Comparing Information Diffusion Structure in Weblogs and Microblogs

Jiang Yang

School of Information
University of Michigan
1075 Beal Ave. Ann Arbor, MI 49109, U.S.
yangjian@umich.edu

Scott Counts

Microsoft Research
One Microsoft Way
Redmond, WA 98052 U.S.
counts@microsoft.com

Abstract

To better understand and characterize the emerging social medium of microblogging we conducted a comparison between Twitter and a weblog network for their respective information diffusion structures. We found systematic differences between the two social media in their contribution, navigation, and interactive structural patterns. Findings revealed the unique role and characteristics of microblogs in the social media design space. Implications are discussed.

Introduction

Microblogging, and Twitter in particular, has become an extremely fashionable form of social media over the past year or so. Similar to weblogs, people post content and share information through following networks. Compared to blogs, Twitter encourages fast updating by limiting post size, restricting the content format to text, and by supporting easy mobile updating. These design differences potentially are creating new ways for people to accumulate and share information.

In terms of how network structure might impact the flow of information in Twitter, earlier analysis (Java et al. 2007) found a high degree of reciprocity in the following network, although more recent work (Huberman et al., 2009) compared the following, followed, and friendship networks and concluded that users' actual interactions are hidden from the "declared" set of friendships seen in following relationships. We update and build on this work, with a more current analysis of Twitter that uses the actual social interaction network rather than the declared network of followers to characterize important aspects of information diffusion in Twitter. To do so, we use blogging as a reference because it is the most established and similar form of social media, and because blogs have been comprehensively studied on a number of related topics like participation patterns (Lento, 2006; Liben-Nowell et al., 2005), network dynamics (Adamic & Glance, 2005; Kumar et al., 2004, 2006), and information diffusion (Gruhl et al., 2004; Leskovec, 2007).

Analyses

Datasets

Our Twitter data source (TW) is one month of the Twitter public timeline, crawled daily through the Twitter API from July 8th 2009 to August 8th 2009. This crawl was augmented with results of a query for tweets that contained the string "http://". Our dataset contains 3,243,437 unique users and 22,241,221 posts[1]. The weblog data (WB) is a set of 59,048 blogs crawled through popular public blog containers (e.g., blogspot.com). There are 342,723 total posts crawled over a period of 5 months.

Contribution Patterns

Distribution of Contribution. Social media are well known for scale-free distribution in content contribution and consumption. We compared TW and WB for their respective distributions of user contribution (number of posts during a month). As shown in Figure 1, they both present mild skewness (TW: α= -1.57, R^2=0.83) although WB is less consistent in the trend and thus results in a poor fit. In general, in terms of quantity of posts, TW users contribute more than WB users by one order of magnitude. We see a bump at around 20 posts for WB, which is a quirk of our crawler and can be ignored, though does seem to be two different stages for the WB distribution, a flatter stage up to about 20 posts and a steeper stage after 20 posts.

[1] This same dataset was also used in entirely independent and non-overlapping analyses in another paper currently under review.

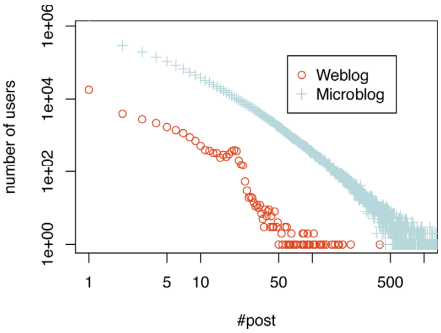


Figure 1: Distribution of contribution

Speed of Posting. We would expect that TW has a higher frequency of both contribution and consumption, since it encourages short posts and easy mobile posting. In addition to contribution in terms of numbers, we looked at how fast users can update new content, which potentially would be important for understanding the information-generating pattern. We used the minimum time interval between any two sequential posts of a user to quantify the capability of people to be fast in posting. Figure 2 shows the comparison, with users grouped by the total number of posts they had during a month.

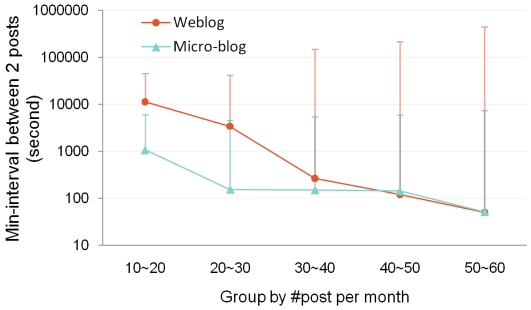


Figure 2: Min-Interval between 2 posts

In general, Twitter users had smaller minimum intervals between posts, especially when posting fewer posts (less than 30 posts per month), where the difference can be as high as more than one order of magnitude. For example, for the group who had 10 to 20 posts, the minimum interval is on average 3 hours for WB and 18 minutes for TW. However, this difference diminishes when people post many times, implying that posting speed for high volume posters reaches a common limit regardless of media type.

Navigation Patterns. We consider any online information medium to play some part of either (or both) of two roles: providing information content directly or navigating people to other sources. For a familiar example, Google, the largest portal of the Web, mostly serves a navigating function. Because a link provides a reference and path to another information source, we analyzed link patterns to quantify this navigating property of the two media.

There are two types of links in TW: explicit URLs (normally in a shortened form) and mentions (citing another user through the @username convention). URLs are also in WB but mentions happen only in TW. Row 1 of Table 1 shows the comparison of the link distribution. Not surprisingly, the rate of posts containing a link in WB is higher, as WB has more space for adding links. However the rate in TW is also striking as almost one quarter of posts had an explicit link and more than one third can potentially route to other users via a mention. This implies that in addition to considering TW as a medium of self-expression and conversation, its capacity for providing reference and navigation on the Web is important.

	WB	TW	Mentions
% posts that contain links	59.9%	24.7%[2]	35.2%
% inbound referring	85.96%	2.2%	35.2%

Table 1: link statistics for two media

Next we compared the destinations that these links point to (Table 1, row 2; and also Figure 3). The portion of links pointing to one another within the medium implies the degree to which a medium is self-sustainable by providing content and navigating within the medium. TW has very few tweets linking to other tweets (2.2%), though mentions (35%) serve this purpose to some extent. In WB, the majority of links actually point to other posts in the WB network. This comparison suggests very different properties of the two media: WB leans toward a function of internal blog content consumption while TW URLs are largely one-directional outwards.

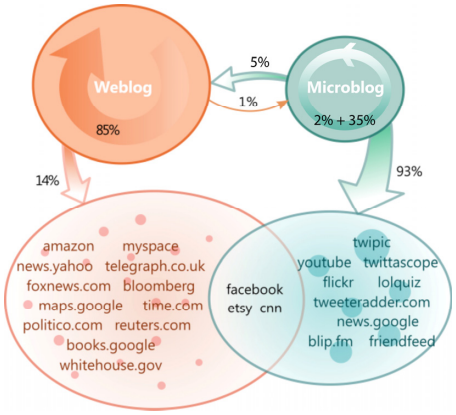


Figure 3: Web of navigation links

We also examined the difference between the sets of destination sites from the two media. Since the two sets are not on the same scale in terms of number of outbound links, we could only compare the distribution and qualitative differences. We count the frequency of each destination site and each blog only counts for one vote for each site. First we found that WB refers to a more diverse and scattered set of destination sites. As illustrated in Figure 3 (showing a sample of the top 15 destination sites), WB directs to many different websites and each of them accounts for only a very small portion of the overall traffic. TW also presents a much higher concentration. For example, the 30 most frequently referred websites by TW comprise 32.18% of total links, while this ratio for WB is

[2] Because our main dataset included additional tweets that matched our "http://" query, we used a different dataset just for this metric. This second dataset consisted of 5 months of the Twitter public timeline (1/1/09 – 6/1/09) sampled every 5 minutes.

only 6.47%. We use the relative dot size and number of sites in Figure 3 to visualize this difference. Table 2 lists the 15 websites most frequently referred by each medium. We see that for WB, each counts for one percent or less of the total number of links. TW is more concentrated, with social networking and media sites like YouTube dominating, including many sites derived from TW itself (e.g, twitpic).

WB		TW	
amazon.com	0.0124	twitpic.com	0.0822
facebook.com	0.0055	www.youtube.com	0.0437
myspace.com	0.0036	twittascope.com	0.0255
news.yahoo.com	0.0031	tweeteradder.com	0.0188
cnn.com	0.0030	NeedFollowers.com	0.0165
telegraph.co.uk	0.0028	lolquiz.com	0.0113
foxnews.com	0.0026	news.google.com	0.0108
etsy.com	0.0021	myloc.me	0.0108
timesonline.co.uk	0.0021	blip.fm	0.0094
maps.google.com	0.0018	www.facebook.com	0.0089
reuters.com	0.0018	www.tweeterfollow.com	0.0077
bloomberg.com	0.0017	www.plurk.com	0.0070
dailymail.co.uk	0.0016	www.ustream.tv	0.0059
politico.com	0.0015	www.flickr.com	0.0052
time.com	0.0015	mypict.me	0.0050

Table 2: top 15 frequent destination sites

Characterizing Network Structure. In these two social media, information dynamics is mainly realized through the network of social interactions. Through those interactions, people spread information, exert influence, and construct social cognition. For WB, hyperlink citations have been used to model the network of social interactions (Gruhl et al., 2004; Leskovec, 2007). In TW, there are two primary interrelationships among users: following and mentioning. Similar to citation in a blog, mentioning is an explicit action that refers or attributes to another user. Therefore, to be as comparable as possible to links in blogs, we constructed a social interaction network for TW based on mentions.

We first compared the global structural characteristics between the WB and TW networks. Table 3 shows that the two social media present distinct characteristics in their global interactive structures (note that WB results based on our sample are consistent with previous reports (e.g., Shi et al., 2007). In general, the WB network is much denser than TW network despite TW having a higher percent of

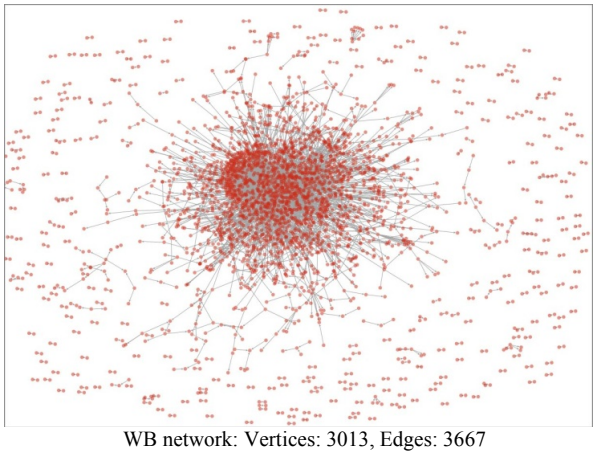
Measures	WB	TW
# nodes	59,048	2,896,784
# edges	198,445	4,557,124
# mutual edges	4613(2.32%)	318,788(7.00%)
Clustering Coefficient	0.0076	0.0031
SCC	14.64%	13.64%
IN	18.23%	14.96%
OUT	12.99%	8.03%
TUBES	1.26%	6.39%
TENDRILS	12.02%	0
OTHERS	40.86%	56.76%

Table 3. Comparing global structures

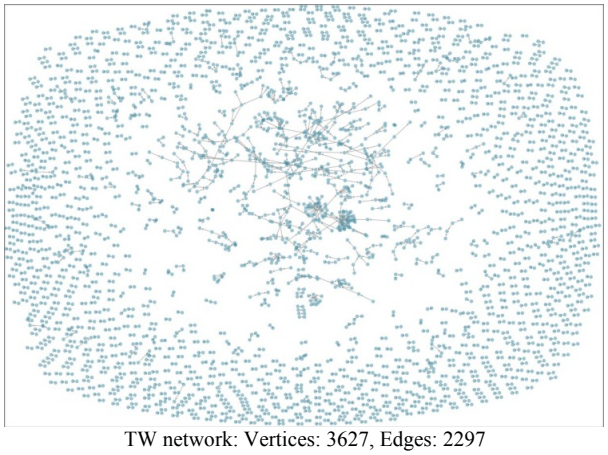
mutual edges, which indicates a slightly higher "reciprocal" nature of TW. We see that the TW network presents lower global connectedness than the WB network, as shown in the Bowtie analysis (SCC: Strongly Connected Component and IN/OUT sets; Broder, et al., 2000).

Interestingly, even with sparse edges and a huge number of nodes, the TW network presents a clustering coefficient of the same magnitude as that of the WB network. The densification law prevalent in many networks suggests that the number of edges grows superlinearly with the number of nodes over time: $e(t) \propto n(t)^{\alpha}$ (Leskovec, 2005) and thus all these measures are sensitive to the size of network. To accommodate this size sensitivity, we compared the WB network with a down-sampled TW network and found that the TW network presents an even higher CC than the WB network at the same scale. This implies that TW is less coherent globally, but likely to form tight clusters locally. Figure 4 visualizes the two networks and we can clearly see the structural difference. The WB network is more coherent and forms a central core while there is little core in the TW network and the nodes tend to cluster in decentralized small structures.

Next we used motif analysis to assess the micro-structural signature of the two networks. We employed FANMOD, a convenient motif detector tool to uncover the significant triad motifs in the network (Milo, et al., 2002). This tool generates a certain number of random networks with the same number of nodes and edges as the original network and tests the difference in terms of frequency of each possible triad motif structure between the original



WB network: Vertices: 3013, Edges: 3667



TW network: Vertices: 3627, Edges: 2297

Figure 4. Visualizing subset of Weblog link-network and Micro-blog mentioning network, used Microsoft C#UNG

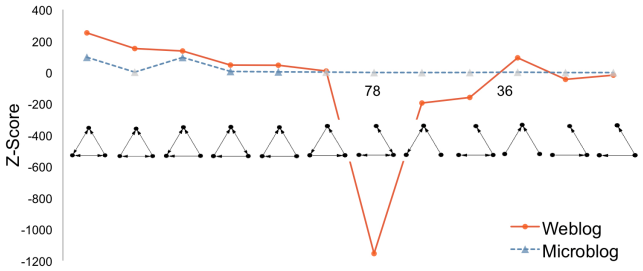


Figure 5. Comparing triad motif (non-sig in grey)

network and the average of all random networks.

Figure 5 compares the degree to which the two networks deviate from random networks. In general, both networks tend to form tightly connected structures (closed triangle), which is consistent with many previous observations on real social networks (Watts & Strogatz, 1998). The WB network shows stronger deviation from random networks. For example, it is extremely significant on the full bi-directed triangle △(which is not shown in the figure because the value is too high to be comparable with others). The WB network is also extremely significant in the negative direction on structure ID=78, where one blog has bi-directed links with the other two, but there is no link between them. Together these results suggest a strong tendency to form full triangle relationships, when there are already two bi-directional pairs there.

Furthermore, triad ID=36 is a very popular structure in the random network (the average frequency was 68.91%). However the WB network has a significantly, if only slightly, higher frequency (68.98%) while the TW is not different from the random network. Thus this motif structure, likely two blogs referring to a third, is dominant in the WB network over the TW network.

Conclusion & Future Work

This work compared several aspects of the respective information structures of weblogs and Twitter. We found that the two forms of social media systematically differ in the contribution pattern, how they navigate on the Web, and the network of social interactions. We found that TW users generally maintain a higher frequency in posting and both WB and TW present superlinear distribution of contribution across users. The WB network is more self-sustained with a large percentage of links pointing to other blog sites. In contrast, most URLs in TW are outbound.

The comparison of the social interaction networks revealed structural differences between WB and TW. The WB network is more coherent globally while the TW network is more decentralized and connected locally. This indicates a flatter social structure in Twitter and a relative inapplicability of some social network analysis algorithms like PageRank. The global disconnectedness also suggests a limited efficiency in larger scale information diffusion. In future work, we would like to extend the comparison to the dynamics of information diffusion such as whether the two forms of social media show different preferences or efficiencies in spreading different types of information.

Acknowledgements

The authors would like to thank Steve Ickman, Paul Johns, Matt Hurst, and Scott Golder who helped with data access and provided discussion insight.

References

Adamic, L. A., & Glance, N. (2005). *The political blogosphere and the 2004 U.S. election: divided they blog.* Paper presented at the Proceedings of the 3rd international workshop on Link discovery, Chicago, Illinois.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph structure in the Web. *Computer Networks, 33*(1), 309-320.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). *Information diffusion through blogspace*. Proceedings of the 13th World Wide Web.

Huberman, B., Romero, D. M., & Wu, F. (2009). Social networks that matter: Twitter under the microscope. *First Monday, 14*.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). *Why We Twitter: Understanding Microblogging Usage and Communities.* Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2004). Structure and evolution of blogspace. *Communications of the ACM, 47*(12), 35 - 39.

Kumar, R., Novak, J., & Tomkins, A. (2006). *Structure and evolution of online social networks.* SIGKDD.

Lento, T., Welser, H. T., Gu, L., & Smith, M. (2006). *The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System*. Paper presented at the WWW Third Annual Workshop on the Weblogging Ecosystem.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). *Graphs over time: Densification law, shrinking diameters and possible explanations.* Paper presented at the SIGKDD.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., Vanbriesen, J., & Glance, N. (2007). *Cost-effective outbreak detection in networks.* 13th KDD.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic Routing in Social Networks. *Proceedings of the National Academy of Sciences, 103*(33), 11623-11628.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science, 298*(5594), 824-827.

Watts, D. J., & Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440–442.