

Contextual Linking Behavior of Bloggers

Leveraging text-mining to enable topic-based analysis

Sofus A. Macskassy

Accepted: May 3, 2011
This edit: May 26, 2011
© Springer-Verlag 2011

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 1.13M blogs posted by 185K bloggers over a period of 3 weeks. These bloggers span private blog sites through large blog-sites such as LiveJournal and Blogger. We show that we can, in fact, tag links in meaningful ways by leveraging topic-detection over the blogs themselves. We use these topics to contextually tag links coming from a particular blog post. This enrichment enables us to create smaller topic-specific graphs which we can analyze in some depth. We show that these topic-specific graphs not only have a different topology from the general blog graph but also enable us to find central bloggers which were otherwise hard to find. We further show that a temporal anal-

ysis identifies behaviors in terms of how components form as well as how bloggers continue to link after components form. These behaviors come to light when doing an analysis on the topic-specific graphs but are hidden or not easily discernable when analyzing the general blog graph.

1 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. These online social network data are large, often contain contextual information such as text or semantically typed relations, are temporal in nature and are also extremely noisy. All of these provide an extremely rich set of data but also introduce new interesting analytic problems for analyzing these large social networks. However, as the data have 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., Adamic and Glance 2005; Leskovec et al. 2007b; Joshi et al. 2007; Nallapati and Cohen 2008; Agarwal and Liu 2008; Hearst and Dumais 2009).

There are many types of online social network data available, which span everything from social networking sites such as Facebook 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 interesting characteristics and each of these have seen their share of attention from researchers in various fields. In this paper, we will focus on

Sofus A. Macskassy
Fetch Technologies
841 Apollo Street, Suite 400
El Segundo, CA 90245, USA
Tel: +310-414-9849
E-mail: sofmac@fetch.com

the blogosphere, where the data consist of blog posts and links between them. What makes this type of data particularly interesting is its richness in that it contains text and explicit links between bloggers. Further, there are millions of people blogging (156 million blogs currently tracked by *blogpulse*¹), 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 key hypothesis of this paper is that leveraging contextual information about the links which are formed will provide new analytic capabilities and insights which were not possible to glean by analyzing the blog-network at large or through standard community detection algorithms. While we expect some behaviors to be the same (e.g., size of clusters following a powerlaw distribution), we also expect others to be quite different. For example, how bloggers link to each other, who the central bloggers are for a topic, how communities evolve, etc. The focus of this paper is to answer just how different these analytics are when considering contextual links versus the blogosphere at large and provide guidance on how to think about how to handle social media, in general.

The key contribution of this work is a study which shows how to apply text mining on blogs to extract topic-specific networks and showing that using social network analysis on these topic-specific networks improves the efficacy and understanding of the behaviors of bloggers. Our approach is to use a topic-detection text-mining method on our blog posts and then “categorize” these blog posts based on the topics they most relate to. From this, we tag links between blogs based on the topics that the categories of the blogs. This results in a large network with contextually tagged relations where we can focus on the small communities that result by only considering relations that are categorized with the same topic. We will show that linking characteristics found using this topic-specific analysis yields insights into bloggers and their linking behaviors beyond what can be done by considering the network as one large aggregate network.

While we use a standard text mining approach (LDA), the way in which we leverage text mining to extract topic-specific networks and analyzing these networks is novel and, as we will show, let us gain new insights into the data which were not possible before.

We will, in addition, repeat some studies used previously in the literature to validate their findings and show how these

findings hold up when analyzing the network with respect to a particular context.

The remainder of the paper is structured as follows: we first put our work in the context of related work. 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 finish with concluding remarks.

2 Related Work

There has been a lot of research on blogs and social media in the past decade. In fact, the International Conference on Weblogs and Social Media (ICWSM)² started back in 2007 and is dedicated to this particular topic.

Of most relevance to this paper is perhaps the rich literature on the general explorative mining of the blogosphere (see, e.g., Joshi et al. 2007; Leskovec et al. 2007b; Agarwal and Liu 2008; Hearst and Dumais 2009). Much has been said about the underlying topology (Shi et al. 2007), demographics (Kumar et al. 2003), structure (Kumar et al. 2006) and evolution (Kumar et al. 2003; Backstrom et al. 2006; Kumar et al. 2006; Leskovec et al. 2007a; Götz et al. 2009) of the blogosphere. Where applicable, I mentioned how our findings were in agreement with these earlier work.

Analyzing the temporal nature of the social network is clearly important. In addition to the broad work on evolution above, one can also focus on the individual level to understand how individuals form and break bonds to communities (see, e.g., Sharara et al. 2010). Improving capabilities in that respect would greatly help understanding the dynamic nature of topics and how the central players move in and out of the spotlight.

More recently, researchers have turned towards better understanding of how information flows through the blogosphere. The methodology used is generally one of understanding information cascades (Leskovec et al. 2007b; Papagelis et al. 2009; Ghosh and Lerman 2011). In other words, what are the specific patterns of diffusion for specific pieces of information. These are generally relatively small patterns, but still informative in order to understand at the micro-level how information might be passed along.

Text mining has been applied to blogs for a variety of reasons. Of interest to generating topic-specific graphs is recent work on enhancing the network by learning user pro-

¹ <http://www.blogpulse.com>, as of February 2011. This is up from 4 million in 2004, to 19.6 million in 2005, 70 million in 2007, and 133 million in 2008 (see <http://www.readwriteweb.com/archives/state-of-the-blogosphere-2008.php>).

² <http://www.icwsm.org/>

files through text classification (Ni et al. 2006). Similarly, others have looked at using user-supplied information (e.g., self-reported interests in the form of keywords) to better understand what makes users link to each other (Bhattacharyya et al. 2010). Combining this type of information with the semantic link information should make a powerful model. Text mining has also been used to identify topics and influence of blog posts (Nallapati and Cohen 2008).

The application of social network analysis (SNA) to blogs and social media is also rich. Recently, has been applied to US political blogs (e.g., Adamic and Glance 2005; Lazer et al. 2010; Hanneman and Shelton 2011; Rosen et al. 2011), where there is a strong tendency for conservative bloggers to link to conservative material, whereas more liberal bloggers spread their links out.³ Clearly, a relevant observation when trying to understand the context and semantics of links.

3 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. These are often not described and we feel that it is important to cover this aspect because data acquisition and cleaning directly impacts the quality of the data analysis. As such, they are important steps in any analysis. In addition, data preparation must be described such that the full process is understood and reproducible.

In the real world, data are not often clean nor in a proper format for analysis. In our particular setting, we are dealing with blog data which consist 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 these 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.

³ In these data, conservative bloggers covers those whose political views are aligned with the Republican party and liberal bloggers are those more aligned with the Democratic party.

3.1 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 because it allows us to ensure timeliness and quality of the extracted data. In addition, we have ready access to a commercial system, FetchBlogs, which already monitors millions of blogs on a basis and extracts new posts as they become available. We here use the FetchBlogs system from Fetch Technologies⁴ to get and archive posts from millions of bloggers for the study in this paper (we describe the particulars of this data below in Section 5.1.) FetchBlogs is based on Fetch Technologies' commercial Agent Platform for large scale information extraction from the web. FetchBlogs works by analyzing a blog-site and automatically recognize where the blog posts are by identifying markers in the HTML or XML structure of the blog page. This is done using machine learning and AI methods to automatically find such markers from looking at a few example pages. As such, the FetchBlogs system is a focused web-extraction agent which only goes to the pages needed to be extracted from. The system is based on many years of research which culminated in the commercial product offered by Fetch Technologies.

The FetchBlogs product is set up to monitor millions of online blogs and extract the clean blog posts together with links and other information. The details of how this tool works is not relevant to the remainder of the paper and we will therefore skip those details, except to highlight some of the issues one needs to pay attention to when extracting blogs (or any other web-content, for that matter).

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.

In addition, particular care must be taken to only extract links within the blog-post itself and not other links to the other bloggers that are not part of the post itself. The FetchBlogs system does this automatically for us as part of its day-to-day operation.

The blogs as extracted as described above contain markups such as bold-faced text, links to multimedia, etc. Many blogs also often contain a comment section where people can comment on the blog content.

⁴ <http://www.fetch.com>

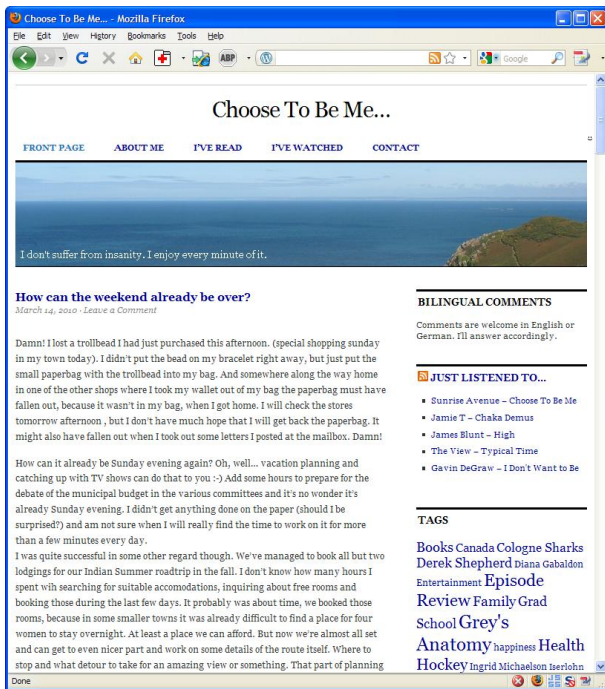


Fig. 1 Example blog from wordpress.com.

While comments and HTML markup (bold-faced text, images, etc.) 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. We extract the raw text from the blogs and keep track of the links within the blog content itself.

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 username as part of the URL.⁵ We represent each such outgoing hyperlink as a directed edge in a large social network.

The working data we are left with are 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 username, the date the post was published online and the length of the post.

3.2 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. In particular, we are interested in meta-level behaviors such as what topics are being posted as well as the

⁵ <http://username.livejournal.com>

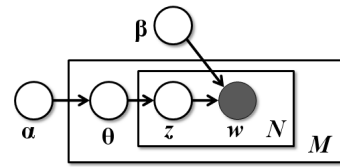


Fig. 2 Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer box represents the M blogs, and the inner box represents the N words for each blog.

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 use an unsupervised topic-detection and clustering algorithm known as *Latent Dirichlet Allocation* (LDA) (Blei et al. 2003), which models a text corpora as a set of topics. We note that the particulars of the topic-detection algorithm is not as important as the high-level concept of how to leverage text-mining to *enrich* the social network and mine this enriched network. In other words, the results of the study in this paper is agnostic to the topic-mining tool. However, understanding how LDA works is important to understand the complete study and we therefore here describe how LDA works at a high level. LDA is a generative probabilistic model of a corpus, represented as a graphical model in Figure 2. The generative process is as follows:

1. For each blog, m , choose a topic distribution θ ($\theta \sim \text{Dir}(\alpha)$).
2. For each word w in the blog, first use θ to choose a topic z
3. Using a β distribution of words, conditioned on z , choose a word.
4. 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 w s in the corpus.

We refer the reader to Blei et al. (2003) for more details.⁶

Given T , a desired number of topics, LDA will find the best parameters to fit the corpora into T topics. One large problem of LDA is settling on the correct number of topics that are truly in the corpora. 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 1,000 to be a good granularity.⁷

⁶ We 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 (McCallum 2002), 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.

⁷ We tested with smaller number of topics down to 100 and found qualitatively similar results as those reported here.

LDA generates a language model over all observed words. However, our blog data set contains over 2.5 million unique words, which makes it quite a large LDA modeling problem. We therefore first reduce the dimensionality of the problem by applying stemming (Porter 1980) 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. This pruning may cause LDA to miss spikes or potential topics, but our study is more focused on the concept and efficacy of being able to extract topics and showing that the contextual analysis used in this paper is indeed worth-while. Further, the 1,000 topics we found are generally broad and discussed by 1,000’s of bloggers so our pruning should not have had an adverse effect on the topics found.

Once we have fit a model (found the optimal values of our parameters), we run an inference step on the same data to get a probability score of how likely it is that each blog post was “generated” by each of the 1,000 topics—i.e., each blog belongs to each topic with some likelihood (many of which are near-zero). Again, we highlight that each of the 1,000 topics is automatically found using LDA and can be represented as the set of words most likely to occur for that topic. We tag a blog with the five most likely topics that it covers, keeping track of the likelihoods because we use them later for additional pruning.

The end result of this step is then that each blog is now tagged with five topics and the likelihood that the blog covers that topic. We use these likelihoods to tag the links.

3.3 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 (Leskovec et al. 2008b). Our approach to handling the problem of large connected graphs is to tag links based on the topics of the blog that is the source of that link.⁸

We note that some blogs are very topic-specific and have only one topic with a high likelihood, whereas other blogs

⁸ We realize that we could also look at the content of the destination blog. However, many of those blogs were not part of the data set used in this study and so we were unable to do that particular analysis.

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 above) down to only the topics that a blog is most closely associated with. We do this for each blog b as follows:

1. $\mathbf{t}_b = \{\}$
2. $\mathbf{t}'_b = \{t_1, \dots, t_5\}$ 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. $i = 1$ (looping variable to loop through t_i)
4. $\mathbf{t}_b = \mathbf{t}_b \cup \{t_i\}$
5. $i \leftarrow i + 1$
6. if $i \leq 5$ and $p(t_i) > 0.9 \times p(t_{i-1})$, repeat from step 4.

The result is that \mathbf{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-username of b and the destination node is the blog-username of the blog-post being pointed to. The tags for this link are \mathbf{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.

4 Analysis

The data set which we will analyze has two forms: the blogs and their meta-information (such as blog username, topics, date, length of post), and the (dynamic) 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.

Our analytic study is broken down into two parts:

1. **General statistics:** We will compute some general statistics on the network to get baseline characteristics such as how active bloggers are, how many topics they tend to write about, etc. We will also include a brief analysis of the topics themselves to provide the reader with some insight into what topics look like and their activity behaviors (spikes, longevity, etc).
2. **Network analysis:** This is the primary part of the study in this paper: how do the topic-specific networks differ from the general network? We ask three questions: do the topic-specific graphs have a different clustering than the general graph? We would expect to get smaller and more manageable components. Second, can we find people in the topic-specific graph which are otherwise hard to find? And third, can we get new insights into temporal behaviors which might be hidden or difficult to tease out from the general graph?

4.1 General statistics

The first part of our study focuses on high-level blogosphere statistics. This study is primarily meant to validate and confirm high-level behaviors which have either been noted or hypothesized in previous studies.

Because the data we collect contain bloggers from a variety of large blog-sites such as LiveJournal and WordPress, this analysis will be both at the micro-level of individual bloggers as well as at the higher-level across blog-sites to see if there are fundamental differences in behaviors of bloggers across sites.

This part of the study contains two parts: posting behavior and topic analysis.

4.1.1 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.

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. Based on public statistics, we expect to see dips on the weekend, where fewer bloggers are active.⁹
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? Based on a plethora of past studies of user behavior, we expect to see that most bloggers are infrequent while a few bloggers post every day (see, e.g., Adamic and Glance 2005; Agarwal and Liu 2008; Hearst and Dumais 2009). 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 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.

4.1.2 Topics

The second part of the general statistics focuses on the “behavior” of topics based on our text-mining results. Treating topics as first-order objects, we explore 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 1,000 general topics all equally popular? Based on common intuition about what people tend to blog about (e.g., current events and general topics), we would expect there to be some topics which are more popular than others amongst the 1,000 topics found. 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 1,000 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.

4.2 Network Analysis

The key part of our study is an analysis of how the social networks generated by the blogosphere at large and those generated by considering only specific topics differ. Specifically, the hypothesis is that the topic-specific networks will induce smaller and more manageable components. If so, then the question becomes whether these smaller components actually provide new insights or enables us to identify behaviors which were not apparent in the larger general network.

⁹ For example, see <http://www.blogpulse.com>

4.2.1 Topological Analysis

Our first question is whether the topic-specific networks are different from the general networks. Specifically, we expect the topic-specific networks to have a larger number small-to-medium sized components than the general network, where we expect we will get one or two very large connected components and a larger number of very smaller components (Leskovec et al. 2008b).

One way to get smaller components would be to use any of the community detection algorithms (e.g., Clauset et al. 2004). The problem with running community and group detection algorithms on large graphs is that they tend to generate a few large clusters (see, e.g., (Clauset et al. 2004)). 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., Newman 2003). 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.

We hypothesize that focusing on contextual networks will lead to a larger number of small-to-medium size components which are easier to analyze in depth, and which are furthermore relevant to a topic which is presumably of interest. We expect these smaller components exactly because we expect that topic-based links are close in the network and only cover the people actually talking about the topic. If we focused on a topic which everyone was likely to actively talk about, then perhaps this would not be the case. However, based on empirical observation of the data and people, in general, we do not believe that there are that many topics which everyone would actively participate in.

Just as importantly, bloggers can now belong to multiple communities (still one per topic, but there are now N topics), something not easily achieved if we do not distinguish between relations or use community detection algorithms. This suggests that the topological metrics are likely to be significantly different.

This part of the analysis is focused on showing that topic-specific networks do, in fact, generate more smaller graphs and that they are different from just randomly sub-sampling the large graph and that they are equally different from using community detection. If we can show this, then we are on the way to show that generating topic-specific networks are important for network analytics.

4.2.2 Finding Key Bloggers

Assuming that we can show that topic-specific networks are fundamentally different, then we can go on to show the first

way in which we can use these networks in ways we could not do on the larger graph.

We will first focus on the problem of identifying important bloggers relevant to a topic. If we can extract out a topic-specific network, then we can identify the important people through various means such as ranking the bloggers by centrality betweenness. The question we ask here is whether we actually *need* a topic-specific network, or whether we can identify the same people from the larger graph. The hypothesis is that the graph structure is fundamentally different to such a point that it is hopeless to identify the same central people from the larger graph, showing that being able to extract out the topic-specific network is important.

4.2.3 Temporal Analysis of Communities

The final part of the network analysis study is an analysis of how the components evolve over the period of observation. We are, in particular, interested in two question: how do the components form and what are the linking behaviors of bloggers after components have merge?

For the former evolution question, we follow prior work and explore how components form. For example, as “new” bloggers merge into a growing component, do they come in a single bloggers or do they form sub-components first? Further, as components merge, is it the smaller components taking the initiative to link to the larger components or vice versa? Prior work in this area have not looked at the directionality of these merges (see, e.g., Kumar et al. 2006).

For the second analysis, we want to know what happens after components merge. Do bloggers start linking more with each other or do they tend to stay closer to their initial component? This will tell us about the cohesiveness of these clusters as they form beyond the standard “closing of the triangle” which is a general behavior of social networks (see, e.g., Wasserman and Faust 1994; Leskovec et al. 2008a; Lazer et al. 2010), but perhaps less so in the blogosphere when we are analyzing topics. In this case it may make more sense to talk about components or groups and monitor how far communication goes outside the initial groups as they start merging into larger communities.

5 Study

We now present our study on a data set of over 1.13 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.

5.1 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 3+ million bloggers and archive any new posts found in that timeframe. This was done as part of an internal blogs study to explore the efficacy of topic-contextual social network mining. This paper is the result of some of the studies we performed on this data. We used this data set because it was timely, it was of good quality and it represented the kind of data that we already collected and we could be sure that any results would carry forward without having to change the collection paradigm.

The bloggers selected for monitoring were a random sample of bloggers already being monitored commercially and reflected the kind of data we would have available on an ongoing data. The bloggers were predominantly from the larger blogsites such as LiveJournal, WordPress, Multiply, Blogger, etc. In the 3-week period 295K of the bloggers had new posts (about 10% of the bloggers being monitored). Of these, 185K bloggers had at least one post with a link to another blogger. We focused on these bloggers in this study because we are predominantly interested in exploring the social networks.

The size and link-metrics of the data set, broken up by the major blog-sites, is shown in Table 1. The “other” category groups together individual blogger sites.¹⁰ We include here statistics on the likelihood that a link from a blogger at a given blog-site links to within the same blog-site (including a link back to the blogger’s own blog). As we can see bloggers tend very strongly to stay within their own blog-site, which agrees with prior studies (see, e.g., Shi et al. 2007).

5.2 Part I: General Statistics

Our first analysis of this data is at the blogosphere level, where we report on general posting behaviors of bloggers across the different blog-sites and on the statistics of the topics found using text-mining. This analysis is done first to understand whether we should consider blog-sites separately and secondly to lay the foundation for the social network mining in the second part of our study.

5.2.1 Posting Behavior

We start by looking at the posting behaviors of the individual bloggers: when are bloggers more active (larger volume of new posts)? Figure 3 shows a plot of the ratio of bloggers that were active in our 3-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 weekends.

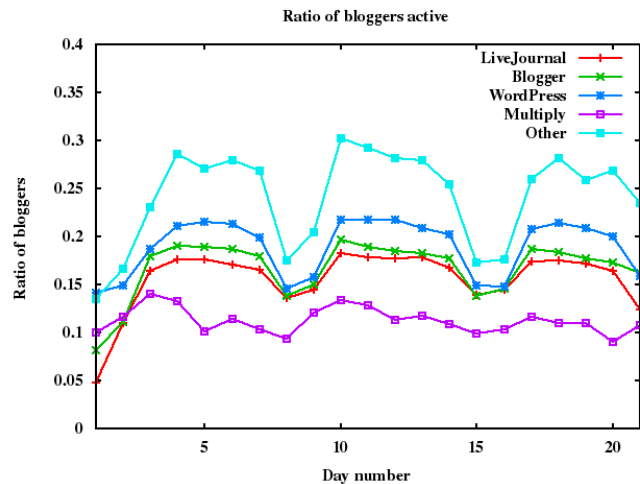


Fig. 3 When do bloggers post? We here plot ratio of all active bloggers that were active each day. We see significant dips on the weekends.

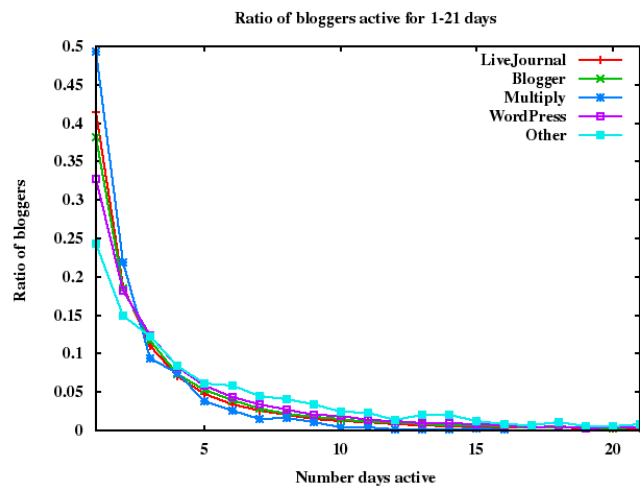


Fig. 4 What is the fraction of bloggers who post 1, . . . , 21 days in the period we monitored the sources. We see a very consistent pattern that most bloggers post less than 4 days, or roughly once a week.

We also see an interesting profile of Multiply, which has a much lower general activity than the other sites despite the fact that it neither has the fewest or the most bloggers overall. This suggests something about the general participation of Multiply as compared to the others.

The next characteristic we look at is the frequency with which bloggers post. Figure 4 shows, for each blog-site, the fraction of bloggers who were active 1 day, 2 days, etc., 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 Multiply had more bloggers active only once than the other sites.

Our last study on the blogosphere is how many topics each blogger covers. We plot in Figure 5, for each blog-

¹⁰ Such as www.thecrazywoman.com

Blog-site	Number of users	Number of links	Number of insite-links	P(insite-link)
LiveJournal	145,740	940,454	911,844	0.970
Blogger	27,826	62,747	56,749	0.904
Wordpress	10,220	60,442	51,975	0.860
Multiply	1,165	67,079	66,775	0.995
Other	471	6,629	4,214	0.636
Google	242	406	41	0.101
Vox	205	1,127	1,089	0.966
Blog-City	12	10	3	0.300
Overall	185,881	1,138,894	1,092,690	0.959

Table 1 High-level characteristics of the dataset used in this study.

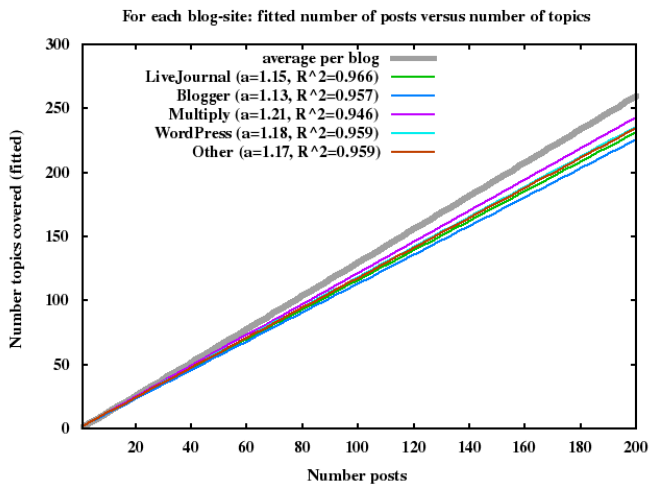


Fig. 5 How many topics each blogger covered. For each blog-site, plot the average number of topics covered by a blogger on that site as the number of posts increase.

site, the average number of posts bloggers had versus the number of topics the bloggers 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 5, 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. We will return to a deeper investigation of topics below in Section 5.2.2.

We also performed a link analysis: degree distributions, the number of links to one-self, the number of links within the same site, reciprocity, etc. Our findings were very similar to that found in (Shi et al. 2007) and are consistent across the different blog-sites. We therefore do not report on them here as they are not pertinent to our further study.

Our first analytic study has shown that the general behavior across the blog-sites are similar enough that we can treat them equally within our study.

5.2.2 Topics

We next turn to the analysis of the 1,000 topics identified using LDA.

We first verified that the topics found by LDA were salient—i.e., that they reflect proper topics or themes. We found that most topics did, in fact, have very strong themes. In particular, we found that only about 40 of the 1,000 topics were completely non-sensical (random words with no coherence) and we found that about 56 topics were foreign language topics (Spanish, German, French, Swedish, Dutch, Korean and a few others). This is not because we targeted specific foreign bloggers or blog-sites, but rather that they were part of our sample of the larger blogosphere which we monitored. In addition, we found that most topics could actually be categorized into a few high-level concepts such as religion, political, entertainment, food, music, technical, weather, travel, book/movie reviews, gaming, movies, finance, depressive, pets, adult, and nature, just to name the categories which seemed the most prevalent.

Tables 2 and 3 show a sample of the words of a few select topics, from the most and least popular (Table 2 and from a few examples of the high-level concepts (Table 3). A larger set of words are shown in Appendix A. As we can see from these tables, the topics (as defined by the words belonging to that topic) often “make sense”, although perhaps the topics are not always very clear as is often the case with LDA-type methods. However, most topics did have a “theme” and as we mentioned above only a few topics were completely non-sensical (one topic, for example, consisted almost entirely of pronoun-like words).

Having verified that LDA did identify salient topics, our first topic analysis looks 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 6 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 that Multiply, a site with the few bloggers have less discussions on topics than the other sites. However, were we to consider all blogs, then all 1,000 topics were discussed at least once per

- (14145) loudtwitter tweet twitter daily tinyurl ...
- (9574) rate author summary pair character word disclaimer ...
- (8643) twitter twitpick loudtwitter helpiranelection iranelect ...
- (7147) haha hahaha damn super dinner lunch eat fun wanna ...
- (6708) icon credit preview hotlink graphic screencape ...
- (6388) video youtube watch clip feature vid vimeo embed fanvid ...
- (5663) album track record song frelease music band sound listen ...
-
-
-
- (493) path world story decade follow wizard harry dark lord ...
- (484) third five fourth six fifth seven sixth eight seventh ...
- (481) level low lower reach factor upper normal mean expect ...
- (474) world chaos war death gather left rebuild wizard eatery ...
- (470) set aside dani complete stacy follow rest scene final ...
- (434) stock gain money profit market company report invest ...
- (389) material length bust item price measure cotton dress...

Table 2 A few example words from the top and bottom (ranked) topics. The number at the beginning of each topic represent the number of blogs tagged with that particular topic. As we can see the most popular topic was tagged in 14145 blogs and the least popular topic was tagged in 389 blogs. A larger sample of words are shown in a table in Appendix A.

(music)	song sing lyric listen music sang ...
(food)	cheese salad pasta tomato sauce dinner chicken ...
(political)	iran iranian elect ahmadinejad mousavi tehran ...
(political)	obama president administration bush barack ...
(book review)	book novel fiction story fantasy romance ...
(gaming)	game wii nintendo mario super video play plai ...
(financial)	money pay save job earn spend paid income ...
(religion)	jesus god christ bible christian gospel ...
(pets)	cat kitten kitty pet vet litter home meow dog ...
(nature)	mountain trail hike climb lake rock valley ...

Table 3 Example words for topics which we categorized into conceptual topics. A larger set of words for these topics can be found in Appendix A.

day. This is interesting because it suggests that specific topics might quickly die out, but more general topics such as those found here all receive consistent attention.

Next, we explore whether 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 7 shows the popularity of topics. We see that the first ten 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 (as shown in Table 2).

Finally, we consider whether topics might have spikes where they are discussed a great deal with lower coverage the remaining time. We plot in Figure 8 the activity for all topics, where each point represents, for a specific topic on a specific day, the ratio of posts about the topic that appeared 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 out-

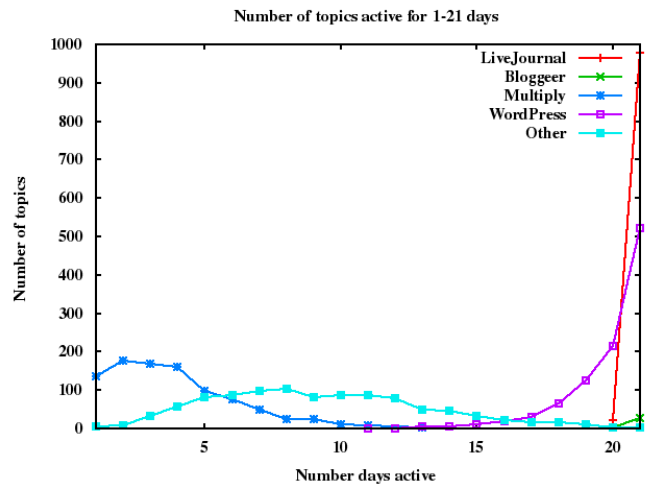


Fig. 6 How many days are topics discussed by each blog-site. As we can see, multiply and other, both of which have very few bloggers, do not have as consistent discussions as the other large sites.

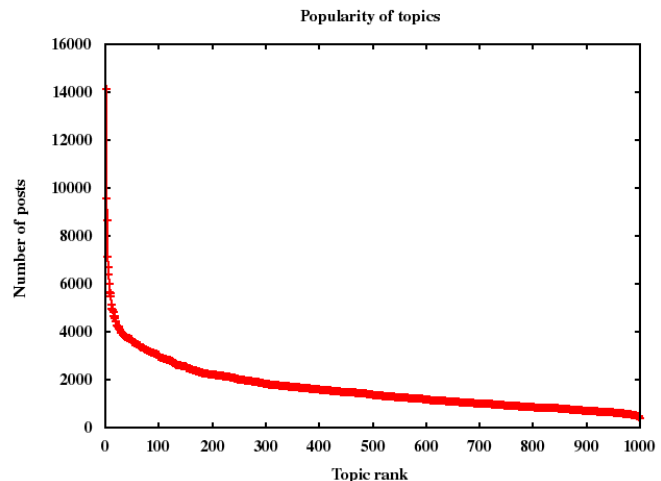


Fig. 7 Popularity of topics: plot, for each topic, the number of posts that were tagged with the topic and sort by number of posts.

lier 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 was 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 was ranked 310 in popularity with 1,795 posts.

Although clearly there are some spikes in the topics as shown in Figure 8, there are interestingly not that many suggesting that for large corpora, there are general enough topics which consistently have chatter. This effect may very

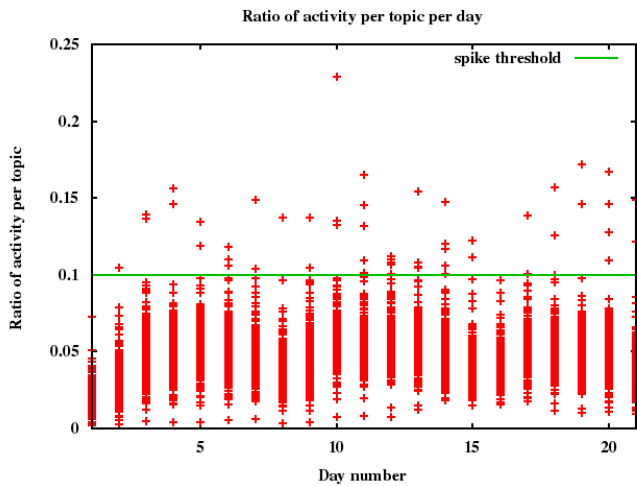


Fig. 8 When were topics discussed and how much? We plot, for each topic, for each day, the ratio of posts about that topic that occurred on that day.

well be partly due to LDA, which has a particular objective function to optimize finding broad topics. In this case, LDA would not be useful for finding short-lived and fine-grained spikes in the data, but rather for finding the broad topics we have shown above.

However, we can still explore the spikes that are present in the data and the question we asked was whether these spikes are due to a few topics with multiple spikes or whether most of the outlier points were due to topics which had high spikes for one or possibly two days. We focused in on all topics which had at least one “spike”, where we define a spike as having more than 10% of the blog-posts for that topic appear on a particular day. There are 51 such spike points on our data, belonging to 21 topics. In addition, we found that the majority of spikes were 1 or 2 days, but that one topic (the Iranian Election) had a full 5-day spike. We show the details of these spikes in Table 4.

We have now shown that the topics found are indeed salient and have longevity. We can therefore analyze our blog data with respect to those topics and contrast different methods to do this analysis to show that leveraging the contextually tagged links provides more precise analytics than not leveraging them.

5.3 Part II: Network Analysis

We now turn towards contextual network analysis. Specifically, we have leveraged text-mining to tag links in the blogosphere and can now break up the social network into topic-specific components. We argue that having very large homogeneous networks (one link type and one node type) makes it very difficult to identify, and analyze the behavior of, small focused groups and communities. Instead, we ar-

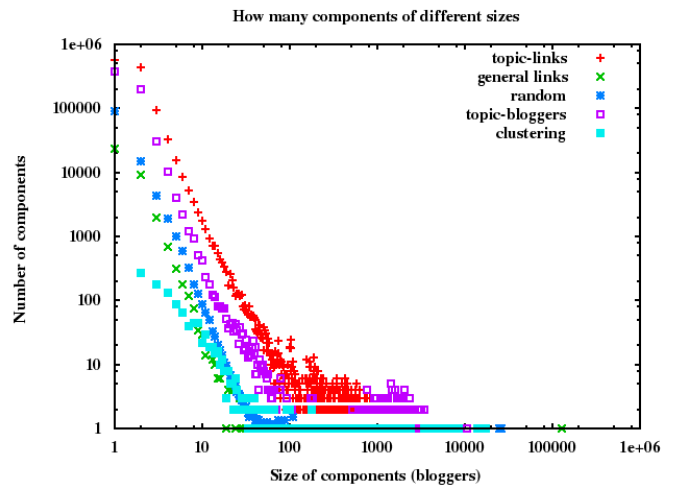


Fig. 9 How many components are found of different sizes, using either general links (general links) or an aggregate of all components found analyzing each topic graph separately (topic-links).

gue, it is preferable to create the social network such that there naturally falls out smaller focused components which can then be more easily analyzed. We here focus on using the tagged links to help extract topic-focused components, showing that these enable us to get insights and analytic outcomes not possible by considering the network as a whole.

We have discussed how we tagged links by the topics of the blog-post from which the links were extracted. The result is a semantically rich network which can be used for more fine-grained analysis. Specifically, we use each of the 1,000 topics found by LDA to induce a different social network consisting only of links tagged with that topic. We can then for each topic find the connected components for the network which is relevant to that topic. There are a couple of advantages for doing this: first, we now get different and smaller components, second, bloggers now readily appear in multiple components, each based on a particular topic, and finally, because we have these topic-specific components, we can more readily identify who the key bloggers are and how that topic evolved.

5.3.1 Finding Smaller Components

We first verify that we do indeed get smaller components. We computed the distribution of component-sizes for the components generated when considering each topic having its own graph consisting only of links tagged with that topic. We compared this distribution to that of the components found when considering all the links as one single graph.

We next wanted to test whether our topic-graphs in fact extracted out topologies different from what would be gotten from just randomly selecting a subset of relations from bloggers. To test for this, we created 100 networks, where for each we randomly selected 5 links from each blogger

Length (days)	Topic no.	Selected topic words
1	365	daniel susan talent teddy boyle britain tracey sing perform simon entertain
	390	eaten school drink kiss ridden fire play drunk snake bottle streak accident prank
	580	record archive detail label free artist myspace improvise rock tracklist gretchen
	669	wish luck rabbit bunny month hope true happy grant lucky forever genie friend
	748	protest police rally riot violenc force arrest support govern student demand activist
2	11	series opera character donald story origin soap duck phantom cartoon final classic
	164	twitter inform help iran oprah reliable blog protest spread internet attack zone
	232	muslim america world peace country islam nation women live palestinian
	412	soldier war serve veteran memory honor service marine american sacrifice hero
	413	anti holocaust jewish museum von white attack terrorist extremist violence
	840	pandora archive electron album avant band barrier gard dadala
	846	favourite pleasure guilty single friend accent tattoo band peace philosophy
3	248	obama speech president muslim world american cairo islam unite east egypt
	272	brow cynic violent tradition experiment peace fantasy attitude culture pullman
	330	election party vote candid european seat labour voter parliament result
	566	father son mother dad family own happy brother child daughter live children
	793	abort tiller pro murder george women kill doctor clinic anti term late
	809	judge court sotomayor supreme justice sonia law nominate white obama
4	493	california marriage court prop gay equal supreme proposition decision legal
	570	suicide david carradine found commit dead death kill kung actor bill rip hotel
5	24	iran iranian elect ahmadinejad mousavi tehran support revolution

Table 4 Description of all topic spikes found in the data, where a spike is denoted by one or more days having more than 10% of the blog-posts on a topic appearing on that day. Selected words for each topic are shown. The first column shows the number of running days for a given topic-spike.

and computed the resulting distribution of components. We then averaged the number of components at each size over those 100 networks to get an ‘average’ expected network.

We also wanted to control for whether we in fact *needed* to extract topic-links. Perhaps we could get the same effect by just extracting the bloggers who post on a topic and induce the network from any links between those bloggers. We note that such a network would consist only of bloggers who post on a topic and would not include other bloggers who were linked to in the context of a topic-discussion. We found that our topic-graph based on tagged links have far fewer links (there was a tenfold difference across all topics), and also includes contextually relevant bloggers regardless of whether they blog on the topic.

Finally, it may also be that the topic-graph perhaps gets at underlying community structure from a different angle. To test for this, we took the largest general component and ran a modularity-based community detection algorithm (Clauset et al. 2004) on it to get the ‘communities’ based on link-structure alone.¹¹

We contrasted each of these five networks to validate that the topic-graphs indeed extracted a different topology. Figure 9 shows the very different behaviors of using topic-tagged links (topic-links) vs. considering all links equally (general links) vs randomly picking 5 links (random) vs. bloggers actively posting on topic (topic-bloggers) vs. the modularity-based clustering algorithm (clustering). Note that in the figure both x and y are logscale.

¹¹ We here use our own implementation, available in our public network analytic tool called NetKit-SRL: <http://netkit-srl.sourceforge.net>

We clearly see that our topic-graphs clearly has many more smaller components than any of the other methodologies. Also, we see that randomly sub-selecting edges had very little effect on the distribution, showing an interesting global effect. For the topic-blogger graph, we see that we did get significant lift in the number of smaller components, but not to the same extent as the topic-links graph. Also, this graph was ‘polluted’ with links which were not on topic. We note that both the topic-links and topic-blogger graphs had the extra advantage that bloggers clearly could participate in more than one component due to the nature of splitting the graph based on topics and due to the fact that each post consisted of 1.3 topics on average. However, this multiplicative is not enough to explain away the significant difference.

It is clear from the figure that the topic-graph gets at a very different topology than any of the other methods. We also see that the distribution of resulting clusters from the modularity-clustering method looks completely different than any of the other methods, having significantly fewer small components and many more larger components than we get from our topic-graphs. In addition, the ‘communities’ found using modularity clustering are very ‘dirty’ with respect to the underlying topics. Specifically, we find that, on average, the ‘majority topic’ in a cluster only covers 15% of the bloggers in the cluster (this goes up to near 70% for the small clusters of size 8-20, but drops to less than 30% for clusters of size 50-150, and continues dropping from there).

This result clearly shows that the topology of the topic-graphs is much more finer-grained than any of the other methods.

5.3.2 Identifying Key People

Having established that the topology is quite different, we can now turn to the question of identifying “important” people for a given topic. Our definition of important is a person with high betweenness centrality. Betweenness centrality is a metric which computes, for a node in a network, how often that node is on a critical (shortest) path between other nodes in the network. We compute betweenness centrality as follows:

$$C_b(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}, \quad (1)$$

where $\sigma(s,t)$ is the number of shortest paths between nodes s and t in the graph ((s,t) -paths), and $\sigma(s,t|v)$ is the number of (s,t) -paths that go through node v .¹² By convention, if $s = t$, $\sigma(s,t) = 1$ and if $v \in \{s,t\}$, then $\sigma(s,t|v) = 0$. By convention, we let $\frac{0}{0} = 0$.

Clearly the topologies are very different between the large graph and the topic-graph, but we here want to know just how different are they with respect to being able to identify key people.

We frame the general problem statement as the following: we know that for any single topic there are multiple components. Let us focus on the problem of finding the most central person in the largest component in the topic-graph. Let us further assume that we somehow can identify all bloggers who blog on a topic. It is not unreasonable to assume that we can generate a “good enough” *ad hoc* query to identify bloggers who blog on a given topic. Given this, perhaps we do not need to extract the topic-graphs as outlined below. Perhaps, we can readily identify the “important” people from the large graph either by considering their volume of posts on the topic, their topic-specific indegree or their centrality.

We answer this question by computing the betweenness centrality of each blogger actively involved in a topic (more than one incoming or outgoing link), both from the topic-graph perspective as well as from the global graph perspective. We then rank the user by their centrality, their volume of posts on the topic, and the number of indegree links.

We first want to gauge whether we can use centrality right away. Clearly their centrality scores will be very different, but will their relative rank? We computed the Spearman’s rank correlation coefficient between the general graph and each of the topic-graphs to get their agreement. The Spearman’s rank correlation coefficient is value between -1

¹² We note that we need to compute shortest-paths for all pairs of nodes to get a centrality score for each node. There are efficient ways for computing this (Brandes 2008). In fact, the all-pairs shortest-path computation can be done in $O(nE)$ time, where E is the number of edges and n is the number of nodes in the graph. We will not include the algorithm here, but refer the reader to the original document (Brandes 2008) for details.

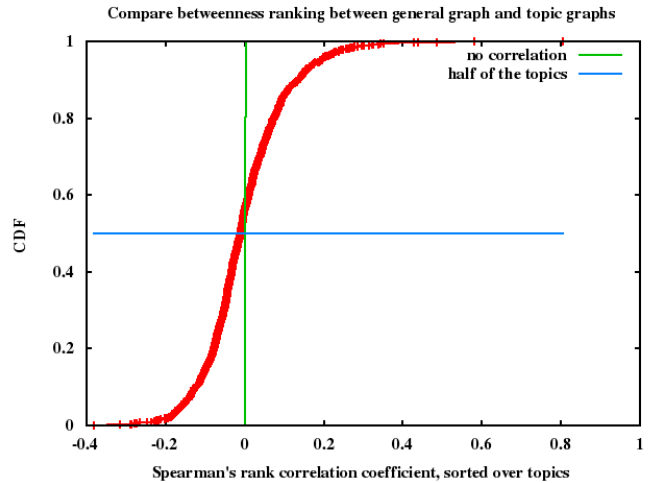


Fig. 10 Comparing betweenness ranks from the general graph to the rankings of each of the topic-graphs. We sort by the coefficient score (-1 to 1) and plot the CDF over the 1,000 topics. We can see a large variance where more than half of the rankings are in fact anti-correlated (down to -0.39) and the best correlation is at 0.8 .

top- K	Found 0	Found 1	Found 2	Found 3	Found 4
5	0.638	0.303	0.055	0.004	–
10	0.411	0.391	0.148	0.047	0.003

Table 5 How many bloggers could we find in the top- K rank in the topic-graph by looking among the top- K in the global rank among the bloggers posting on that topic. The two rows are for searching in the top-5 and top-10. The columns are how often 1, 2, 3, or 4 people were found in the top- K in the general graph who were also in the top- K in the topic graph (when ranking using betweenness centrality).

and 1 , which shows how well correlated two rankings are. A value of -1 means they are perfect reversals and a value of 1 means they are identical. The coefficient is computed as follows:

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}, \quad (2)$$

where d_i is the difference in rank of node i^{th} in the two rankings and n is the size of the list. The correlation value ranged from -0.39 to 0.80 . We show in Figure 10 a plot of all the coefficient values, sorted by value on the x -axis and y being the CDF. The horizontal line shows the point where half the rankings have been accounted for and the vertical line shows the point where there is no correlation (the left of the line is anti-correlation and the right of the line is positive correlation). As is clear from the figure, the vast majority of topics had little correlation and in fact more than half were anti-correlated! Clearly, this is not a good strategy.

To make the problem more concrete, we ask how often we could find a blogger who is among the top 5 most central people in the topic-graph using the rankings of the general graph. For example, if one were to search the general graph for bloggers posting on a topic, rank them based on their

betweenness centrality, and return the top-5. How many of those would also be in the top-5 list among the bloggers when computing betweenness centrality on the topic-graph. Table 5 shows the answer, where we can see that for the top-5, almost 64% of the time you will not find even one blogger in the general search, 30% of the time you would find one, and you would never find more than 3. If we increase to top-10, these numbers become slightly better. However, we still find only one person 39% of the time, and we do not necessarily know which in the top-10 was actually the most central in the topic-graph because the rankings are so varied that many of the others returned in the top-10 were far lower in the topic-graph ranking. Thus, using the general graph is just not reasonable.

Unfortunately, using the simpler ranking methods such as volume and indegree are equally useless. From Figure 5 we know that bloggers tend to write about many things. In other words, they tend not to write about the same topic often. That means that many bloggers will have the same high volume and we cannot pick one over the other. We have the same problem with indegree. The network is relatively sparse and indegree will not work either.

5.3.3 Temporal Linking Behavior

We now turn to the last analysis of our topic graphs: the temporal behavior of linking and what we can learn from it. We will specifically be asking two questions: how do components evolve over time (following the study in Kumar et al. (2006)), and once components “merge,” what is the future linking behavior like?

The former question will inform us as to how topics evolve over time. Prior work suggests that it is primarily small components which merge with other components until they finally merge with the large giant components. We first want to verify that this holds, but secondly we also want to understand *who* does the merging? Is it the case that the large component links to the smaller component or vice versa? Perhaps both at the “same” time (to the extent we can measure it, since we have daily snapshots). We further want to understand whether the behavior is different for the topic graphs. We already know the topology is different, but is the merging behavior different as well when we take context into account?

Second, we are interested in understanding what happens *after* a merge. Do bloggers start linking more with each other, or do they still tend to link “locally”? How much does the merge really affect future behavior?

Component Evolution

We start by exploring component evolution use methodologies in the literature (cf., Kumar et al. 2006). Specifically, we want to know first whether we get the same evolution pattern which have been observed before. We analyzed

our data on a per-day snapshot. We looked, for each new day, at the bloggers who started blogging at that day and all the links formed on that day. Links between “new” bloggers were kept separate to identify new components. Links between bloggers (old or new) which connected two previously unconnected components is considered a merge in the directionality of the link. For example, if a blogger from a component of size 10 links to another blogger in a component of size 100, then that means the smaller component *merged into* the larger component (and vice versa if the directionality of the link is reversed).

We first analyze the general graph (all links are treated together) to get the general behavior. Table 6 shows how many times different size components merged with each other over the 3-week timespan of this study. From prior studies (e.g., Kumar et al. 2006), we know that small components tend to merge with large components. However, we here also see in *which direction* these merges take place and interestingly enough we see that the merges go both ways, quite strongly. We also see that very few merges take place in the middle region. This is largely due to the fact that we end up with one giant component (128K bloggers) and the next largest component is only 791 bloggers large. However, we see that this large component is quite good at linking into the smaller component with almost the same frequency. We also see quite a bit of activity (both ways) for components all the way up to size 8-20.

The question then becomes whether this behavior is the same for the topic graphs. We performed the same merge analysis across all the topic graphs and show the aggregate (sum total) merges over all the topic graphs in Table 7. There are two very interesting observations which spring out at us right away. First, there is a lot more activity in the middle region, with a lot of merges being done across the board. Second, we see that the merging is much more skewed, where there is a very strong trend of the smaller components linking into larger components, but not the other way around, where the larger components perform almost not merges, except to the very smallest components. This is quite a different behavior which was completely hidden when looking at the large graph and this behavior seems intuitively more realistic than what we see in Table 6. This again strongly suggests that topic graphs are a different way of analyzing what is going on, bringing new insights into the behavior of bloggers which was hidden in the general graph behavior.

However, temporal linking analysis should not stop once components have been merged. We have a unique opportunity to keep monitoring the network to see how bloggers link *after components merge*. We first ask about how introverted bloggers are. Specifically, do they tend to continue to link primarily in their own initial component, or do they start linking outside the initial component. Further, will we be able to see a behavior that’s equally true in the general graph

	1-2	3-7	8-20	21-49	50-149	150-499	500-1,499	1,500-4,999	5,000+
1-2	6,620	1,356	339	91	51	25	2	1,453	30,839
3-7	5,154	482	113	27	28	20	4	148	2,733
8-20	1,647	132	37	12	5	5		38	398
21-49	387	45	16	1	2		1	11	73
50-149	174	38	8	4	2	1		8	27
150-499	138	48	6		1			2	9
500-1,499	215	11	2						1
1,500-4,999	620	69	18	4	6	2			
5,000+	41,231	1,622	229	52	13	7	1		

Table 6 What size components merged with what size components over the three week timespan of this study. Cell (i, j) means that a component of size $(1, j)$ merged into (i.e., linked to) a component of size $(i, 1)$. For example, we see that a component of size 50-149 merged into a component of size 21-49 4 times.

	1-2	3-7	8-20	21-49	50-149	150-499	500-1,499	1,500+
1-2	42,631	31,209	27,272	29,226	58,399	96,280	45,116	2,730
3-7	21,586	3,440	1,904	1,852	3,487	5,932	3,114	250
8-20	5,095	815	373	298	477	820	462	51
21-49	3,212	638	234	144	235	390	211	11
50-149	2,874	294	131	94	163	200	103	4
150-499	5,838	190	42	49	75	87	16	2
500-1,499	4,904	193	52	5	5			
1,500-4,999	674	40	11	4				

Table 7 What size topic-specific components merged with what size topic-specific components over the three week timespan of this study.

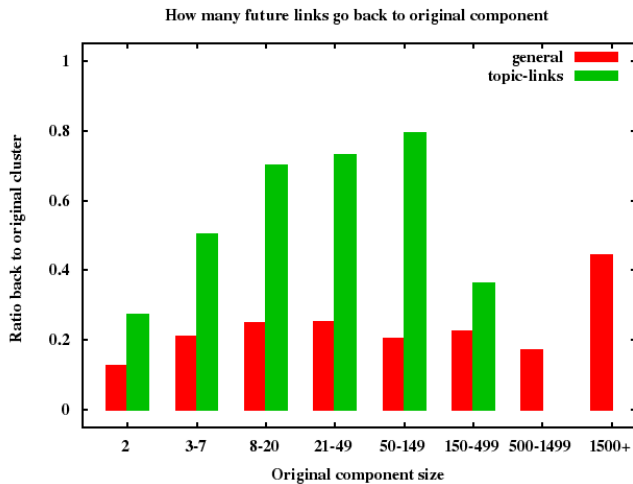


Fig. 11 What is the average ratio of new links (after a component has merged with another component), which point to a blogger within the same initial component. The x -axis is grouped by the sizes of the original component.

as in the topic-specific graphs? We plot, for each component C —after it has been merged with other components—the ratio of post-merge links which were internal to C (i.e., the link was from a blogger in C to another blogger in C). We ignore any singleton components (of which there is a large number), because we do not consider links to one self. Figure 11 shows the statistics for both the general graph and the aggregate statistics over the topic-specific graphs. The Figure clearly shows a significant difference in the two types of graphs. Again we see that the general graph, even though its

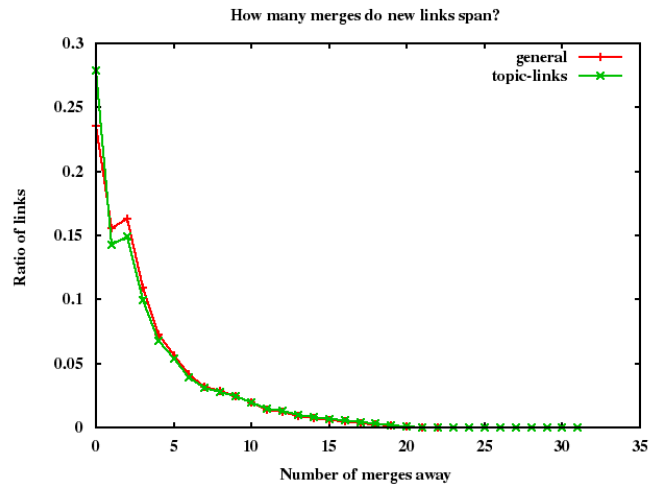


Fig. 12 How far do links travel in terms of merges? We see a near-identical pattern for the general and topic-specific graphs, although the topic-specific graph still has a higher ratio at a link-distance of 0.

ratios are quite high, still hides the much larger propensity for linking back to one's own starting component which we see in the topic-specific graphs.

This indicates that although components merge, there is still a very large propensity to stay close to the initial group. In fact, we can take this one step further and ask: out of the links which do not go to the initial group, how far do they go? More concretely, as we consider how components merge over time into larger components, we can define the distance a link makes as how many merges it needs to cross to get to initial component of the recipient blogger. Clearly the more

merges the link crosses, the further away that other blogger is with respect to this evolving component. Figure 12 shows the ratio of links being made at different distances. As we can see in the Figure the topic-specific graphs have a higher ratio at a distance of 0 (representing the introvert links we discussed above), but the behaviors are otherwise near-identical.

In summary, we found that there are clear linking behaviors that come out from doing topic-specific analysis which are impossible to get out of the analysis of the large general graph which hides these characteristics.

We have seen in all three parts of our network analysis study that the topic-specific graphs lend a new analytic power to dig in deeper to behaviors which are hidden by the larger general graph.

6 Conclusion

We proposed to enhance social network analysis on social network data sets which contain user-generated 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 answered by analyzing the generic social network alone.

We first described our methodology for taking social networks with user-generated text, pre-processing it using text-mining in the form of topic-detection, tagging links by the topic-context within which they were made and then analyzing the resulting enriched social network by only including the links for a particular topic.

We focused specifically on the questions of how the enriched networks differed from the general network. We showed that the topology was quite different quantitatively if not qualitatively in that the topic-specific graphs had far more small-to-medium sized components and distinctly lacked the giant component found in the general network. We showed that we could not find topic-specific central people using the standard graph, but that we had to analyzing the topic-specific graph specifically. Finally we showed that, in terms of the dynamics and the evolving components, the topic-specific graphs had significantly different characteristics which were completely hidden from view when analyzing the large general graph. In all of our cases we found that using the topic-specific graphs enabled us to get new insights into the behavior of bloggers.

This is clearly only a first step towards more indepth social network enrichment. For example, we here used an *ad hoc* method for selecting the best tags for a link, given the entirety of the blog post. Clearly, we should be able to fine-tune the context of a link to be more precise than using the entire blog post. In addition, we did not at all consider the

recipient when tagging a link. Intuitively one would assume that the blog post being linked to might carry more information about the context of the link. Further, we have not considered the profile of the blogger. If we analyze more blog-posts of a given blogger, we may be able to derive an in-depth profile of the blogger which also should be helpful in more in-depth social network analysis and mining.

References

- L. A. Adamic and N. Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.
- N. Agarwal and H. Liu. Blogosphere: Research issues, tools, and applications. *SIGKDD Explorations*, 10(1):18–31, July 2008.
- L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '06*, pages 44–54, New York, NY, USA, 2006. ACM. ISBN 1-59593-339-5. doi: <http://doi.acm.org/10.1145/1150402.1150412>. URL <http://doi.acm.org/10.1145/1150402.1150412>.
- P. Bhattacharyya, A. Garg, and S. Wu. Analysis of user keyword similarity in online social networks. *Social Network Analysis and Mining*, pages 1–16, 2010. ISSN 1869-5450. URL <http://dx.doi.org/10.1007/s13278-010-0006-4>. 10.1007/s13278-010-0006-4.
- D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.
- U. Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, May 2008.
- A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70, 2004. 066111.
- R. Ghosh and K. Lerman. A framework for quantitative analysis of cascades on networks. In *Proceedings of Web Search and Data Mining Conference (WSDM)*, February 2011.
- M. Götz, J. Leskovec, M. Mcglohon, and C. Faloutsos. Modeling blog dynamics. In *AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.
- R. Hanneman and C. Shelton. Applying modality and equivalence concepts to pattern finding in social process-produced data. *Social Network Analysis and Mining*, 1:59–72, 2011. ISSN 1869-5450. URL <http://dx.doi.org/10.1007/s13278-010-0009-1>. 10.1007/s13278-010-0009-1.
- M. Hearst and S. Dumais. Blogging together: An examination of group blogs. In *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media*, 2009.
- A. Joshi, T. Finin, A. Java, A. Kale, and P. Kolari. 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.
- R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proceedings of the 12th international conference on World Wide Web, WWW '03*, pages 568–576, New York, NY, USA, 2003. ACM. ISBN 1-58113-680-3. doi: <http://doi.acm.org/10.1145/775152.775233>. URL <http://doi.acm.org/10.1145/775152.775233>.
- R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, New York, NY, USA, 2006. ACM. ISBN 1-59593-

- 339-5. doi: <http://doi.acm.org/10.1145/1150402.1150476>. URL <http://doi.acm.org/10.1145/1150402.1150476>.
- D. Lazer, B. Rubineau, C. Chetkovich, N. Katz, and M. Neblo. The coevolution of networks and political attitudes. *Political Communication*, 27:248–274, 2010.
- J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1, March 2007a. ISSN 1556-4681. doi: <http://doi.acm.org/10.1145/1217299.1217301>. URL <http://doi.acm.org/10.1145/1217299.1217301>.
- J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *Proceedings of the Seventh SIAM International Conference on Data Mining (SDM)*, 2007b.
- J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2008a.
- 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. arXiv:0810.1355v1, 2008b.
- A. K. McCallum. MALLET: A machine learning for language toolkit. <http://mallet.cs.umass.edu>, 2002.
- R. Nallapati and W. Cohen. Link-plsa-lda: A new unsupervised model for topics and influence of blogs. In *Proceedings of the 2nd International Conference on Weblogs and Social Media*, 2008.
- M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.
- X. Ni, X. Wu, and Y. Yu. Automatic identification of chinese weblogger's interests based on text classification. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 247–253, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2747-7. doi: <http://dx.doi.org/10.1109/WI.2006.47>. URL <http://dx.doi.org/10.1109/WI.2006.47>.
- M. Papagelis, N. Bansal, and N. Koudas. Information cascades in the blogosphere: A look behind the curtain. In *AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.
- M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.
- D. Rosen, G. Barnett, and J. Kim. Social networks and online environments: when science and practice co-evolve. *Social Network Analysis and Mining*, 1:27–42, 2011. ISSN 1869-5450. URL <http://dx.doi.org/10.1007/s13278-010-0011-7>. [10.1007/s13278-010-0011-7](http://dx.doi.org/10.1007/s13278-010-0011-7).
- H. Sharara, L. Singh, L. Getoor, and J. Mann. Understanding actor loyalty to event-based groups in affiliation networks. *Journal of Advances in Social Networks Analysis and Mining*, 2010.
- X. Shi, B. Tseng, and L. Adamic. Looking at the blogosphere topology through different lenses. In *AAAI Conference on Weblogs and Social Media (ICWSM)*, 2007.
- S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge: Cambridge University Press, 1994.

A Larger List of Topic Words

(music) song sing lyric listen music sang chorus heard voice hear singer dance sung tune theme play verse version favorite karaoke catchy wrote remind sound video ballad lullaby anthem favourite album repeat heart youtube beauty humming guitar songwriter pop soundtrack radio rap inspire choir rendition compose band piano artist harmony

(food) cheese salad pasta tomato sauce dinner chicken bread mushroom onion potato green fresh pepper lunch garlic soup cream spinach dress olive slice cook food bean veggie grill sandwich lettuce mix egg carrot asparagus roast delicious meal recipe serve bake red tortilla broccoli italian spaghetti vegetable dish rice menu homemade

(political) iran iranian elect ahmadinejad mousavi tehran support revolution protest vote regime iranelection million happen street twitter khamenei presidential report ayatollah mahmoud candid country result inform president update fraud change govern world supreme islam won hossein hour mire universe count police kill peace speak moussavi political bbc leader victory facebook

(political) obama president administration bush barack white clinton house american washington policy political nation congress campaign america elect govern hillary democrat presidential george secretary office former czar promise press support biden unite senate vice critic mccain republican federal bill york agenda official michell stimulus chief politician kennedy economy public inauguration

(book review) book novel fiction story fantasy romance character author genre mystery reader write series plot science writer written review page publish history set literary world sci enjoy thriller protagonist heroine adventure paranormal horror narrate tale suspense detect literature style trilogy recommend crime romantic murder urban description fan theme detail sequel

(gaming) game wii nintendo mario super video play play zelda console arcade bro tetris metroid controller legend system fit final sega galaxy kart release buy sonic gamer plu sport smash ninja videogame origin classic dsi announce fantasy playstation pac motion luigi yoshi fun lite gameboy ddr snes screen rpg version

(financial) money pay save job earn spend paid cash income pay financial debt afford loan amount paycheck buy month dollar bill rent expense extra spent hard able pocket car sell cost worth payment finance broke invest account waste salary borrow sum budget poor payday buck send rich credit free buy

(religion) jesus god christ bible christian gospel word scripture verse church disciple faith teach lord kingdom paul follow testament matthew believe john preach truth live spirit passage spiritual biblical call revelation apostle heaven corinthian message command author mean luke speak world son resurrect prophet roman sin act father eternal holy

(pets) cat kitten kitty pet vet litter home meow dog paw food adopt feline fur claw strai tail cute box scratch house lap pur carrier feral hiss poor purr tabby cuddle outside shelter owner jump lick animal rescue hide adore brought snuggle neuter fluffy collar tiny fed eat mouse sleep

(nature) mountain trail hike climb lake rock valley view mile park canyon peak top road ridge feet trip river waterfall forest snow summit creek mount walk hill beauty cliff steep nation head glacier ski tree drive rocky stop reach drove water north fall hiker below route foot landscape picture camp

Table 8 Larger set of words for different categorical topics. Expanded version of Table 3.)

(14145)	ship automate loudtwitter tweet twitter daily twitter tinyurl plurk yay ugh tonight omg twitterp ramble congratulate today chirp digest twitter tweety ftw wtf blip awake tweeter twat hooray argh tweep woo bleh woot sleepy fml honk shipment woohoo birdy twt nutshell eep roundup magpie
(9574)	rate author summary pair character word disclaimer warn count fic note prompt spoiler fandom written beta none drabble mine own belong challenge genre slash imply ficlet angst feedback fiction profit mistake table series wordcount language category sequel season length pov fluff versus porn fanfic borrow crossover inspire timeline
(8643)	twitter tweet bit twitpick loudtwitter automate daily tinyurl ship follow via twitter tumblr update spymaster tonight playspy wow squarespace twittascope yay lol avatar iphone followfriday thank add helpiranelection awesome overlay tweetdeck twurl pic retweet iran support green democracy tomorrow spam trend iranelect omg assassin bkite ugh tweeti wtf
(7147)	haha hahaha damn super dinner okay lah meet till met realise nice home omg anyway mum hahahaha tsk alot lol shop school tmr heh caus singapore lunch bought thank wah funny reach wait hahah eat miss plu tire fun wanna session serious girl lor gonna photo hehe tuition abit
(6708)	icon credit comment preview please banner teaser base wallpaper hotlink graphic screencap batch header misc enjoy rule resource stock journal texture request manipulate textless theme text alter claim edit multifandom appreciate spoiler soo various super-natural create iyo general rest promo table bauble sample upload fanart multi angelamaria pic picspam
(6388)	video youtube watch http clip feature vid footage relate link upload channel vimeo embed emb edit song vlog fanvid music user tube cnn playlist amv documentary tribute quality imeem audio camera vodpod stream segment vidder parody disable videotape spoof nsfw hulu funniest slideshow trailer shockwav subtitl blooper mtv videograph
(5663)	album track record song release music band sound listen pop single artist rock debut product vocal label lyric hit solo singer studio chart feature download songwriter cover fan title vinyl compile tune version classic tour genre guitar perform punk mix heard disc origin career include collecte influence ballad indy
	• • •
(493)	unknown path road world ahead bring story decade follow wizard harry dark cast heal potter continue goes discover war entire future happen game gone current crossroad join lord truly hand chose friend meet mystery spent pursue broken night rule eager fell wound rise relief pain pave vivian death feet
(484)	third five fourth six fifth seven sixth eight seventh follow ten half include final nine twice grade past eighth previous count set remain rest row note ninth entire consider repeat complete earlier total except tenth twenty consist middle begin nearly mention ahead month straight quarter twelve consecutive eleventh plus
(481)	level low lower reach factor upper normal mean expect below range rate middle mid result limit advance extreme lowest drop effect basic rise demand reason current chance consider increase moderate standard intermediate improve amount usual relate continue due depend average change raise similar typic move combine degree goes adjust
(474)	world chaos war death gather left rebuild wizard eatery torn remain strike seek live follow fleur leave ago seven move plan determine choose power vision cast damage individual auror game rubble fenrir recuperation ministry minister endless current forth constant aftermath figure escape final call apart crack begin destroy tear
(470)	set aside dani complete stacy follow rest scene final help marcy timer begin ready couple trish separate note prepare extra previous close five add forth finish otherwise setup piece main entire straight line special fine stage decide expect record middle standard specify gather rough earlier able easy limit adjust
(434)	stock gain money profit market company report invest trade recommend rich share buy price month million opportune free play reader call special investor alert pick follow world research top easy exactly wealth simple chance service guarantee tell junior own fortune double cap trader potential set pay hit won issue
(389)	material length bust item price measure cotton name code dress belt shoulder remark sleev inner waist detail chiffon pre top line addition black register lycra mail satin elastic white silk lace sgd strap zip ruffle pink denim neck spree front crinkle email elegant please silky wide charge beige blouse

Table 9 Larger set of words for the top and bottom (ranked) topics. Expanded version of Table 2.