Flash Floods and Ripples: The Spread of Media Content through the Blogosphere

Meeyoung Cha **Juan Antonio Navarro Pérez** **Hamed Haddadi**
Max Planck Institute for Software Systems (MPI-SWS)
Kaiserslautern/Saarbrücken, Germany

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 also disseminate new content and novel ideas to communities of interest. But how does content spread across these communities, what kinds of content spreads, and at what rate?

This paper presents an analysis of the network structure and the spreading patterns of media content in the blogosphere. Based on 8.7 million posts in 1.1 million blogs across 15 major blog hosting sites, we show that the network structure of blogs is different from that of other online social networks: most links are unidirectional and the network is sparse. We also find that user generated content, often in the form of videos or photos, is the most common type of content shared in blogs. Focusing on the distribution of 10,000 popularly linked YouTube videos in blogs, we demonstrate how propagation patterns depend on content category. For example, the latest political news video immediately catches the attention of bloggers and fades away after a week, while a music video propagates slowly over a long period of time.

Introduction

Blogging has become a significant part of today's Internet culture. Throughout the blogosphere, millions of people share their thoughts with the world and present their viewpoints about diverse topics like technology, music, politics, and travel. Blogs are an important social media platform, where bloggers connect to each other to share and disseminate ideas, norms, and content. Blogs can be connected explicitly (e.g., listed as friend blogs in a 'blogroll') or implicitly (e.g., inter-connected by comments or links, by shared interests, or by use of the same tags).

In this paper we study the trends in the use of blogs as a social medium. Using the HTML links embedded in blog posts from a large data set of blog feeds, we extract the implicit social relationships between blogs and construct the blog graph (Kumar et al. 2003). This allows us to examine 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.

Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

This paper is based on the Spinn3r data set (ICWSM 2009), which 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 for the study. 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, showing 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.

Finally, we study the content spreading pattern in conjunction with the blog graph. We identify 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.

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. Finally, we summarize the results and conclude.

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 (Kumar et al. 2003). They called the graph defined by links between blogs the *blog graph*. Shi et al. compared the structure of the blog graph, using multiple snapshots in time, against those of the web and social networks (Shi, Tseng, and Adamic 2007).

A rich set of related work focused on the interplay between the blogosphere structure and information dissemination. Gruhl et al. studied the diffusion of information in the blogosphere based on the use of keywords in blog posts (Gruhl et al. 2004). Adar and Adamic used the explicit use of HTML links between blogs to track the flow of information (Adar and Adamic 2005). More recently, Leskovec et al. developed algorithms to identify blogs which give the most up to date information on stories that propagate in the blogosphere (Leskovec et al. 2007).

A few studies focused on the use of blogs as a social medium. Bhagat et al. studied the demographics of multiple blog domains and characterized the interaction between blogs and the web (Bhagat et al. 2007). Adamic and Glance measured the interplay between the liberal and conservative political blogs during the 2004 U.S. election (Adamic and Glance 2005).

Compared to the work above, this paper focuses on two unstudied aspects in the blogosphere. First, we study the link structure and the content sharing trends across *multiple* domains and language groups. Second, we examine the posting of YouTube video links to determine the topics of popular videos and their spreading patterns in the blog graph.

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.

We first define the terminology. There are a number of *blog hosting sites* which allow an individual or a community 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.

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 site URL, post URL, language (identified by Spinn3r), and 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 posts, 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.

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. For this, we sorted the names of the individual 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. Our heuristic is based on the assumption that popular blog domains likely publish many more web feeds, 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.

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

Table 1: Summary of data set from 15 blog domains

Blog domains considered	Number of blogs	Number of posts	Posts per blog average	Posts per blog median	Dominant language
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

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.

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. 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.

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, 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 post counts from the less active half of all blogs. We see that 20% of the most active blogs account for 70% of all posts,

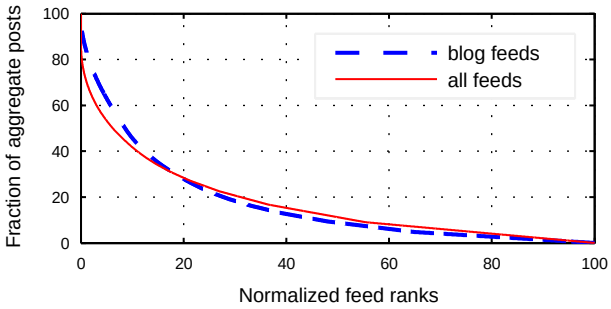

Figure 1: Testing Pareto principle in blog post behavior

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 five categories: (*i*) content aggregators or recommenders, that recommend other blogs or re-post content from other feeds; (*ii*) community blogs, where multiple people of a special interest group produced content; (*iii*) microblogging, where posts are typically brief text updates compared to lengthy journals; and (*iv*) spam blogs or splogs.

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

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. 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?

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.

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.

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 2 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 (Broder et al. 2000): 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 (Shi, Tseng, and Adamic 2007). 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.

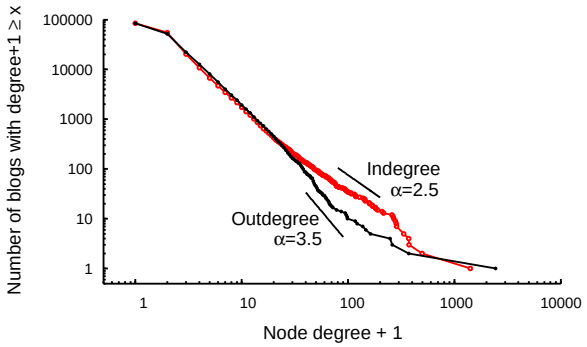

Figure 2: Degree distribution of the blog graph

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 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% (Shi, Tseng, and Adamic 2007). 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 (Cha, Mislove, and Gummadi 2009).

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 computes 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. The density value of a node is typically 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 (Lento et al. 2006). 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 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 3 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 (Mislove et al. 2007) for a randomly selected sample of 10,000 nodes. The median difference between the density values of blogs and nodes in a tree is 0.0023, while the median difference between nodes in the Facebook network and in a 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.[1]

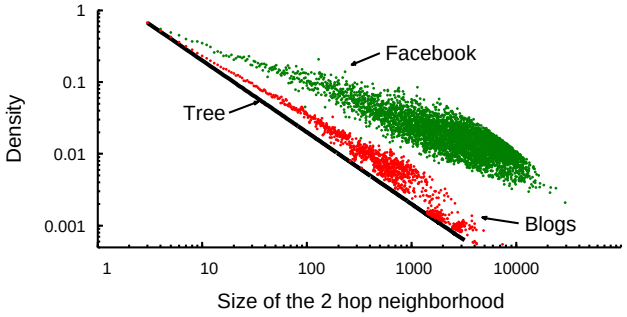

Figure 3: Comparing density of the blog graph with others

An alternative metric to determine the "treeness" of a graph is the so called *circuit rank* (Gibbons 1985). 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.

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.

[1]While the plot diverges for large x values, the impact of this difference is not significant since the y axis is in log scale.

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.

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.

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.

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. 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).

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 10 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 miniblogging messages from twitter.com are the most popular among livejournal users.

Table 2: Top 10 linked websites

Rank	Web link domain	# posts
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

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. 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 "homemade" content. These findings are consistent with the ones reported on earlier studies on web content in blogs (Bhagat et al. 2007).

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. Thus, we focus on links that point to YouTube videos and characterize their content sharing patterns among blogs.

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 4. 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 from youtube.com and obtained information about the uploader, view counts, tags, category, and duration. 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. In the remainder of this section, we present analyses of these 10,000 videos.

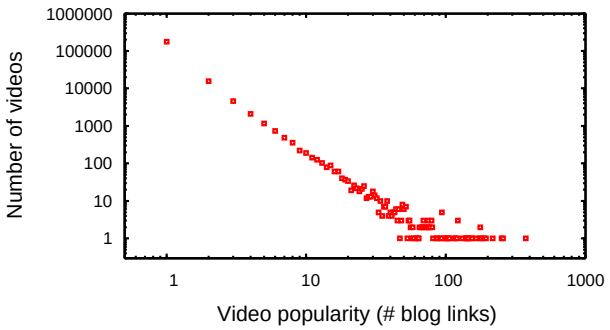

Figure 4: Popularity of YouTube videos in the blogosphere

Categories of the linked videos We examine which video categories are popular among blogs. Table 3 displays the top 10 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	% of videos	% 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 uploaders in the category of news and politics "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. 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[2] and quickly removes

[2] `http://www.youtube.com/t/community_`

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.

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.

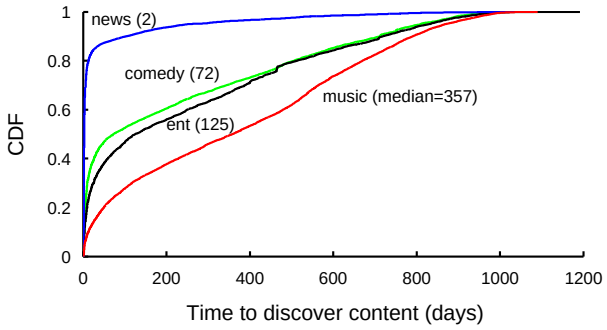

Figure 5: Age distribution of videos in the blogosphere

Figure 5 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 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.

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 6 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.

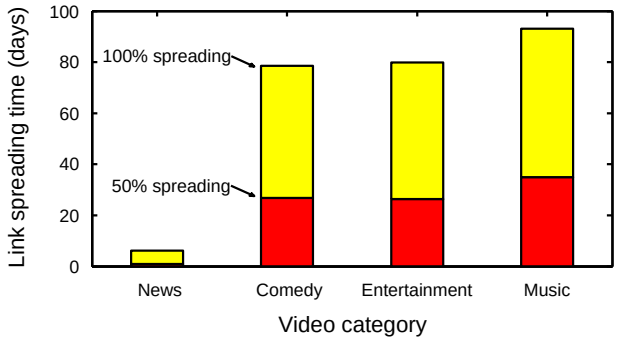

Figure 6: Time lag in the spread of videos in the blogosphere

The bar plot in Figure 6 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.

Content spreading over the blog graph

Finally we analyze the content spreading patterns along the blog graph. While bloggers can independently link to the same YouTube video, we are interested in scenarios in which information about YouTube videos propagated in the blogosphere. We assume that a YouTube video can spread in the blog graph if the following two conditions are met: (*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 7. 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 edges indicates the direction of information flow. Edges that fail to meet the time

guidelines

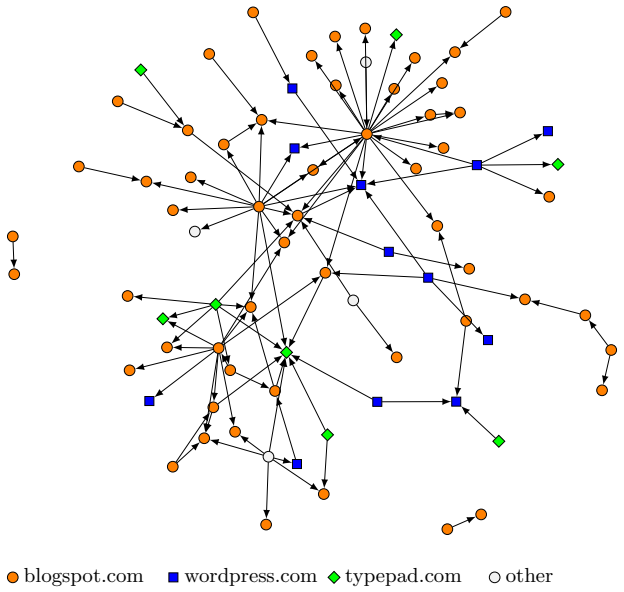

Figure 7: Diffusion of a political YouTube video in the blog graph. Node styles denote different blog domains.

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 clearly demonstrates that a large-scale diffusion of information can occur along the links in the blog graph at a rapid rate and across domains.

Concluding Remarks

In this paper we conducted an analysis of two months worth of web feeds from 15 popular blog hosting sites on the Internet. 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. Such characteristics clearly differentiate blogs from other social networks.

Content sharing is common in the blogosphere. In particular, user generated videos and photos are the most often linked content. Our study of the diffusion of YouTube videos showed that the spreading patterns vary by video category: news videos spread on the order of hours to days, while music videos spread over several months. As a result, old music videos can get rediscovered among bloggers even a year after upload. These findings indicate that blogs, as a social medium, encourage the interaction of the Internet users with media and content providers by forming communities of interest in the World Wide Web. 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. In the future, we would like to determine the diffusion patterns of other types of content (e.g., photos), the impact of community structure on spreading, and the set of words or text strings in the post 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

Adamic, L. A., and Glance, N. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In *ACM SIGKDD Intl. Workshop on Link Discovery*.

Adar, E., and Adamic, L. A. 2005. Tracking Information Epidemics in Blogspace. In *ACM Intl. Conf. on Web Intelligence*.

Bhagat, S.; Cormode, G.; Muthukrishnan, S.; Rozenbaum, I.; and Xue, H. 2007. No Blog is an Island – Analyzing Connections Across Information Networks. In *ICWSM*.

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; and Wiener, J. 2000. Graph Structure in the Web. *Computer Networks* 33(1).

Cha, M.; Mislove, A.; and Gummadi, K. P. 2009. A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. In *Proc. of the WWW*.

Gibbons. 1985. *Algorithmic Graph Theory*. Cambridge University Press.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information Diffusion Through Blogspace. In *Proc. of WWW*.

ICWSM. 2009. ICWSM 2009 Spinn3r Dataset. In *Proc. of the 3rd Intl. Conf. on Weblogs and Social Medi*.

Kumar, R.; Novak, J.; Raghavan, P.; and Tomkins, A. 2003. On the Bursty Evolution of Blogspace. In *Proc. of WWW*.

Lento, T.; Welser, H. T.; Gu, L.; and Smith, M. 2006. The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop. In *Proc. of the 3rd Annual Workshop on the Weblogging Ecosystem*.

Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; and Glance, N. 2007. Cost-effective Outbreak Detection in Networks. In *ACM SIGKDD*.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and Analysis of Online Social Networks. In *ACM IMC*.

Shi, X.; Tseng, B.; and Adamic, L. A. 2007. Looking at the Blogosphere Topology through Different Lenses. In *ICWSM*.