Noname manuscript No.
(will be inserted by the editor)

The Spread of Media Content Through Blogs

Meeyoung Cha · Juan Antonio Navarro Pérez · Hamed Haddadi

Received: date / Accepted: date

Abstract Blogs are a popular way to share personal journals, discuss matters of public opinion, pursue collaborative conversations, and aggregate content on similar topics. Blogs can be also used to disseminate new content and novel ideas to communities of interest. In this paper, we present an analysis of the topological structure and the patterns of popular media content that is shared in blogs. By analyzing 8.7 million posts of 1.1 million blogs across 15 major blog hosting sites, we find that the network structure of blogs is "less social" compared to other online social networks: most links are unidirectional and the network is sparsely connected. The type of content that was popularly shared on blogs was surprisingly different from that from the mainstream media: user generated content, often in the form of videos or photos, was the most common type of content disseminated in blogs. The user-generated content showed interesting viral spreading patterns within blogs. Topical content such as news and political commentary spreads quickly by the hour and then quickly disappears, while non-topical content such as music and entertainment propagates slowly over a much long period of time.

1 Introduction

Online social media like MySpace, Twitter and Facebook have emerged as a powerful communication tool for people to share and exchange information, ideas and thoughts on a wide variety of topics [31]. Information about real-world events are

Meeyoung Cha
Graduate School of Culture Technology, KAIST, Korea
E-mail: meeyoungcha@kaist.edu

Juan Antonio Navarro Pérez
Technische Universitat Munchen, Germany
E-mail: navarrop@mail.in.tum.de

Hamed Haddadi
Queen Mary, University of London, UK
E-mail: hamed@ee.ucl.ac.uk

shared in the form of text, hyperlinks, images, audio files, and other objects. Researchers have identified this active information sharing by Internet users as a new generation of journalistic conventions [28]. There is great value for mining this type of information for economical incentives and for understanding the psychological and social behavior of individuals on a large scale.

Several recent work has devoted to the problem of capturing the trend in adoption of social media. For instance, recent works using short messages on Twitter website use this data as a stream and produce events of a particular type such as news [33], media updates [35], local events [42], and earthquake alarms [32]. Sentiments embedded in short text updates in social media have been shown to effectively predict and even precede the daily stock price variation [7]. Likewise, the blogosphere has been shown effective in capturing up-to-date news [25]. In fact, a non-negligible fraction of news items shared in these social media are known faster than the traditional, authoritative news sources [38].

Weblogs, or blogs as they are referred to, have the longest history among the popular social media. Initially, a small number of blogs started out as an online diary and they rapidly became more prominent on the World Wide Web after year 2000 [6]. Blogs are often rich in hypertext links to other sites, which reflect a record of a user's latest browsing that is made available for others' interest. With the popular use of hyperlinks, blogs soon became an important social media platform, where 'bloggers' connect to each other to share and disseminate ideas, norms, and content. By 2003, more than two million weblogs cited other websites and interestingly a large fraction of them were known to focus on public affairs [5]. The practice of information sharing and the new journalism conventions have now spread to other social media such as Twitter and Facebook, and such a culture has become a significant part of today's Internet.[1]

In this paper, we study the trends in the use of blogs as social media to share and exchange information. We seek to contribute to the understanding of the 'new generation of journalistic conventions' as seen in blogs. More precisely, we are interested in knowing how blog users are connected to each other and what kinds of content are shared in the network of blogs over what time scale. Using the HTML links embedded in blog posts from a large data set of blog feeds, we extract the social relationships between blogs and construct the blog graph as proposed by Kumar et al. [21]. This allows us to examine the dynamic interactions between bloggers. We then study the types and the topics of content that is shared through such interactions. Our goal is to understand how a specific content (e.g., YouTube video) propagates in the blog graph and how do the spreading characteristics differ when comparing a video of a recent political event, against a music video.

This paper is based on the Spinn3r data set that was published in [39]. The data consists of web feeds collected during a two month period in 2008. The data set includes posts from blogs as well as other data sources like news feeds. We discuss our methodology for cleaning up the data and extracting posts of popular blog domains

[1] While over the recent years microblogging services like Twitter has become extremely popular, the traditional form of blogs still has a much larger number of users. Hence, we mention that it is equally important to study the blogging conventions as well as those in the newer microblogs. See our discussion in Related Work.

for the study. For the wider community use, we share the data parsed for this work at `http://navarroj.com/research/tools/`. Because the Spinn3r data set spans multiple blog domains and language groups, this gives us a unique opportunity to study the link structure and the content sharing patterns across multiple blog domains. For a representative type of content that is shared in the blogosphere, we focus on videos of the popular web-based broadcast media site, YouTube.

Our analysis, based on 8.7 million blog posts by 1.1 million blogs across 15 major blog hosting sites, reveals a number of interesting findings. First, the network structure of blogs shows a heavy-tailed degree distribution, low reciprocity, and low density. Although the majority of the blogs connect only to a few others, certain blogs connect to thousands of other blogs. These high-degree blogs are often content aggregators, recommenders, and reputed content producers. In contrast to other online social networks, most links are unidirectional and the network is sparse in the blogosphere. This is because links in social networks represent friendship where reciprocity and mutual friends are expected, while blog links are used to reference information from other data sources.

Second, concerning the interaction between different blog domains and language groups, we find that a significant portion of links span different blog domains. (Our notion of a language group is identified for each blog, rather than for each blog domain.) This result shows that blog interactions are not limited by the domain of the blog hosting sites. However, when it comes to language groups, we see few links between blogs of different languages. When they do occur, links between different languages tend to be unidirectional: the most common form is a non-English blog pointing to an English-written blog.

Third, media content spreads according to two broad patterns: flash floods and ripples. The first group includes topical content such as news, political commentary, and opinion. Like flash floods, these types of content spread quickly by the hour and then quickly disappear. This demonstrates the role of blogs as a social medium that helps and influences how opinions form and spread on current issues. The second group includes non-topical content such as music and entertainment. Like ripples, old content (produced more than a year ago) can get rediscovered and again start gaining the attention of bloggers, albeit at a slow rate.

Fourth, we study the content spreading pattern and content type in conjunction with the blog graph. Unlike our initial expectation, the type of content that were talked about popularly in blogs was surprisingly different from that from the mainstream media: content in Web 2.0 sites such as YouTube and Flickr was the most common type of content disseminated in blogs. We identified the top 10,000 YouTube videos that were linked by blogs in the blog graph. As a case study, we describe how one popular political video spread across the blogosphere and demonstrate a rapid, large-scale diffusion of media content along the blog graph.

Finally, we confirm the characteristic of a fast diffusion process in sharing news content among bloggers by focusing on 9 popular events that happened in 2008. These popular events include celebrity death news, society events, as well as the U.S. presidential election. We find that breaking news spark a rapid rise on the number of blog posts mentioning the topic, while the interest also rapidly dies out. For popular societal and political topics, we examined continued interests among bloggers

in debating about them. This result demonstrates the active role of blogs as a social medium, where people identify, discuss, filter, and disseminate interesting media content.

The rest of the paper is organized as follows. We first review related work. We then define terminology used in this paper, introduce the data set, and describe our methodology for parsing the data. Next we present two sets of analyses. The first analysis is about the structure of the blogosphere and the second analysis is about the patterns of content sharing and content diffusion on the blogosphere. We also examine the times it take for popular topics to spread in blogosphere. Finally, we summarize the results and conclude.

2 Related Work

A number of previous studies on blogs have looked into the structure defined by both explicit and implicit interactions between bloggers. Kumar et al., for example, focused on the evolution of the link structure in blogs over several years and proposed tools and models to study the communities formed by blogs [21,22]. They called the graph defined by links between blogs the *blog graph*. Using the concept of the blog graph, their goal was to study the evolution of connected component structure and microscopic community structure. Shi et al. similarly tracked the structure of the blog graph using multiple snapshots in time, and examined the topological characteristics such as the degree distributions and clustering coefficients [36]. Lento et al. investigated a person's tendency to keep blogging and their embedding in the online social network [24].

Building upon these studies, our study on the blog link structure expands our understanding about the topological characteristics across *multiple* blog domains (unlike within a single blog domain, which other work focused on). Furthermore, our main emphasis is on the interplay between the connection structure of blogs and its impact on content spreading, which the above studies did not consider.

Several studies focused on the interplay between the blogosphere structure and information dissemination, which is the topic that is more closely related to this paper. Gruhl et al. studied the diffusion of information in the blogosphere based on the use of keywords in blog posts [18]. Adar and Adamic used the explicit use of HTML links between blogs to track the flow of information [2]. We use the same methodology of examining HTML links in this work. Leskovec et al. developed algorithms to identify blogs which give the most up to date information on stories that propagate in the blogosphere [26]. They used the quotation marks and considered quoted text as a clean piece of information that spread in the network.

Compared to the work above, this paper focuses on two unstudied aspects in the blogosphere. First, we study what kinds of content (e.g., mainstream news, user generated videos or photos) are shared popularly in blogs. We investigate this by examining the URL domains of the content shared in blogs. Second, we examine how the topic (e.g., sports, music, news) of the content affects its spreading pattern within the blog network. We mention that all the above work focus on a single kind of content topic.

Other studies focused on the use of blogs and micro-blogs as a social medium. Bhagat et al. studied the demographics of multiple blog domains and characterized the interaction between blogs and the web [4]. Adamic and Glance measured the interplay between the liberal and conservative political blogs during the 2004 US election [1]. They found that liberal and conservative blogs rarely link to each other. Zhou investigated the communicative processes of political blogs and their implications in a specific region, China [44]. He found that the use of blogs gave quick responses to political events, by encouraging discussions on politically sensitive topics. Etling et al. conducted the link analysis on the Arabic blogosphere, based on 35,000 Arabic-language blogs, and studied how online practices are embedded in local political contexts [14]. They found that most blogs are focused on domestic political issues, while concern for Palestine was the one issue that united the entire network. The differences between topic-specific blogs and general blogs were also highlighted by Macskassy in [27].

More recently, a number of studies focused on the popular microblogging service, Twitter. Shamma et al. [35] demonstrated that Twitter can be used to track the popularity of planned political events such as the presidential inauguration speech. Sankaranarayanan et al. [33] demonstrated an exciting application of Twitter in identifying late breaking news. Yardi and boyd examined tweets about two geographically local events and checked whether geographic proximity can provide real-time information and eyewitness updates about the event [42]. Our group also looked into Twitter to investigate the social communication patterns of users and explore representative measures of user influence [11]. We found that topological characteristics such as indegree of a user alone does not capture other meaningful measures of influence, such as one's ability to spawn many retweets or mentions.

These studies emphasize the increasing role of online social networks like blogs as a social communication platform. With the widespread use of the Internet, the active use of social media will likely be a continued trend. Recently, microblogging services like Twitter has become extremely popular, because of its easy usage and accessibility in small mobile devices. Likewise, a great number of research is being conducted on this particular medium. Nonetheless, compared to Twitter's 100 million users (as of early 2011), blogs still remain as a much more popular social medium (already 130 million users owned blogs in 2002 and 350 million Internet users were reading blogs in 2008) [40]. Furthermore, because blog is an established platform, it had enough time to evolve and consolidate, involving less caveats with demographic bias than other young platform like Twitter (founded in 2006) [30]. Hence, we believe that understanding the types of popular content in blogs and their spreading pattern is of interest to the service providers, advertisers in social media, as well as to the research community.

3 Methodology

This section describes the data set and our methodology for cleaning up and extracting relevant blog feeds. This section also presents the high-level characteristics of

the data set. For the wider community use, we share the data parsed for this work at `http://navarroj.com/research/tools/`.

We first define the terminology. There are a number of *blog hosting sites* which allow an individual or a group of users to create a *blog*. These hosting sites provide, for each blog, a *web feed* which contains the latest entries (or *posts*) that have been published in the blog. Web feeds are also referred to as the RSS documents. Internet users can subscribe to web feeds of their favorite blogs in order to get updates whenever new content is published. However, not all web feeds originate from blogs. Various other content producers and aggregators, like web forums and online newspapers, also make their content available through web feeds.

3.1 Spinn3r data set

The data set, provided by the Spinn3r web service company, consists of 44 million web feeds crawled during a two month period between August 1st and October 1st, 2008. Because the data set includes all the posts available in the corresponding feeds at the time of the crawl, data for blogs with infrequent posting may include posts that were published long before the time of the crawl. Spinn3r groups individual feeds into 'tier groups' based on the influence rank (computed by their internal algorithm). Due to the massive scale of the data, in this study, we exclude any web feed of tier group "none" and focus on all web feeds that were assigned proper numbered tiers.

We parsed the XML documents describing each post to extract information such as the site URL, the post URL, language (identified by Spinn3r), and the time posted (or the time crawled if the former is not available). We scraped the content of all posts in order to search for links to web documents and embedded content such as images or videos. We discarded non-HTTP URLs and links that did not have a valid URL format. Since some blogs publish only summaries on their feeds, but not the full content that appears on the blog, we missed some of their HTTP links. This is a limitation of our study, imposed by the data set available to us.

In total, we identified 9,691,253 blog posts that were published in 1,225,720 feeds in 21,419 different web domains. The most active feeds include web domains such as craigslist.org, a popular website for classified advertisements, yahoo.com, which provides feeds for various news topics, and mckinseyquarterly.com, an online journal of business and management related articles.

3.1.1 Extracting the top 15 blog domains

Because our focus is on blogs, we need to identify blogs from the mixture of blog and news sources in the data set. In order to ensure that the blog posts we analyze are from individual blog users and not by popular media sources, we sorted the names of the web domains by the number of feeds they publish, and visited the domains with the most feeds to manually identify if they are blog hosting domains. In this way, we identified the top 15 blog domains, which we use in the rest of the paper. These 15 blog domains contained 90.7% of the entire Spinn3r data. Our heuristic is based on the assumption that popular blog domains likely publish many more web feeds,

Table 1 Summary of data set from 15 blog domains

Blog domains considered	Number of blogs	Number of posts	Posts per blog		Dominant language
			average	median	
myspace.com	390,812	1,217,757	3.1	1	English
live.com	321,730	1,161,103	3.6	2	CJK
wordpress.com	254,225	1,666,165	6.6	3	English
exblog.jp	72,376	1,127,383	15.6	13	Japanese
livejournal.com	66,598	2,120,474	31.8	28	English
blogspot.com	31,412	863,950	27.5	13	English
vox.com	22,572	234,794	10.4	6	English
yculblog.com	10,684	84,433	7.9	6	CJK
blogfa.com	8,386	64,377	7.7	9	Farsi
typepad.com	8,054	159,056	19.8	11	English
blog.com	3,915	4,021	1.0	1	English
over-blog.com	3,366	31,227	9.3	5	French
cocolog-nifty.com	804	17,899	22.3	13	Japanese
blogs.com	675	24,916	36.9	16	English
canalblog.com	803	17,428	21.7	14	French
Total	1,196,412	8,794,983	7.4	2	English

each one of them originating from an individual blog, compared to news sites where each feed might represent one of a predefined set of topics. Other ranking algorithms are discussed in [9]. However in this paper we focus on the contents of the blogs as opposed to differences in the ranking methodology.

Many blogs have their own web domains, but use standard blogging sites to host their posts. In order not to miss such blogs, we extracted the domain information from the post URL, rather than from the site or the feed URL. In order to clean up the data set, we further removed feeds originating from FAQs, forums, automated tag aggregations, and news sites (e.g., news.wordpress.com), which are clearly not representative of a typical blog. We identify each of the remaining feeds as an individual *blog* of the corresponding domain.

Table 1 displays the list of the selected 15 blog domains and their statistics. In total, we identified 1,196,412 blogs and 8,794,983 posts, respectively. The ranking of blog domains in the Spinn3r data set differs from other Internet statistics. According to alexa.com, blog.com is ranked much higher and exblog.jp and vox.com are ranked much lower. Figure 1 displays the Complimentary Cumulative Distribution Functions (CCDF) of the number of blog posts compared to all the posts for different publishers in the dataset. We observe that the figure has a power-law nature for all the feeds, however there is a sharp fall at the tail. The distribution for the blog posts displays a decay for the blogs with high posts and the steep rise in the tail is present. This shows that bloggers do not maintain a high level of activity as the ordinary publishers when it comes to mass publishing, a characteristic that is likely to keep a blog informative and useful and different in nature form content aggregation websites.

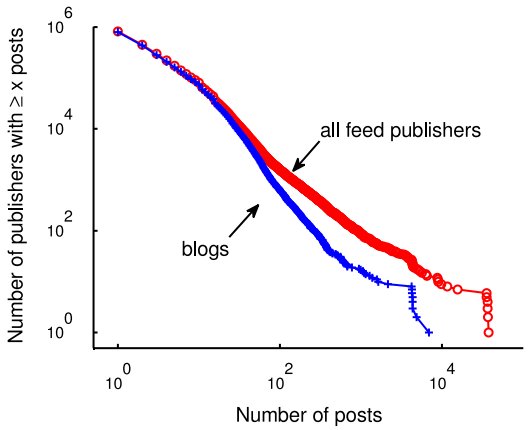

Fig. 1 Posting rate of feed publishers and blog users.

3.2 High-level characteristics

Here, we present the high-level characteristics of the blog feeds based on the language and the posting rate. In terms of languages, most blogs are written in English. However, the data set also included blogs written in Farsi, French, Spanish, and CJK (Chinese, Japanese, or Korean). Table 1 displays the dominant language for each blog domain.

In terms of the posting rate, we saw low content production trend (Figure 2). The average number of postings during a two-month period is 7.4 and the median is 2, indicating that most bloggers post only once or a few times a month. The content production rate varied largely across the blog domains. Blog domains livejournal.com, blogspot.com, and blogs.com showed the highest average posting rate. Myspace.com and live.com, which had the most blogs, showed low posting rate.

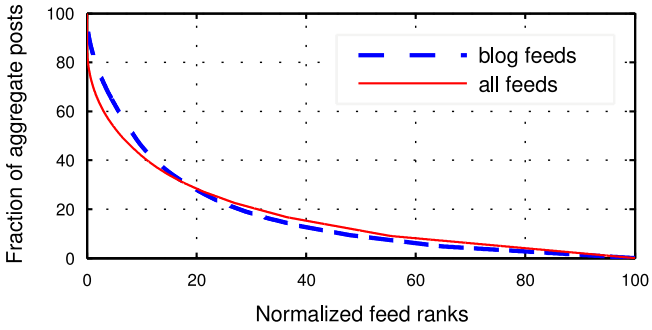

Fig. 2 Testing Pareto principle in blog post behavior

Furthermore, individual blogs varied widely in the number of posts they produced. To examine the variation, we plotted the number of posts for the r-th least active blogs in Figure 1. The horizontal axis represents the blogs sorted from the most active to the least active (from left to right), with blog ranks normalized between 0 and 100. The figure represents a cumulative plot on the horizontal axis, i.e., a value of 50 represents the total blog post counts generated from the less active half of all the blogs. We see that 20% of the most active blogs account for 70% of all posts, while the remaining 80% of blogs account for only 30% of posts. This skewed contribution of individual feed sources shows that the Pareto principle applies to the posting behavior. The Pareto principle (or the 80-20 rule) is widely used to describe the degree of skew in a distribution. The posting behavior of all web feeds, also shown in the graph, shows a similar skewed distribution.

We have manually visited the top 50 active blogs to understand which blogs are active in their posting behavior. We found that the active blogs tend to fall into one of the following three categories: (*i*) content aggregators or recommenders, that recommend other blogs or re-post content from other feeds; (*ii*) multi-owner blogs, where a group of people of a special interest produced content; and (*iii*) spam blogs or splogs. We find the usage and postings trends of active blogs is similar to those found in micro-blogging websites such as Twitter, where posts are typically brief text updates and are spread to a large number of individuals [11].

The remainder of this paper focuses on the structure and the content sharing patterns for the blogs listed in Table 1.

4 Linkage Structure of the Blogosphere

Blog posts often include HTML links to web pages such as videos, news articles, and posts from other blogs. The goal of this paper is to study these links in order to understand how blogs are dynamically connected and what type of content is shared among them. As a first step to answering these questions, in this section, we focus on the structural properties of the blog graph. We construct the blog graph following the methodology in [21]. Blog graphs serve as a fabric for information diffusion and spreading in blogs. Here we analyze the properties of the blog graph from two angles. First, what are the graph properties of the blog graph? Second, how are users across multiple blog domains and language groups connected in the blog graph?

4.1 Constructing the blog graph

We construct the blog graph as follows. There is a directed edge from node A to node B if any post in blog A links to a post in blog B. Even when blog A has explicitly cited blog B, we do not assume that blog B necessarily knows about blog A. We discard any HTML link to a blog that is not in the Spinn3r data set, even if the blog belongs to one of our 15 blog domains. Thus, we focus on the fraction of the blog graph for which we have full visibility, both for incoming and outgoing links. Our data set generated a network of 85,013 nodes with 129,079 edges, which accounts

for 7.1% of the blogs in Table 1. The remaining blogs are singletons and are not connected to any other nodes.

While blogs can be connected explicitly through 'blogroll' or the list of other blogs that a blogger lists as being interesting, the Spinn3r dataset does not include any information about blogroll. The blog graph, based on HTML referencing, represents implicit social relationship of blogs. A similar implicit relationship could be extracted from blog comments, trackbacks, and tags.

4.2 Structural properties of the blog graph

The first question we want to answer is what are the properties of the blog graph structure. For this, we examine global network properties such as the node degree distribution, reciprocity, and density. We compare the structure of the blog graph to the ones formed in online social networks.

4.2.1 Node degree distribution

We examine the degree distribution of all 85,013 nodes in the blog graph. The average number of edges per node is 1.5 and the median is one for both indegree and outdegree. Figure 3 shows the indegree and outdegree distributions. The horizontal axis represents the node degree and the vertical axis represents the cumulative number of blogs of degrees greater than or equal to a given degree. The two distributions exhibit a similar shape, forming a straight line in the log-log scale—a characteristic behavior of the power-law distribution. However, the two distributions differ in their shapes for degrees greater than 30. Except for the largest degree node, high degree nodes are more prevalent in the indegree distribution.

The tail degree exponent α of the power-law distribution $p(x) = cx^{-\alpha}$ is less steep for indegree ($\alpha = 2.5$), than for outdegree ($\alpha = 3.5$). A strikingly similar pattern was shown for the web [8]: α values are 2.1 for indegree and 2.7 for outdegree. Recently Shi et al. found a similar pattern in the blogosphere, although their outdegree distribution was curved [36]. These results—the high exponent of the outdegree distribution and larger indegree—reveal important insights about the blog graph structure: Shi et al. explain that while it is possible for one blog to attract a lot of attention (indegree) at a particular time, it is less likely that a single blog will lavish as much attention (outdegree) on as many different blogs in the same time period.

Our results about the power-law degree distribution of the blog network is in agreement with previous findings on the blog network, as discussed in [36, 37, 21]. Such a highly skewed distribution has important consequence in the varying degrees of information spreading efficiency individual bloggers could have. Our main focus is not to confirm such heavy-tailed node distribution, but to investigate later how such connectivity pattern affects the spread of content in the blog network.

4.2.2 Degree correlation and reciprocity

Next, we examine two other important graph measures: degree correlation and reciprocity. To see if nodes with high outdegree also have high indegree, we compare

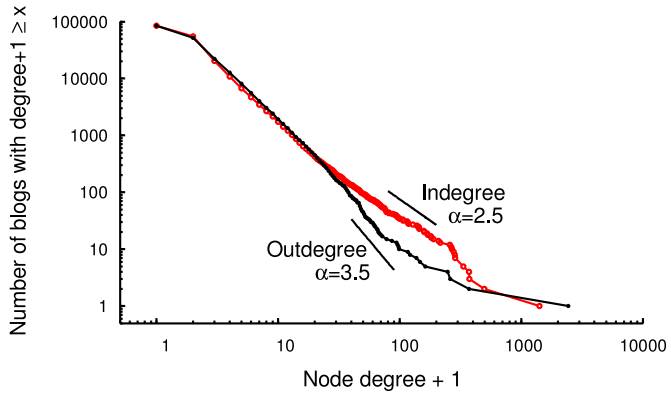

Fig. 3 Degree distribution of the blog graph

indegree of a node against outdegree. The Pearson's correlation coefficient between indegree and outdegree is 0.0664, indicating a weak correlation. Many blogs have no incoming links, but they linked to many other blogs. Other blogs had outdegree of zero, yet were linked to by tens of other blogs. The blog with the highest outdegree is "Blogs of the Day" from wordpress.com. It linked to 2,434 other blogs and received 43 incoming links. The blog with the highest indegree is "I Can Has Cheezburger?", also from wordpress.com, which contains funny pictures of cats and received 1,409 links. However, this blog did not have any outgoing links.

Overall, only 6% of the blog links are bidirectional. This could be because bloggers typically add HTML links to unilaterally cite information from other blogs and websites. More "interactive" actions such as comments and trackbacks have been shown to increase the level of reciprocity, up to 20% [36]. Unlike the blogosphere, online social networks exhibit high reciprocity. Many social networks like Facebook and Orkut, in fact, allow only bidirectional links. Even in social networks with unidirectional links, high reciprocity has been shown. For instance, in Flickr, 70% of the links are bidirectional [13].

4.2.3 Density

To better understand the structure of the blog graph we plotted several subgraphs of the blog graph and frequently observed tree-like local linkages. To measure the extent to which the local structure of the blog graph resembles that of a tree, we computed its *density*, which quantifies how dense or sparse a graph is. The density is defined, in the undirected version of the blog graph, as the ratio of the observed number of edges divided by the maximum possible number of edges.[2] This method is similar to the directed version of the density measure and due to the large size of the dataset we use the undirected version of the method. The density value of a node is typically

[2] The density measure could be also defined under directed graph [34], whose results are very similar to that in the undirected graph. We use the concept in undirected graph, in order to compare the results with that of other (undirected) social networks like Facebook.

calculated as the density of a subset of the entire network, consisting of all nodes and edges within a k hop distance of the node [23]. For each node, if the number of nodes in a k-hop neighborhood is N and the number of edges in the neighborhood is E, density is defined as:

$$D = \frac{2E}{N(N-1)} \ .$$

For each node in the blog graph, we calculate the density based on its 2-hop neighborhood and compare the value with the density of a synthetic tree. A synthetic tree with N nodes has $N-1$ edges and so has a density of $2/N$, which corresponds to the smallest density possible of a node. Figure 4 shows the density of each node in the blog graph as a function of its neighborhood size N. The axes are in log-log scale. For comparison, we also show the density of the Facebook network [29] for a randomly selected sample of 10,000 nodes. The median difference between the density values of blogs and nodes in a synthetic tree of the same node size is 0.0023, while the median difference between nodes in the Facebook network and in a synthetic tree is 0.0156, one order of magnitude larger. The density of the blogs shows a strong linear correlation to the density of a tree; the correlation coefficient is 0.9811. Note that while the plot diverges for large x values, the impact of this difference is not significant since the y axis is in log scale.

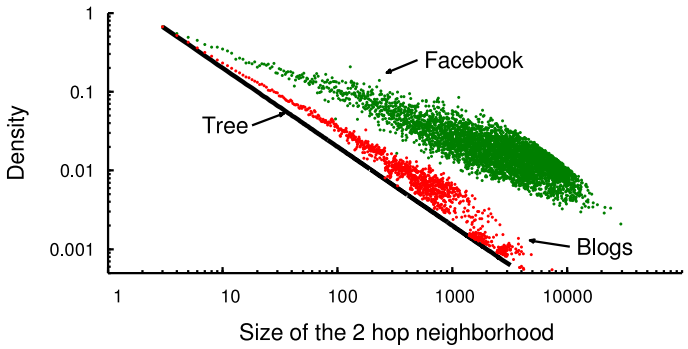

Fig. 4 Comparing density of the blog graph with others

An alternative metric to determine the "treeness" of a graph is the so called *circuit rank* [16]. Circuit rank represents the number of edges that must be removed from the undirected graph in order to make the graph cycle-free. This value is calculated as $E - N + C$, where E is the number of edges, N is the number of nodes, and C is the number of connected components in the graph. In the case of the blog graph, removing 31% of the links is enough to turn the graph into a set of trees, while in the more densely connected Facebook network, 95% of the links have to be removed. This further supports our finding that, unlike other social networks, the connection between blogs is sparse and more similar to a tree. This indicates that two connected bloggers are less likely to share a common friend than in other social networks like

Facebook. Rather the blog network forms a shape that is most efficient for information propagation.

4.3 Connection across different blog profiles

So far we have examined the structural properties of the blog graph. Here, we investigate the role that different blog profiles (i.e., blog domains and language groups) play in determining the structure of the graph. In particular, we are interested in knowing whether bloggers are less likely to form links to blogs hosted in other domains or written in different languages. To answer this question, we measure the fraction of links that are formed between blogs of different domains and languages.

Our analysis reveals that blog interactions do occur beyond the boundaries of blog hosting sites: 66% of the edges in the blog graph join blogs from different domains. This also suggests that analyzing the blog graph based on the data from a single blog domain will miss a lot of the rich linkage structure in the blogosphere.

To analyze the effect of language groups in the formation of links, we examine the language for each node in the blog graph. Note that language group is identified for each post by Spinn3r. Because individual blog can have posts written in different languages, we assigned to each blog the most common language that was used by the blog.

Our results show that, indeed, language is a barrier for link formation. In total, 93% of the edges join nodes of the same language. The remaining 7% or nearly 7,000 edges join blogs of different languages. This means that, as opposed to blog domains, language imposes a barrier that can effectively partition the network and prevent the flow of information. Not surprisingly, a large fraction (35%) of the links between blogs that speak different languages occur when a non-English blog points to a blog written in English. In blogs with high variability of languages, the language barrier becomes less dominant, however multi-language blogs are rare due to the conversational nature of blogging.

4.4 Summary

In this section we observed that the blog graph has three structural properties: (*a*) the node degree distribution is heavy-tailed, (*b*) links are not reciprocal, and (*c*) the network structure is sparse and, compared to other social networks, closer to that in a tree. Nodes with high indegree may represent popular media sources or trendsetters among bloggers. A sparse structure may indicate that bloggers have a clear preference for the blogs that they follow up or recommend. With respect to blogs with different profiles, we saw that blogs from different domains interact freely, while language imposes barriers that can potentially prevent the flow of information on the blog graph.

5 Content Sharing in the Blogosphere

The previous section focused on structural properties of the blog graph, which affect its efficiency for information dissemination. In this section we focus on the types of content bloggers talk about and the patterns of content sharing in the blogosphere. Our goal is to gain an insight into how different types of content affect the shape of the blog graph.

We present the following three sets of analyses. First, using information about web links embedded in blog posts, we examine which websites bloggers frequently link to. Second, we pick YouTube videos as a representative type of content that is shared in blogs and study the characteristics of the popularly linked YouTube videos. Third, we correlate the spreading pattern of YouTube videos with the blog graph and check whether any video caused a large-scale diffusion.

5.1 Commonly linked websites in the blogosphere

We describe the high-level properties of the HTML links embedded in blog posts, based on the data set of 8.7 million blog posts. While the usage of links varied widely across blog domains and among individuals, nearly 40% of the posts contained at least one HTML link. Figure 6 presents the CCDF of the blogs with links and all the blogs. We observe a sharp fall at the tail of the distribution for the most active blogs, bringing the top publisher to a tie in number of links posted.

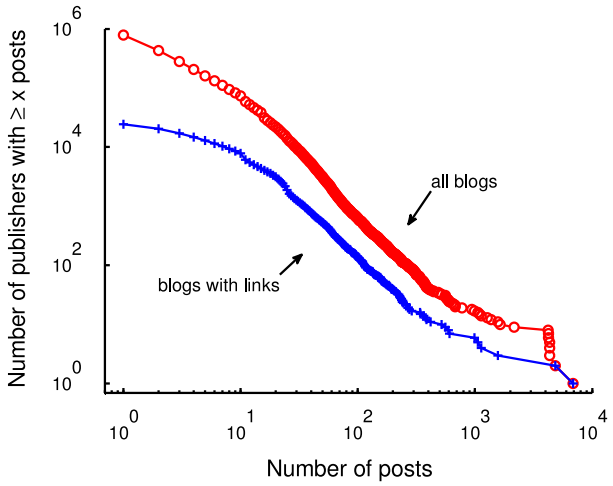

Fig. 5 Posting rate of feed publishers and blog users.

This prevalent usage of links is due to self-links: 60% of the posts with at least one link contained a self-link, referring to one's own blog. A self-link typically appears when a blogger explicitly cites one of his or her previous posts or has uploaded

multimedia content. A self-link can also be added automatically by the blog hosting site (e.g., providing links for readers to comment).

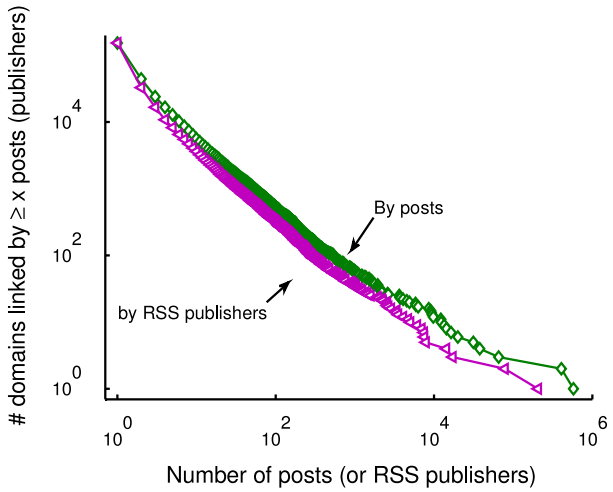

Fig. 6 Posting rate of feed publishers and blog users.

Interestingly, posting links to blogs is not dominated by use of RSS feeds or by directly posts. Figure 6 displays the CCDF of blog postings by these two methods. We observe that both methods follow a heavy tailed distribution, with RSS feed posts narrowly coming second to direct posts. However for the tail of the distribution, the direct posts present the majority of links.

To examine the types of content shared in blogs, we exclude any link that points to known blog domains and focus on HTML links to external websites. The 20 most popularly linked websites include content sharing sites, online shopping websites, mainstream news media sites, web portals, and social media sites like wikipedia.org and digg.com. Table 2 displays the top 15 websites along with the total number of blog posts that linked to the corresponding website. The top websites differed from one blog domain to another. For instance, the number of links to websites such as reuters.com and technorati.com is highest among blogspot users, whereas links to microblogging messages from twitter.com are the most popular among livejournal users.

The top 4 sites in the list are websites for sharing user generated videos and photos, indicating that bloggers like to talk about multimedia content. Online retail website amazon.com ranked fifth, indicating that bloggers also frequently talk about products like books, songs, and videos. These findings are consistent with the ones reported on earlier studies on web content in blogs [4].

We initially expected blogs to link to content in mainstream media like newspaper websites. Although links to mainstream media like nytimes.com and bbc.co.uk do appear in the top list, the number of blog posts linking to them is an order of magnitude smaller than links to user-generated or “home-made” content. The extreme popularity

Table 2 Top 15 linked websites

Rank	Web link domain	# blog posts with links
1	youtube.com	206,803
2	photobucket.com	140,194
3	flickr.com	135,327
4	imageshack.us	41,997
5	amazon.com	36,379
6	nytimes.com	33,801
7	twitter.com	30,572
8	technorati.com	27,583
9	tinypic.com	23,899
10	bbc.co.uk	20,893
11	imdb.com	16,649
12	cnn.com	16,640
13	digg.com	15,348
14	facebook.com	11,817
15	wikipedia.com	4,676

of user generated content in blogs is one of the main differences between the 'new' media and 'traditional' media, because traditional media (such as news websites) predominantly link to established, authoritative news institutions [28]. In contrast, blog users are open to citing less authoritative sources, i.e, other user generated content. This difference shows the usage of blogs as a new generation of journalistic conventions that could potentially break the general conservatism and rigidity of many journalism's practices and act as a new outlet that shares more long-tail content.

While different blog domains prefer different websites for linking photos and news, YouTube is ranked first for almost all blog domains and it received the most number of links in total. YouTube videos are also some of the most popular shared links among other social networking and micro-blogging sites such as Facebook and Twitter. Thus, we focus on links that point to YouTube videos and characterize their content sharing patterns among blogs.

5.2 Content sharing patterns of YouTube videos

Here we focus on HTML links to YouTube videos and examine three aspects of content sharing patterns in the blogosphere: (*a*) what are the topics and categories of videos that are popular; (*b*) what is the age of the shared videos (i.e., are old videos rediscovered through blogs); and (*c*) how quickly do links to the same video spread in the blogosphere.

Our data set includes a total of 279,081 HTML links to 202,658 distinct YouTube videos, indicating that some blog posts linked to multiple videos in YouTube. Interestingly, the number of HTML links to YouTube videos in the blogosphere follows Zipf's law, as shown in the log-log graph of popularity distribution in Figure 7. This hints us at the existence of a large-scale diffusion of YouTube videos. The most popular video received links from 375 blog posts.

To understand the characteristics of the popularly shared videos, we downloaded the metadata of the top 10,000 YouTube videos using its Data Application Program-

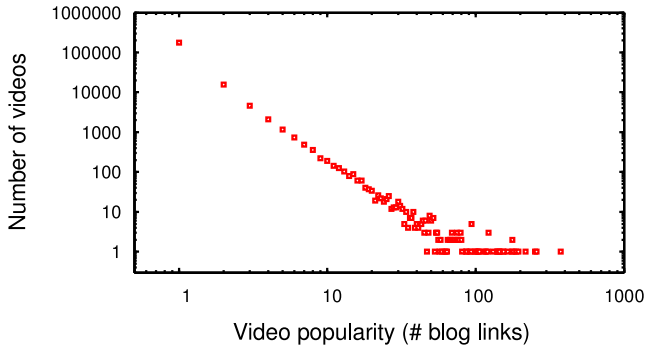

Fig. 7 Popularity of YouTube videos in the blogosphere

ming Interface (API).[3] Using the YouTube API, we wrote a Python script to automatically download information about the uploader, view counts, tags, category, and duration of all 10,000 videos. Each one of these popular videos received at least 3 links from blog posts, and all of them together received 68,826 or 25% of all links to YouTube. Although recommendations also play a role in YouTube video popularity [43], in this paper we focus on links directly shared on blogs. For the wider community use, we also share the YouTube video information we crawled. In the remainder of this section, we present analyses of these 10,000 videos.

5.2.1 Categories of the linked videos

We examine what kinds of video categories are popular in the blog network. Table 3 displays the top 10 user-assigned categories based on the number of videos, among the top 10,000 videos, that were linked by blogs in the data set.

Table 3 Top 10 video categories

Category	Perc. of videos	Perc. of links
Music	23.5	18.4
(taken down)	22.3	19.9
News & Politics	19.6	27.2
Comedy	8.9	9.7
Entertainment	8.8	8.0
Film & Animation	4.7	3.8
People & Blogs	2.9	2.5
Science & Technology	1.5	2.4
Pets & Animals	1.2	1.7
Education	0.9	1.4

Music videos accounted for the largest number of videos, but videos on news and politics received the most links. Two channels in the category of news and politics

[3] `http://code.google.com/apis/youtube/overview.html`

"BarackObamadotcom" and "JohnMcCaindotcom" collectively received the most links on their uploaded videos, indicating that high popularity of this category is due to the U.S. Presidential Election in 2008. Nearly a quarter of the videos in the top list were *taken down*. YouTube has strict policies with regards to content ownership and community awareness[4] and quickly removes videos in breach of its terms of service. Yet it is interesting that these videos had already gained huge popularity in the blogosphere before being removed.

5.2.2 Age of the linked videos

Our next focus is on the age of the linked videos. We are interested in knowing whether bloggers are keen on the latest produced content or rediscover old content. To check this, we examine the time between the video upload (in YouTube) and the blog linking. We observe large variations across individual videos as well as different video categories. Due to space limitation, we show the results for only the top 4 video categories: music, news, comedy, and entertainment.

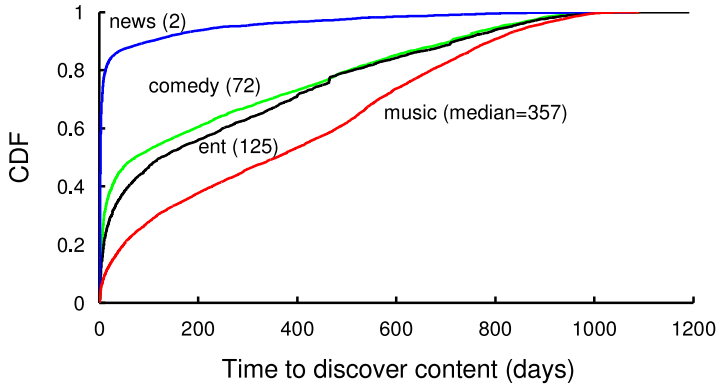

Fig. 8 Age distribution of videos in the blogosphere

Figure 8 displays, for each category, the distribution of the video age at the time each link was formed. The horizontal axis represents the time difference between the video upload and the blog linking. The vertical axis shows the cumulative distribution plot (CDF) of the number of links. Next to each plot we show the median age of liked videos in units of days. The median age for videos in the news category is 2 days old, and some links appeared within a few seconds to minutes of the video upload. Very few news videos were linked after a year of being uploaded. This demonstrates that news videos that spread in the blogosphere are topical and young. The other video categories show a pattern of a much delayed discovery; the median age of a comedy video is 72 days at the time it was linked by a blog. The median age of videos for entertainment is 125 days and is 357 days for music! This indicates

[4] `http://www.youtube.com/t/community_guidelines`

that bloggers post about recent events when it comes to news and politics, but also enjoy rediscovering old content (nearly one year old) for other video topics. This is in contrast with previous findings on the age of content on YouTube [12], as the blogs help users in rediscovering old content.

5.2.3 Diffusion time lag in blog links

Given that videos are discovered at different rates depending on the their topics and categories, we next examine how are the links of the same video correlated in time. To understand this, we first sort the blog posts based on the blog post time. Then we calculate the time taken for the video spreading as two values, which we call *half-spreading time* and *full-spreading time*. The former is defined as the number of days that diffusion of a video took, starting from the first post of the video, to the 50% of all links to the video to appear. The later is the number of days between the posting of the first and the last blog post that had a link to a given video. Figure 9 shows the median values of the half times and full times of videos for the 4 video categories. Recall that, although the Spinn3r trace spans only two months, Spinn3r's web crawler can discover posts that were much older than two months for blogs that published posts infrequently.

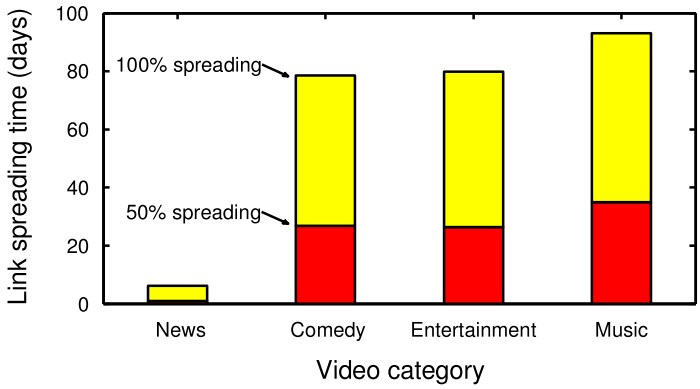

Fig. 9 Time lag in the spread of videos in the blogosphere

The bar plot in Figure 9 shows that most news videos gain their popularity within the first few days of diffusion. The median half-spreading time is one day and the median full-spreading time is one week. This indicates a fast diffusion process of the news category, where users respond to popular videos within few hours. Other categories show a much delayed spreading pattern. The median half-spreading time is around 30 to 40 days for comedy, music, and entertainment categories, and the full-spreading time is more than 2 months. This means that bloggers are more relaxed in following up on non-topical videos for these categories.

The distinct time variations seen in information spreading across different topics has not been reported in the literature, hence needs further investigation. The two

broad patterns resemble *flash floods* and *ripples* in their speed of propagation. Like flash floods, topical content such as news, political commentary, and opinion spreads quickly by the hour and then quickly disappears. This demonstrates the role of blogs as a social medium that helps and influences how opinions form and spread on current issues. Non-topical content such as music and entertainment resembles ripples. Like ripples, old content (produced more than a year ago) can get rediscovered and again start gaining the attention of bloggers, albeit at a slow rate.

An immediate implication of the flash floods and ripples-like spreading patterns is on content recommendation and advertising. First, in the long-tail content systems like YouTube and Flickr, content recommendation becomes an important task. Our study confirms that topical content have a much shorter popularity-cycle than the non-topical content. Hence, a content recommender system could take into about the type and age of content and its estimated popularity-cycle. Second, the popularity-cycle also could affect advertiser's strategy on deciding where to insert online advertisements. Once the peak popularity has passed, it might be more viable to place advertisements on non-topical content that will continue to gain user attention over an extended period of time, such as baby photos (rather than a topical political photo footage).

5.3 Content spreading over the blog graph

Finally, we present a theoretical analysis of the content spreading patterns along the blog graph. While bloggers can independently link to the same YouTube videos and discover them externally, we are interested in scenarios in which information about YouTube videos propagate within the blog network. While the Spinn3r data [19] we used does not have a complete coverage of all the blogs in the Internet, it is by far the largest realtime data repository of blogs and provide representative subset of blogs. Therefore, our study provides a meaningful lower bound case of the spreading that could happen in the blog network.

We assume that a YouTube video can spread in the blog graph if the following two conditions are met:[5] (*i*) information can flow in the direction of edges in the blog graph, but not in the reverse direction; and (*ii*) information can flow from one blog to another blog in a time-increasing order of their link posting. These conditions mean that a video can spread from node A to node B if there exists a directed edge from A to B and if A posted the video link prior to B.

In total 2,401 or 24% of the YouTube videos had any spreading in the blog graph (i.e., linked to by at least two bloggers who are directly connected in the blog graph under the diffusion conditions). These are the videos whose spreading was potentially aided by the linkage structure.

We show the diffusion pattern of the video that was propagated most widely in the blogosphere in Figure 10. The video, uploaded by YouTube user "JohnMcCaindotcom", is related to the U.S. presidential election. For clarity, we only show the nodes and the edges that are related to the diffusion of the video. The direction of

[5] In this dataset we are unable to verify if the videos were discovered independently by a user, or were shared as a result of a recommendation by another blogger.

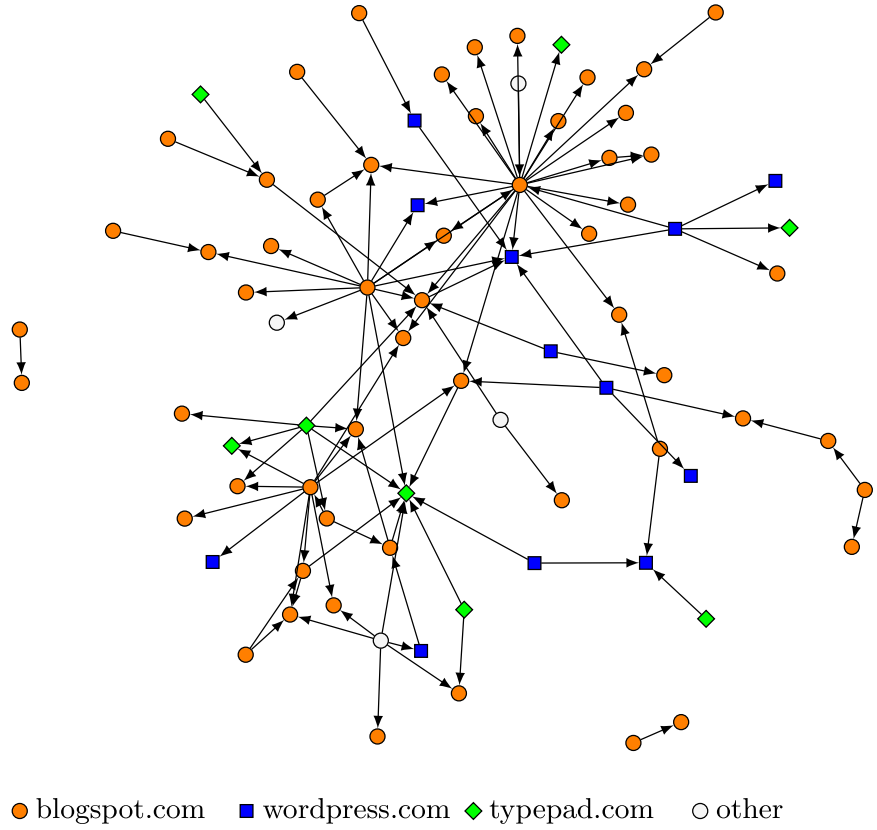

Fig. 10 Diffusion of a political YouTube video in the blog graph. Node styles denote different blog domains.

edges indicates the direction of information flow. Edges that fail to meet the time ordering of video link posting are removed. The video received HTML links from 79 blogs that had 105 edges between them (in the appropriate time order). The diffusion network formed a large connected component and two disconnected node pairs. It took less than a week for blogs to form the large connected component. We also see that the video spread across multiple blog hosting domains (e.g., blogspot.com, wordpress.com, typepad.com). This example demonstrates that a large-scale diffusion of information can occur along the links in the blog graph at a rapid rate and across domains.

6 The Spread of Various Topics

In this section, we repeat the analysis on the spreading times of different topical categories (in Section 5.2) to more general blog posts by examining the topics that blogs covered. This step is to ensure that the results we observed in the previous section are not limited to blog posts with links to YouTube videos.

Unlike in the case of using YouTube video links, the discussion of the same topic among blogs cannot be correlated explicitly due to a network-wide cascade. This is because (1) the prominent topics were covered widely by various media sources like newspapers giving a high chance for the blog users to be exposed to the topics outside the blogosphere and (2) mentioning the same keyword such as a celebrity name by neighboring bloggers does not necessarily guarantee that the blog content is on the same event. Hence, we do not focus on the connectivity of the blog users, but rather on the time it took for a given topic to be mentioned by blog users.

Therefore, "spreading" of topic in this section indicates the general process in which individual blog users are adopting and discussing a particular topic, similar to the context of diffusion of innovations [31].

6.1 Popular topics in 2008

In order to find popular topics discussed in the blogosphere, we focused on a set of prominent events that occurred in 2008 by consulting various media sites and publicly available lists.[6,7] Among the prominent events, we gave particular attention to those that occurred during Aug 1 – Oct 1, which corresponds to the timespan mainly covered by the Spinn3r data.

Table 4 Summary of 9 topics studied

Category	Event ID	Description	# posts	# bloggers
Celebrity	Mac	Actor Bernie Mac dies at 50 from pneumonia on Aug 9	11,717	11,693
	Duchovny	Actor David Duchovny enters rehab for sex addiction on Aug 28	2,098	2,097
	Newman	Actor Paul Newman dies at 83 on Sep 27	5,728	5,727
Society	Stock	The Dow fell to an astounding all time low	42,248	41,864
	Tibet	The breakout of Tibetan unrest on March 14	14,589	14,435
	Olympics	The 2008 Summer Olympics held in Beijing	44,179	43,825
Politics	Election	The presidential campaign in the U.S.	42,388	42,044
	Obama	The first African American to be voted into presidency	113,747	112,736
	Palin	Governor of Alaska in the U.S. related to the election	113,110	112,651

Table 4 shows the list of events that we studied. To identify blog posts related to these events, we chose keywords describing the events such as the celebrity names for most events and 'beijing'+'olympic' and 'tibet' for the other events. The table also shows the number of blog posts and the number of distinct blog users who mentioned the topic. In high-level, these events can be categorized into 3 topics: celebrity, society, and politics. Only the celebrity events had specific starting dates, which allowed us to examine their first phase of spreading. Other events were popularly talked about throughout the year.

[6] http://www.tvguide.com/special/best-of-year-2008/photogallery/headlines-1000425

[7] http://www.faqs.org/shareranks/2361,Hottest-Headlines-of-2008

6.2 Spreading time of different contents

We first observe the frequency of blog posts on each of the 9 topics. Figure 11 shows the number of daily blog posts on four of the topics. News on celebrity deaths, as shown in Newman and Mac, cause large bursts of blog posts that contain celebrity names; the peaks coincides with the dates these celebrities died. This finding demonstrates a fast diffusion process of sharing news content among bloggers. Upon the start of any breaking news, we observe a burst of blog posts that cover the event. We also see that the number of blog posts on these topics drops quickly over time, indicating the short interest span of users.

On the other hand, topics on society and politics such as Tibet and Obama exhibit continued interests among bloggers over the two months period. The topic of Tibet was dealt by a fewer number of bloggers than other society events, generating only between 50-750 new blog posts every day. Its post frequency is harder to predict, because debates on Tibet within blogs were often generated by news media that published articles on Tibet in a non-regular fashion. The topic of U.S. presidential candidate Barack Obama sparked interests with a more regular pattern than Tibet. The posting rate shows a weekly pattern, where the weekdays generated more blog posts than weekends.

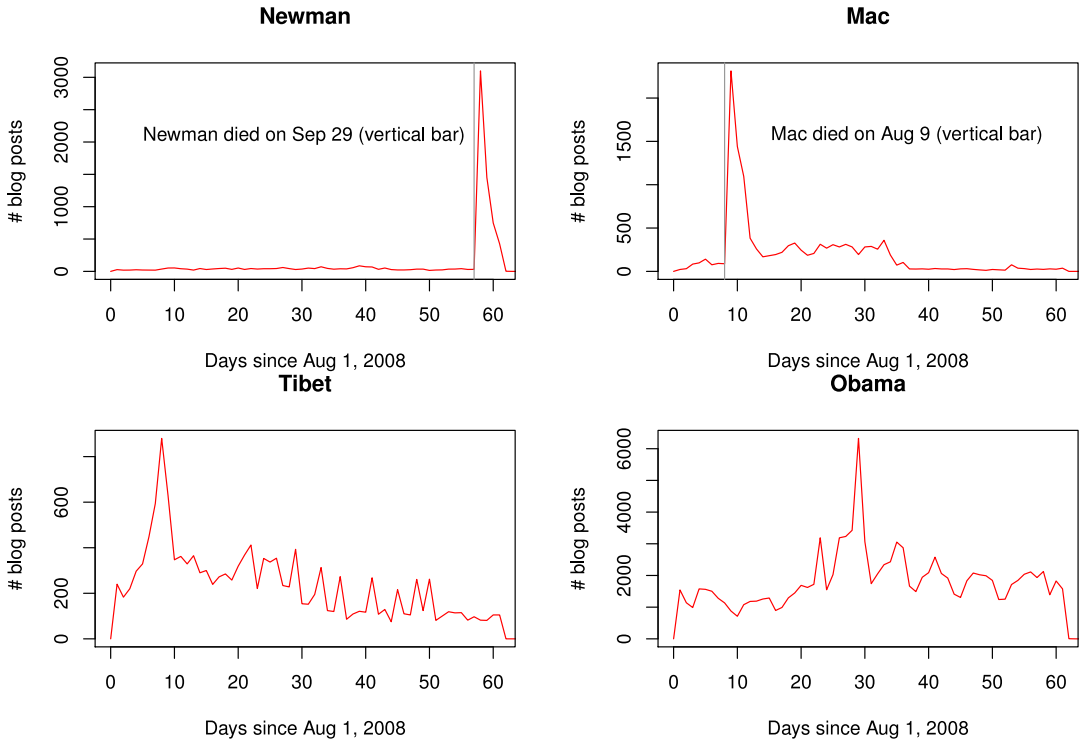

Fig. 11 The number of daily blog posts on different topics in the blogosphere

Given that topics in Table 4 are discussed with varying frequencies and patterns, we next examine the characteristic times of the topic spreading. Similar to our analysis in Section 5.2.3, we examine the *half-spreading time* and *full-spreading time* of each topic. The half-spreading time is defined as the number of days that spreading of a topic took among bloggers, starting from the first post on the topic, to the 50% of all posts to appear. In case of the celebrity events, which are event-driven, we only consider the blog posts that appeared on the starting date of the event and onward. For all other events, we take Aug 1st of 2008 as the starting date. The full-spreading time is the number of days between the posting of the first and the last blog post on a given topic.

Figure 12 shows the spreading times on the 9 topics. The half-spreading time of all three celebrity death news is less than a week, similar to the rapid spread of YouTube videos on news (see Section 5.2.3). In case of the Newman event, the half-spreading time took less than 1 day and the entire event was discussed for only 4 days, which is an artifact due to the data containing up to Oct 1st, 2008. The other two celebrity death events, while their half-spreading time is quick, continues to spark some level of discussion among bloggers for a much longer period of time as shown in the full spreading time.

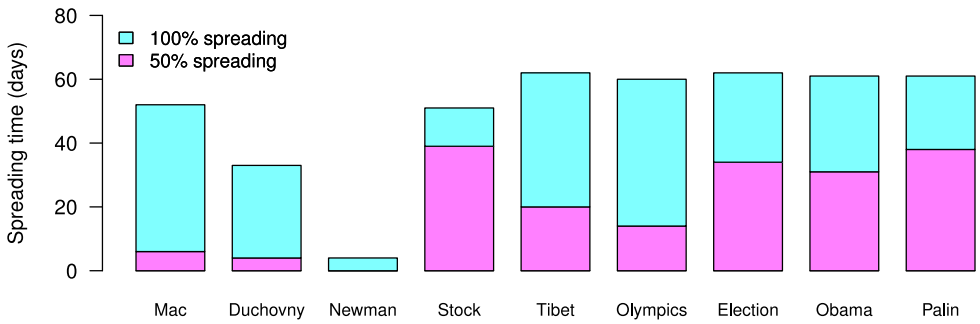

Fig. 12 Time lag in the spread of 9 different topics in the blogosphere

The three society events—Stock, Tibet, and Olympics—and the three political events—Election, Obama, and Palin—in contrast show a much larger half-spreading time of 14–39 days. Given that the maximum spreading time is 62 days (Aug1–Oct1), the half-spreading times closely resides on nearly the half of this range for political events, which indicates that the topic on politics were discussed almost regularly throughout the trace period. This is because the collection time of Spinn3r data was close to the data of the U.S. presidential election, November 4th, 2008. Hence, many bloggers were engaged in these topics in this period. The findings in this analysis again confirms the role of blogs as a social medium, where people actively spread breaking news and discuss issues on society and politics in at least two very distinct and recognizable patterns.

7 Concluding Remarks

In this paper we studied the trends in the use of blogs as a social medium to share and exchange information and sought to contribute to the understanding of the new generation of journalistic conventions. We conducted an analysis of two months worth of web feeds from 15 popular blog hosting sites on the Internet. This accounts to massive data on 9,691,253 blog posts, 1,225,720 blog feeds in 21,419 different web domains. Based on the HTML links embedded in blog posts, we constructed a blog graph or the blogosphere that captures the implicit social relationship between blogs and studied its network structure. We also analyzed the patterns of content sharing in the blog graph.

We demonstrated that the blogosphere has unique structural properties that distinguish blogs from other social networks. Similar to social networks, the node degree distribution is heavy-tailed. However, unlike users in social networks, bloggers do not exhibit a strong inter-personal relationship; only 6% of the edges in the blog graph are bidirectional. Most of the links point to a small set of popular blogs and, in some sense, demonstrate attachment to their particular preferences. As a result, based on the density and the circuit rank measures, the overall structure is sparse and closer to the shape of a tree. The low level of reciprocity and density clearly differentiate blogs from other social networks in that they allow bloggers to tune their subscription lists to other interesting bloggers' updates and effectively disseminate information to their followers.

We also examined the popularity of content sharing in the blogosphere. In particular, user generated multimedia content was the most frequently shared content among bloggers. This is in sharp contrast to the old media that predominantly cite authoritative sources. Our study of the diffusion of YouTube videos showed that the spreading patterns vary by topics: topical content such as news, political commentary, and opinion spread on the order of hours to days (similar to flash floods), while non-topical content such as music and entertainment videos spread over several months (similar to ripples). As a result, old music videos can get rediscovered among bloggers even a year after upload.

Our finding about the extreme popularity of media content indicates that blogs, as a social medium, encourage the interaction of the Internet users with media and content providers by forming interest groups in the World Wide Web. The effects of topics and clustering of blogs based on interests are more broadly investigated in [3]. Our finding about the interesting time differences in spreading across topics indicates that such information sharing groups form and disappear dynamically over different time scales, similar to findings on micro-blogging site Twitter [11] (perhaps depending on the urgency of the information and human response times [20]). We leave the question of investigating the detailed mechanism behind the complex diffusion process across topics as future work. In conclusion, our work demonstrates that blogs, coupled with media sites, act as channels for distributing content, where users can generate content, discuss it in blogs, and pass it around in different forms such as web links, web feeds, and tweets.

Findings in this paper open up new research directions and help us better understand the new journalistic conventions. These findings are also essential for media

blogs and advertisers on understanding the dynamic nature of content spreading and the patterns of propagation of news and content in blogs and other social media comparatively. In the future, we would like to determine the diffusion patterns of other types of content (e.g., photos, rumors, conventions) [10] as well as the specific roles users take in disseminating content (e.g., leaders, spammers) [15]. We would also like to investigate the impact of local community structures on spreading (e.g., whether the ripple-like spreading over a long period of time is due to information having to cross "multiple" different communities) [41, 17]. We are also interested in investigating the changes in the set of words or text strings that describe a given object (e.g., HTML link of a photo or a video) over time. Such studies will help us extract meaningful information about dynamics of the opinion formation and popularity of linked content in the blogosphere.

References

1. Adamic, L.A., Glance, N.: The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In: Proceedings of the ACM SIGKDD International Workshop on Link Discovery (2005)
2. Adar, E., Adamic, L.A.: Tracking Information Epidemics in Blogspace. In: Proceedings of the ACM International Conference on Web Intelligence (2005)
3. Agarwal, N., Galan, M., Liu, H., Subramanya, S.: Clustering of blog sites using collective wisdom. In: A. Abraham, A.E. Hassanien, V. Snel (eds.) Computational Social Network Analysis, Computer Communications and Networks, pp. 107–134. Springer London (2010)
4. Bhagat, S., Cormode, G., Muthukrishnan, S., Rozenbaum, I., Xue, H.: No Blog is an Island – Analyzing Connections Across Information Networks. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (2007)
5. Blogcount: Wandering Through the Weblog Cemetery, `http://www.dijest.com/bc/`, Internet Draft (2003)
6. Blood, R.: Weblogs: a History and Perspective, `http://www.rebeccablood.net/essays/weblog_history.html`, Internet Draft (2000)
7. Bollen, J., Mao, H., Zeng, X.J.: Twitter Mood Predicts the Stock Market. Elsevier Journal of Computational Science pp. 1–8 (2011)
8. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph Structure in the Web. Computer Networks **33**(1) (2000)
9. Bross, J., Richly, K., Kohnen, M., Meinel, C.: Identifying the Top-dogs of the Blogosphere. Springer Social Network Analysis and Mining Journal pp. 1–15 (2011)
10. Centola, D., Macy, M.: Complex Contagions and the Weakness of Long Ties. American Journal of Sociology **113**, 702–734 (2007)
11. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring User Influence in Twitter: The Million Follower Fallacy. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (2010)
12. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.Y., Moon, S.: I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In: Proceedings of the ACM Internet Measurement Conference (2007)
13. Cha, M., Mislove, A., Gummadi, K.P.: A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. In: Proceedings of the International World Wide Web Conference (2009)
14. Etling, B., Kelly, J., Faris, R., Palfrey, J.: Mapping the Arabic Blogosphere: Politics and Dissent Online. New Media & Society (2010)
15. Fazeen, M., Dantu, R., Guturu, P.: Identification of Leaders, Lurkers, Associates and Spammers in a Social Network: Context-dependent and Context-independent Approaches. Springer Social Network Analysis and Mining Journal pp. 241–254 (2010)
16. Gibbons: Algorithmic Graph Theory. Cambridge University Press (1985)
17. Granovetter, M.: The Strength of Weak Ties. American Journal of Sociology (1973)
18. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information Diffusion Through Blogspace. In: Proceedings of the International World Wide Web Conference (2004)

19. ICWSM Spinn3r Blog Data. http://www.icwsm.org/2009/data/
20. Johansen, A.: Probing Human Response Times. Physica A **388** (2004)
21. Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: On the Bursty Evolution of Blogspace. In: Proceedings of the International World Wide Web Conference (2003)
22. Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: Structure and Evolution of Blogspace. Communications of the ACM (2004)
23. Lento, T., Welser, H.T., Gu, L., Smith, M.: The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop. In: Proceedings of the Annual Workshop on the Weblogging Ecosystem (2006)
24. Lento, T., Welser, H.T., Gu, L., Smith, M.: The Ties That Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System. In: Proceedings of the Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2006)
25. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-Tracking and the Dynamics of the News Cycle. In: Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (2009)
26. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective Outbreak Detection in Networks. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2007)
27. Macskassy, S.: Contextual Linking Behavior of Bloggers: Leveraging Text Mining to Enable Topic-Based Analysis. Springer Social Network Analysis and Mining Journal pp. 1–21 (2011)
28. Matheson, D.: Weblogs and the Epistemology of the News: Some Trends in Online Journalism. New Media & Society (2004)
29. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and Analysis of Online Social Networks. In: Proceedings of the ACM Internet Measurement Conference (2007)
30. PEJ: New media, old media, *Pew Research Journalism. http://www.journalism.org/analysis_report/twitter* (2010)
31. Rogers, E.M.: Diffusion of Innovations. Free Press (2003)
32. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In: Proceedings of the International World Wide Web Conference (2010)
33. Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: TwitterStand: News in Tweets. In: Proceedings of the ACM International Conference on Advances in Geographic Information Systems (2009)
34. Scott, J.: Social Network Analysis: A Handbook (2nd Edition). Sage Publications (2000)
35. Shamma, D.A., Kennedy, L., Churchill, E.: Summarizing Media Through Short-Messaging Services. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work (2010)
36. Shi, X., Tseng, B., Adamic, L.A.: Looking at the Blogosphere Topology through Different Lenses. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (2007)
37. Shirky, C.: Power Laws, Weblogs, and Inequality. Networks, Economics, and Culture (2003). Aula, Helsinki, Finland
38. Solis, B.: The Information Divide Between Traditional and New Media, http://tinyurl.com/ya8etcf, Internet Draft (2010)
39. Spinn3r: Blog Dataset. In: International AAAI Conference on Weblogs and Social Media (2009)
40. Technorati's State of the Blogosphere, 2010. http://technorati.com/state-of-the-blogosphere/
41. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press (1994)
42. Yardi, S., danah boyd: Tweeting from the Town Square: Measuring Geographic Local Networks. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (2010)
43. Zhou, R., Khemmarat, S., Gao, L.: The Impact of YouTube Recommendation System on Video Views. In: Proceedings of the ACM Internet Measurement Conference (2010)
44. Zhou, X.: The Political Blogosphere in China: A Content Analysis of the Blogs Regarding the Dismissal of Shanghai Leader Chen Liangyu. New Media & Society (2009)