see a linear and flat line, then all topics are equally represented. However, if we see a log-scale curve, then we can see that a few topics are dominating.

- 3) Finally, we consider the activity of topics—do they generally have spikes of activity or are they generally active. We explore this by plotting, for each topic, how many posts discussed the topic on any given day. Whether a topic has spikes or not is interesting, because if we see no spikes, then that would suggest a way to rapidly identify if a topic is being picked up. On the other hand, if we do see spikes, that would suggest that topics are discussed quite a bit for a little while and then diminish. If the topic is still active the whole time, then that is interesting because it suggests that the topic did not completely leave the collective attention.

C. Linking Behavior

The third analytics we will report on in our study below is the linking behavior of bloggers and blog-sites and how these behaviors can be used to identify close-knit groups in the blogosphere beyond what is currently possible.

Specifically, we will report on the following:

- 1) How often do bloggers link: What is the likelihood that a blog has a link to another blog?⁵ We can compute this likelihood by computing the ratio of blogs that have links.
- 2) For a given link, is it more likely to be to a previous post by the same blogger or some other blogger? In other words, are links more likely to be self-links? We compute this metric by computing the ratio of links that are links to one self.
- 3) If the link is to another blogger, is it more likely to be on the same blog-site as the blogger or to a different blog-site? This will tell us something about the cohesiveness of a blog-site. If bloggers are more likely to link to within the same blog-site, then it indicates that bloggers of similar profiles are more likely to select the same type of blog-site and that blog-sites do indeed cater to different types of bloggers.
- 4) If the link is to another blogger, is the link reciprocated? In other words, how many links are reciprocated. One common myth is that bloggers are becoming increasingly good at reciprocating links and that this is important to get networked and more readers. We here compute the ratio of links that are reciprocated to gauge the prolificness of reciprocity.

D. Group Detection

The final analytics of the social network which we will investigate is group behaviors. Our question is whether focusing on topic-specific relations lead to smaller, yet informative, groups than if considering all links as equal. For example, we expect that if we consider all links as equal that we will get a

⁵We can also check for any link at all and explore links to non-blog sites, but we leave that for another paper study as we are more interested in the social network in this particular paper.

few very large connected components and a larger number of smaller components. The problem with community and group detection on large components is that community detection algorithm (see, e.g., [11]) tend to generate a few large clusters. If we start with a large component, then we will end up with smaller communities detected within the component, yet these are still very large. We could certainly repeat the community detection algorithm to find smaller and smaller communities (see, e.g., [12]). However, these smaller communities are likely to consist of relations that were created from very different blogs and may suggest underlying strong communities but will lose much of the peripheral bloggers that are important to a specific topic. Rather, if we focus on relations that are about a specific topic then we will find much smaller connected components and hence even smaller and tighter communities. Of more importance, bloggers can now belong to multiple components and communities, something not easily achieved if we do not distinguish between relations. We focus in this paper only on the analysis of the connected components.

We here perform a comparative study of the difference in the number and size of components found using generic links versus topic-specific links. One of the key differentiators we will look for is the size of the components and the number of components that bloggers belong to. We also consider the number of blogs making up a component. For example, are components generally created by a few blogs with a large number of links, or are components more spread out.

IV. STUDY

We now present our study on a data set of over 1.25 million blog posts, spanning multiple blog-sites. We will in this study explore each of the analytic questions put forth in the previous section.

We will first describe in some detail the data that we are using in the study and then address each of our analytic questions on the data.

A. Data

We are in this study using blogs gathered over a period of three weeks (May 23, 2009 through June 12, 2009), where we used the FetchBlogs product to monitor over 295K bloggers from a variety of blog sources. We monitored bloggers from blogspot, myspace, livejournal, multiply, wordpress and a few private blog sites, extracting all new blogs written in the timeframe we monitored the bloggers.

The size and metrics of the data set is shown in Table I. As we can see, we got basically no links from myspace and our link analysis will therefore not include that site.

B. Posting Behavior

Our first analysis is of the behaviors of the individual bloggers. We first look at when bloggers post. Figure 3 shows a plot of the ratio of bloggers that were active in our three week period that were also active on any particular day. As we can see, all sites have a very similar profile with a significant dip in activity on the three weekends. We also see an interesting profile of both myspace and multiply, where both sites have much lower general activity than the other sites despite the fact that they neither have the fewest or the most bloggers overall.

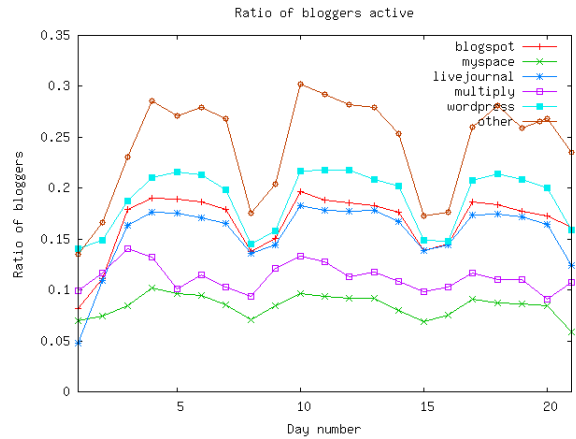


Fig. 3. When do bloggers post? We here plot ratio of all active bloggers that were active each day. We see significant dips on the weekends.

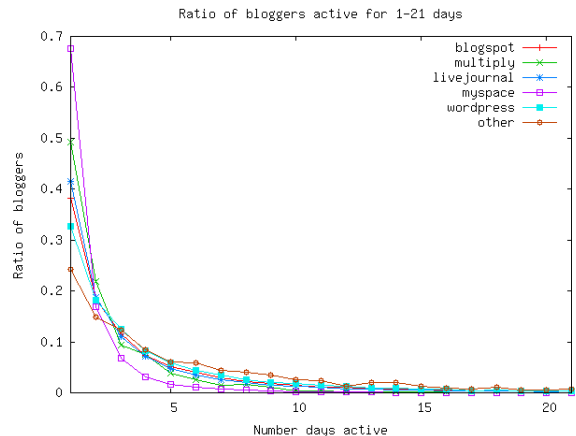


Fig. 4. What is the fraction of bloggers who post 1, ..., 21 days in the period we monitored the sources. We see a very consistent pattern that most bloggers post less than 4 days, or roughly once a week.

This suggests something about the general participation of those sites as compared to the others.

The next question we ask is what is the frequency with which bloggers post? Figure 4 shows, for each type of blog-site, the fraction of bloggers who were active 1 day, 2 days, ..., through the full period. The graph shows a consistent pattern across all sites, namely that the majority of bloggers were not that active and only posted once in the three week period we are considering here. Consistent with Figure 3, we see that both myspace and multiply had by far the most bloggers only active once.

As a follow-up to the two first analytics, we explore whether there is a difference in the size of the blog posts for each site as well as whether there may be a correlation with the number of posts a blogger has. Figure 5 shows a plot of all bloggers, plotting the number of posts a blogger has against the average size of a blog-post for that blogger. Note that the y -axis is a log-scale for clarity. We see a very consistent and interesting pattern across all sites (except for myspace), where the distribution of average blog-post sizes (in bytes) centers (roughly) around 1000 bytes, with livejournal having by far the most prolific bloggers and also a slightly lower overall size in blog-posts. We see an interesting extraction artifact with the myspace bloggers: all blog-posts were less than 203 bytes. It

metric	overall	blogspot	myspace	livejournal	multiply	wordpress	other
# blogs	1242598	255194	124674	761881	3250	87100	10499
# blogs-with-links	323379	26113	1	271438	1912	21349	2566
# bloggers	298638	51586	61311	167250	883	16174	1434
# bloggers-with-links	97386	10137	1	79811	631	6157	649
# links	1474259	81111	2	1208052	79582	94333	11179

TABLE I
METRICS FOR THE DATASET USED IN THIS STUDY.

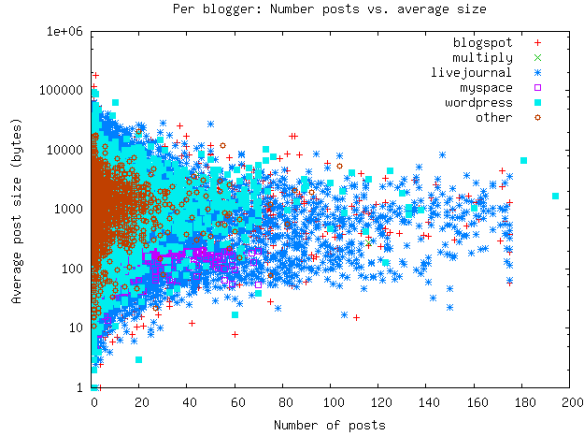


Fig. 5. This shows for each blogger, the number of posts the blogger had and the average size of the post. We see a relatively consistent pattern across all sites. Note that the y -axis is a log-scale.

turns out that the rss feed we were using only shows the first 200 bytes (followed by a “...” if the post is longer), and one has to go to the page itself to get the full post. Unfortunately, we needed not only a login but also permission to view many of the posts and hence we did not get the full blogs for myspace. We found out about this problem too late to address it in this study.

Our last study on individual bloggers is how many topics they cover. We plot in Figure 6, for each blogger, the number of posts the blogger had versus the number of topics the blogger covered across all those posts. We found that blogs on average were tagged with 1.3 topics given the algorithm outlined above in the data preparation step, and plot the line $1.3 \cdot x$ as a comparison trend. As we can see in Figure 6, the more posts a blogger has, the more topics were covered. Although the trend does not completely follow the expected line if all new posts covered completely new topics, we still see a significant upward trend, which is almost identical across all blogsites. This suggests that bloggers do tend to blog across a wide spectrum of topics and are not likely to stay focused on a few topics.

C. Topics

We wanted to explore the “behavior” of the 1000 topics we identified using LDA. Our first topic analysis looked at whether topics tended to only have activity for a few days as with many specific news items, or if these general topics found by LDA might be discussed for a longer period of time. Figure 7 shows how many topics were discussed 1 day, 2 days, all the way up to 21 days. When breaking it up by blogsite, we see the two sites with the fewest bloggers have less discussions on topics than the other sites. However, were we two consider all blogs, than all 1000 topics were discussed at least once per

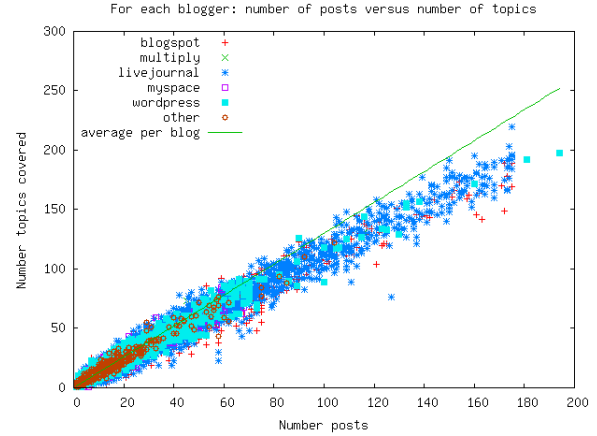


Fig. 6. How many topics each blogger covered. For each blogger, plot the number of posts the blogger had versus the number of topics the blogger covered in the posts.

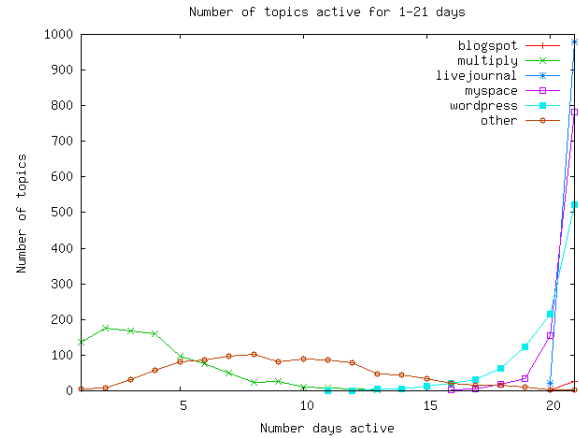


Fig. 7. How many days are topics discussed by each blog-site. As we can see, multiply and other, both of which have very few bloggers, do not have as consistent discussions as the other large sites.

day. This is interesting because it suggests that specific topics might quickly die out, but more general topics, even as large as 1000 over three weeks, all receive consistent attention.

Next, we explore where some topics are more popular than others. We have just seen that all topics are discussed consistently, but are some topics more popular than others? Figure 8 shows the popularity of topics. We see that the first 10 topics receive a large amount of posts, and then the curve rapidly changes to a linearly degrading curve, suggesting only a few very popular topics. Interestingly enough, the most popular topic is all about twitter.

Finally, we consider whether topics might have spikes where they are discussed a great deal with very low coverage the remaining time. We plot in Figure 9 the activity for all topics, where each point represent for a specific topic on a specific day the ratio of posts about the topic that appeared that day.

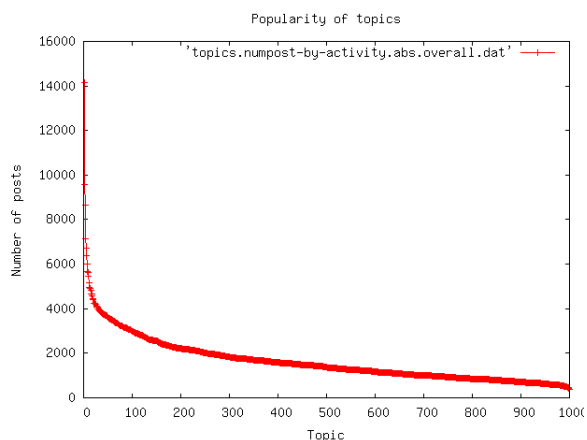


Fig. 8. Popularity of topics: plot, for each topic, the number of posts that were tagged with the topic and sort by number of posts.

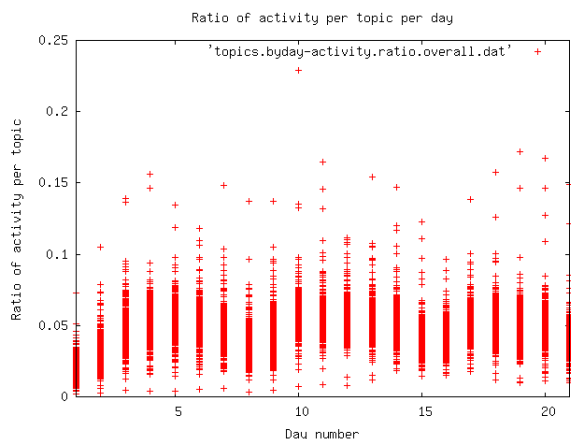


Fig. 9. When were topics discussed and how much? We plot, for each topic, for each day, the ratio of posts about that topic that occurred on that day.

We are in particular interested to see if there are any spike-points, points that show the ratio of coverage is very large. We see very few of those spikes, with one outlier on day 10, who had more than 22% of the stories on that topic appear on that day. We were curious about the topic and identified the spike to belong to topic number 570, which turned out to be about David Carradine’s death back in June 2009. This was a very specific topic about this particular incident, yet it appeared throughout the 3-week period (before his death). However, the ratio of posts before his death and towards the end of this period were very small. This suggests that the topic may have been picked up by related topics (e.g., Quentin Tarantino, accidents, martial arts, kill bill the movie, etc.). This topic turned out to be ranked 310 in popularity above, with 1795 posts.

D. Linking Behavior

We next turn to analysis of the linking behavior of bloggers. We are specifically interested in understand their linking behavior with respect to how many links, how often they link, and to whom they link.

We first take a look at the linking behavior of bloggers and blogs in general. Table II shows across all blogsites as well as for each blog-site, various likelihoods. “P(blogger-link)” shows the likelihood that a blogger has at least one blog with

metric	overall	blogspot	livejournal	multiply	wordpress	other
P(blogger-link)	41.0%	19.7%	47.7%	71.5%	38.1%	45.3%
P(blog-link)	28.9%	10.2%	35.6%	58.8%	24.5%	24.4%
P(self-link)	32.2%	50.9%	24.4%	98.7%	57.5%	58.7%
P(site-link)	63.6%	39.4%	72.5%	0.9%	31.1%	1.6%

TABLE II
WHAT IS THE LIKELIHOOD OF A BLOGGER HAVE AT LEAST ONE LINK? A BLOG HAVING AT LEAST ONE LINK? THAT A LINK IS BACK TO THE SAME BLOGGER OR THE SAME BLOG-SITE?

metric	blogspot	myspace	livejournal	multiply	wordpress	other
blogspot	90.3%	4.1%	0.6%	0.6%	3.9%	0.5%
livejournal	1.2%	1.1%	96.9%	0.1%	0.3%	0.4%
multiply	0.1%	0.0%	0.0%	99.5%	0.0%	0.4%
wordpress	8.9%	1.6%	0.3%	0.1%	88.6%	0.5%
other	13.2%	4.1%	0.3%	0.0%	22.1%	60.2%

TABLE III
WHERE ARE NON SELF-LINKS GOING? CELLS ON THE DIAGONAL (LINKS GOING BACK TO THE SAME BLOG-SITE) ARE BOLD-FACED.

a link one link in their blog (the ratio of bloggers to bloggers with at least one blog with a link). As we can see, over 40% of bloggers have at least one blog with a link and “P(blog-link)” shows that almost 29% of all blogs have links. “P(self-link)” shows the ratio of links that were a blogger linking to herself or himself. Finally, “P(site-link)” shows the likelihood that if the link was not a self-link that it was at least pointing to the same blog-site. The table shows some very interesting patterns. First, we see that although multiply has a lot of bloggers linking, they are almost exclusive links to themselves (at 98.7% of all their links!). While not generally quite as extreme as multiply, we do see a strong trend for self-links and links back to the same blog-site and with the exception of blogspot, linking seems quite ubiquitous.

We next consider where all the remaining links go, after removing the self-links. Table III shows for each blog site where the remaining links end up. As we can see, almost all links end up in general on the same blog-site, indicating that homophily [13] is a strong force when bloggers select the blog-site they will use. Of the blog-sites, only the non-major blog-sites showed some restraint, having significant number of links going to blogspot, myspace and wordpress.

Finally, we consider reciprocity. How likely is it that a link will be reciprocated? In other words, if blogger A points to blogger B, what is the likelihood that blogger B will have a link to blogger A? We computed the likelihood of reciprocity for bloggers from blogspot, livejournal and wordpress, the three sites where we had enough bloggers to compute these likelihoods with any confidence. We found that likelihood of reciprocity lay in the range of 4–6%, which is quite good and does show that reciprocity is a practice that many use.

E. Group Detection

Our last analysis considers the effect of our text-mining and tagging on how the social network is broken up into connected components. We argue that having extremely large connected components will make it very difficult to identify salient small close-knit groups or communities. Instead, it would be preferable if the social network itself is created such that there are more smaller components which can then be more easily analyzed. By tagging links by the topics

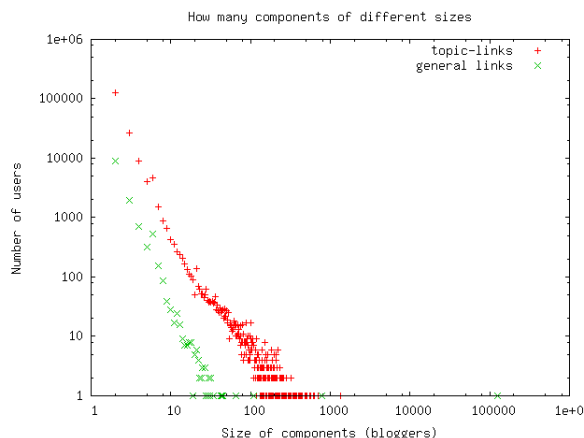


Fig. 10. How many components are found of different sizes, using either general links (general links) or an aggregate of all components found analyzing each topic graph separately (topic-links).

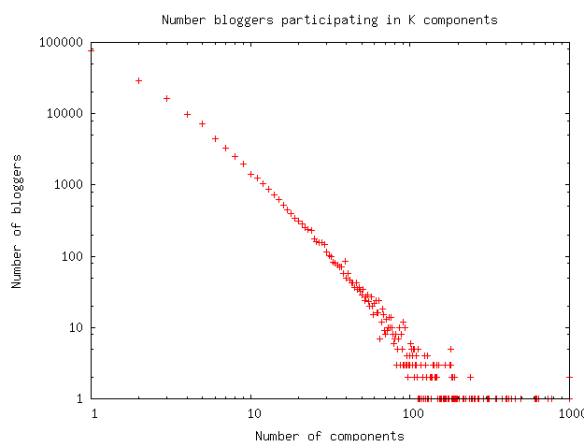


Fig. 11. How many components do bloggers participate in?

of the blog-post from which the links were extracted, we create a semantically rich network which can be used to find such smaller components. Specifically, if we consider that each of the 1000 topics found by LDA induce a different social network consisting only of links tagged with that topic, then we can for each topic find the connected components for that graph. The advantages for doing this is twofold: first, we now get more smaller components and hopefully no giant components, and second, bloggers now readily appear in multiple components, each based on a particular topic. This alleviates the problem of finding multiple smaller overlapping groups in a large homogeneous network.

Figure 10 shows the very different behaviors of using topic-tagged links or considering all links equally (not both x and y are logscale). In the latter case, we see fewer components of smaller sizes and then large and one giant component. However, in the former case where we use the topic-tagged links, we get far more smaller components, each of which are much easier to handle and analyze.

Finally, we verify that using the topic-tagged links and separately analyzing 1000 topic-graphs for their components result in bloggers participating in multiple components. This is an important aspect of this methodology because bloggers are likely to belong to multiply groups and if we were to treat all links as equal, all these groups would be joined

together in one giant component, making it hard to separate out the various overlapping groups. Figure 11 shows how many bloggers ended up in various groups (note the x -axis is logscale.) Specifically, we see how many bloggers ended up in 1, 2, ..., 1000 components (1000 is the max, one per topic-graph). We see a very nice graph, where many bloggers participate in 10's and 100's of components (with one blogger ending up in 1000 components.)

V. CONCLUSION

We proposed to enhance social network analysis on social network data sets which contain text by mining the textual content to identify topics and then enrich the underlying social network with these topics. We argue that this enrichment allows us to do finer-grained analysis of the social network and enables us to ask questions that cannot be asked of the generic social network alone.

We first described our methodology for taking such data and pre-processing it into a format that is readily mined and analyzed. Our methodology included the complete process from gathering and extracting data, to using text-mining and topic-detection to identify topics, and to finally created and tag the social network.

We then described the new analytic questions we could ask of this enriched data, which include in-depth queries on the topics as well as topic-analysis of the social network.

We showed the efficacy of our analytic methodology on a real-world dataset we collected from the blogosphere, which consisted of over 1.25 million blogs and nearly 300,000 users. Our analysis spanned behavioral analysis of bloggers and topics, to finding smaller closer-knit groups in the social network by leveraging the semantically enriched links.

REFERENCES

- [1] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: divided they blog," pp. 36–43, 2005.
- [2] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst, "Patterns of cascading behavior in large blog graphs," 2007.
- [3] A. Joshi, T. Finin, A. Java, A. Kale, and P. Kolar, "Web 2.0 Mining: Analyzing Social Media," in *Proceedings of the NSF Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation*, October 2007.
- [4] R. Nallapati and W. Cohen, "Link-plsa-lda: A new unsupervised model for topics and influence of blogs," 2008.
- [5] N. Agarwal and H. Liu, "Blogosphere: Research issues, tools, and applications," July 2008, vol. 10, no. 1, pp. 18–31.
- [6] M. Hearst and S. Dumais, "Blogging together: An examination of group blogs," 2009.
- [7] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research (JMLR)*, vol. 3, pp. 993–1022, 2003.
- [8] A. K. McCallum, "MALLET: A machine learning for language toolkit," 2002, <http://mallet.cs.umass.edu>.
- [9] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, July 1980.
- [10] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," 2008, arXiv:0810.1355v1.
- [11] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, 2004, 066111.
- [12] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.
- [13] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, vol. 27, pp. 415–444, 2001.