Predicting the Speed, Scale, and Range of Information Diffusion in Twitter

Jiang Yang
School of Information
University of Michigan
1075 Beal Ave. Ann Arbor, MI 48109, U.S.
yangjian@umich.edu

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

Abstract

We present results of network analyses of information diffusion on Twitter, via users' ongoing social interactions as denoted by "@username" mentions. Incorporating survival analysis, we constructed a novel model to capture the three major properties of information diffusion: speed, scale, and range. On the whole, we find that some properties of the tweets themselves predict greater information propagation but that properties of the users, the rate with which a user is mentioned historically in particular, are equal or stronger predictors. Implications for end users and system designers are discussed.

Introduction

The practice of microblogging, characterized by short status updates posted frequently to social media sites, has gained significant usage and attention in recent months. Twitter, with its sole purpose of sharing short statuses to a largely public audience, arguably is the best known example of microblogging. With significant recent growth and attention, Twitter may make microblogging on par with social networking and blogging as a form of social media. With this rise in usage and popular media attention, it is important to understand exactly how Twitter users are interacting with one another and how information propagates through Twitter.

Large scale network analyses of Twitter have been relatively scarce. Java et al. (2007) examined the follower network and reported high degree correlation and reciprocity in the follower network and revealed there is great variety in users' intentions and usages on Twitter. Huberman, et al., (2009) examined tweeting behavior in relation to the following networks of Twitter users and what they refer to as the friend network. A friend is another user to which the user has directed two or more tweets (using the "@username" convention). Their results show that the number of tweets is more strongly related to the number of friends than the number of followers, suggesting that users' actual interactions reflect a different network than the following relationships suggest. Thus the interaction network, rather than the follower network, is preferable for network analyses of Twitter.

While studies of Twitter have been sparse, similar analyses are much more prevalent in the blog arena. Most relevant to the present work are network structure and information diffusion analyses. For example, links have been used to detect communities (Adamic & Glance, 2005; Tseng, et al., 2005), and a variety of factors such as geography (Liben-Nowell, et al., 2005), age and common interests (Kumar, et al., 2004), and existing friends (Backstrom, et al., 2006) have been correlated with link formation. In terms of information diffusion, the efforts include but are not limited to: identification of topics over time (Gabrilovich, et al., 2004; Havre, et al., 2000), tracking topic flows (Adar & Adamic, 2005; Adar, et al., 2004; Leskovec, et al., 2009), and modeling the dynamics of adoption cascades (Gruhl, et al., 2004; Leskovec, et al., 2007, 2009).

In summary, while network properties of Twitter have not been studied extensively, previous work including considerable work on blogging networks, suggests that the active interaction network is of higher value than the follower network, particularly with respect to analyses of information diffusion. We build on this by constructing interaction networks based on @username mentions to extract network structural properties and attributes of users

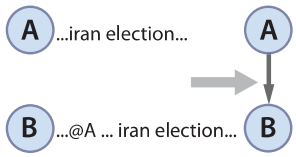

Figure 1: Topic-constrained diffusion link between two Users.

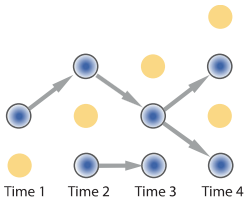

Figure 2: Building a diffusion network.

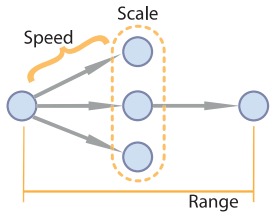

Figure 3: Three measures of local diffusion tree

and content that predict information propagation within these structures.

Analysis

Data and Methods

Data. Our primary data source is one month of the Twitter "spritzer" feed, a sample of the public timeline, crawled daily through the Twitter API from July 8th 2009 to August 8th 2009. Our crawler augments these data with results of an additional query of the standard Twitter search for the string "http://". Our dataset contains 3,243,437 unique users and 22,241,221 posts.

Method. We focused our analyses on mentioning, the practice of referring to another user in a tweet via the "@username" convention. In contrast to following, mentioning represents an active user interaction. In our definition, mentioning includes all uses of @ (e.g., retweet, reply). Given our focus on information diffusion, this inclusive definition captures most comprehensively the social links between users. We note here that retweeting has been shown to take different forms, such as "RT" and "via" (boyd, et al., 2010), but most if not all retweets still include "@username" and thus should be included in our analyses.

Diffusion Network. To measure how topics propagate through network structures in Twitter we constructed a diffusion network based on @username mentioning, with a constraint of topical similarity in the tweet. That is, we build a link from **A** to **B**, if **B** mentions **A** in her tweet that contains **topic C** that **A** had talked about earlier (Figure 1). Given the lack of explicit threading in Twitter, this is the optimal approximation of the path of person **A** diffusing information about topic **C**.

Figure 2 shows how we then built the diffusion network with timestamps. All posts that contain the topic keywords (e.g., "Iran election") are labeled with timestamps and track the diffusion links as defined above. Blue gradient colored nodes (outlined) are counted into the network while other (yellow) nodes are those who just mentioned the topic but without linking to any ancestor node.

Local dynamics: speed, scale, and range. We developed models for three dimensions of diffusion networks in Twitter (Figure 3): ***speed***, whether and when the first diffusion instance will take place; ***scale***, the number of affected instances at the first degree; and ***range***, how far the diffusion chain can continue on in depth.

Speed The most straightforward question when seeing a post about a particular topic, is how the followers would be influenced and retweet, reply, or otherwise mention the initial tweet in their own tweets about the same topic. This question involves two parts: whether one would mention at all and if so, when will this mention happen. Employing survival analysis, both questions can be addressed in a single model: we predict when a tweet containing a topic is likely to be mentioned by another tweet also containing the topic.

We then use the Cox proportional hazards regression model (Cox & Oakes, 1984) to quantify the degree to which a number of features of both users and tweets themselves predict the speed of diffusion to the first degree offspring. For instance, aspects of each individual author, such as their activity level in tweeting and mentioning and being mentioned may also predict diffusion speed. In terms of characteristics of tweets, we examined whether the tweet contains a link, whether it itself is a mention, and what we call stage: whether the tweet comes at an earlier or later stage in the topic lifespan. To simplify the stage variable, we divided tweets based on their timestamp into two sets: before and after 10 days following our earliest observation of the topic.

We ran regression analyses on these variables predicting whether and when a tweet produces its first offspring node over different topics. As an example topic see "Iran Election" in the third column of Table 1. We see that when the author is more active in posting (nPost) and has a higher rate of being mentioned (MentionedRate), the

Topic	Apollo	Iran Election	Google Voice	Harry Potter	Bing	Chrome OS	Swine Flu	Ice Age 3
nPost		1.0004**		1.0007***		1.0006***		
nMention		1.0006**		1.0006.	1.0013**	1.0004*		1.0178*
nMentioned	1.0020***		0.9987***	1.0027***	1.0007***	1.0001**	1.003***	
MentionedRate	1.3785***	1.1479***	2.4490***	1.0447***	1.1664***	1.0875***	1.091***	5.1330***
isMention		1.2077**		2.2106***				
haveLink			2.5876***	0.6944***	1.5730***	1.2895**	1.301**	
stage	0.1653***	0.3372***	2.2156**	0.3934***	0.6893***	0.6052***	1.131**	3.1194*
R^2(max possible)	0.028(0.473)	0.067 (0.975)	0.059 (0.777)	0.009(0.245)	0.016(0.597)	0.01(0.738)	0.016(0.588)	0.028(0.192)
Reporting exp(coef) with p-value. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1								

Table 1: Predicting whether & when a post will get mentioned by an offspring node about the same topic. Only significnt effects are shown. Values above 1.0 indicate a positive relationship between the predictor and speed of influence. Values below 1.0 indicate a negative relationship.

present tweet will gain offspring in a shorter time. When the post is a mention per se (isMention), it has a higher chance to continue the diffusion. Stage (when the tweet is tweeted) also counts for a significant effect. For this topic, posts in an earlier stage (Jul 8-17) are more likely to produce an offspring in a shorter time. Finally, whether the tweet contains links does not affect the ability to generate offspring nodes for this topic.

For almost all topics in Table 1, the author's rate of being mentioned by other people (MentionedRate) is an important predictor for whether and how fast her tweet on this topic would be mentioned. The time when the tweet is posted (indicated as stage) is also a frequent predictor. For many cases, earlier posts can be more effective in producing offspring (in the table, when coefficient<1). However, there are also opposite cases, such as with the Google Voice and Ice Age 3 topics for which tweets later in the observation period generated offspring tweets more rapidly. These results suggest that a topic might have a different diffusion efficiency at different time stages of its life cycle. That is, when information is diffused through the network, the speed and efficiency would vary over time versus being linear over time. To understand how information diffusion efficiency varies over time would be an interesting area for future exploration. Similarly, the presence of link(s) in a tweet may increase the likelihood of producing offspring nodes, but the direction of the effect is not stable, as it is positive for most topics, but negative for the Harry Potter topic. This implies an interaction between topic properties and tweet properties, suggesting a role for additional text analysis in future work.

Scale Next we turn to the question of scale: for each tweet, how many people mentioned the same topic as first degree child nodes in the diffusion network? Here each user is only counted once for their first post about a given topic. Because the majority of posts did not produce a child, we only predict based on tweets that had at least one child node. Further, we used the logarithm of those variables given significant skew from a normal distribution.

Table 2 presents regression results on our sample trending topics. R^2 of the regression is presented in the second to last row and the correlation coefficient between the predictor and log(nChild) is presented in each cell with significance codes. In general, these regressions yield much better prediction power than for the speed analysis. The activity level of the user and number of times she is mentioned are stable predictors and account for the majority of the variance. For example, the correlation coefficient between log(nChild) and log(Mentioned) is 0.63 for the Iran Election topic. In addition, including links in tweets often generates more child nodes.

Range As a final metric of local diffusion, we measure the range of influence as indicated by the number of hops in a diffusion chain. To do so, we trace a topic from a given start node to its second and third degree of offspring nodes, and so on. As shown in Figure 2, the length of the chain indicates how far the original node diffuses in depth.

First we investigated general patterns of these diffusion chains. For most of these topics, more than half the ancestor nodes fail to produce offspring of the first degree, and less than 30% continued to the second degree. After 5 hops away, for most topics, less than 5% of ancestor nodes still continue producing offspring. In addition, the various topics yielded significant differences in chain life (Survival Difference test, $p<0.0001$). Topics like "Iran Election" tend to have longer chain life than topics like "Apollo".

Similar to our analyses for speed and scale, we again examined aspects of users and tweets that may predict greater range of diffusion. Table 3 presents the predictors of the length of a topic chain within a diffusion network. Consistently, greater activity in posting and being mentioned are often predictors of longer diffusion hops across topics. Interestingly, we also see that a tweet being a mention itself, occurring in later stage, and containing link (except for Harry Porter) predicts longer chains.

Topic	Apollo	Iran Election	Google Voice	Harry Potter	Bing	Chrome OS	Swine Flu	Michael Jackson
Log(nPost)	0.1726**	0.1415***	0.2024***	0.0685.	0.2331***	0.2444***	0.1416**	0.1342**
Log(nMention)		0.2516***		0.0812**	0.1781***	0.1212***		0.0845.
Log(nMentioned)	0.4565***	0.6270***	0.4001***	0.2943***	0.4467***	0.5821***	0.3789***	0.3916***
MentionedRate	0.4071***	0.0941***	0.4701***	0.1371***	0.3862***	0.4271***	0.1835***	0.3092***
isMention	-0.1374*			0.0767**		-0.0620*		
haveLink		0.0654*	0.1837***	0.1634***	0.0920*	0.0576*		0.1128**
stage			0.1511**			-0.0570*		
R^2	0.3357	0.4192	0.3108	0.1567	0.251	0.4643	0.1966	0.219
Reporting correlation coefficient. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1								

Table 2: Predicting number of child nodes one can produce. Only significnt effects are shown.

Topic	Apollo	Iran Election	Google Voice	Harry Potter	Bing	Chrome OS	Swine Flu	Michael Jackson
nPost	0.9999.	0.9986***		0.9970***	0.9996***	0.9998*	0.9992***	0.9997*
nMention		0.9967***	1.0022*	1.0018***			1.0020**	
nMentioned			0.9945**		0.9991*	0.9952***	0.9984*	0.9964***
MentionedRate	0.6919***	0.7336***			0.8650***	0.7303***	0.8518***	0.9585.
isMention		0.7281***	0.5780*	0.6859***	0.5650***	0.8618*	0.6630**	0.6205***
haveLink			0.5118***	1.0765***	0.8420***	0.9052***	0.6743***	0.8897***
stage	0.9313*	0.6280***	0.1902***	0.5348***	0.3860***	0.3277***	0.6519***	0.3452***
R^2(max possible)	0.043(1)	0.083(1)	0.168(0.993)	0.040(1)	0.115(1)	0.140(1)	0.055(1)	0.185 (1)

Table 3: Predicting length of influence chain of ancestor nodes. Only significnt effects are shown.

Discussion

Our analyses investigated @username mentions in order to utilize the "hidden" network of actual user interactions in Twitter rather than the potentially very passive follower network. We constructed diffusion network by scoping it to specific topics to measure aspects of how the network impact information diffusion in Twitter. In particular, we focused on the properties of the network that predict the speed, scale, and range of information propagating through Twitter.

First, for speed (how quickly will a tweet produce an offspring tweet), the amount a user is mentioned is a good predictor of producing offspring rapidly, although across our eight sample topics, the regression equations predict only a small amount of variance. Interestingly, in some cases, tweets appearing later in our observation of a topic yielded offspring more quickly. This suggests that system designers are wise to not simply assume that the earliest tweet about a topic is the most important, but instead should continue to watch the topic for tweets with the greatest amount of influence.

In terms of scale (number of child nodes one can produce), again the amount a person is mentioned is the best predictor of producing more child nodes. In this case the correlation is quite strong, as high as .63 for the Iran Election topic. This analysis (see Table 2) revealed a few surprises in terms of variables not predicting the generation of greater numbers of child nodes. First, containing a link does tend to correlate positively with generating more children, but the correlations are not terribly strong. The same holds for whether the tweet is itself a mention. This suggests that looking exclusively at the properties of the tweet itself, while useful, is not necessarily the best strategy for predicting whether a tweet will generate offspring. Instead a combination of properties of the tweet and tweeter is suggested.

Finally, for range (number of hops in the diffusion network), a few predictors stand out. As we have seen consistently, the mentioned rate of the tweeter is a significant predictor of tweets traveling longer distances in the network. As with speed, tweets that came later in the observation often were more influential, in this case traveling further in the network. Again this suggests that for uses like surfacing tweets for search results or for various other analysis purposes, not simply searching for the first or even the earlier tweets on a topic, will help uncover the most influential content. We do see evidence for the inclusion of links in tweets reaching further across the network, and thus suggest easy queries (e.g., tweets with "http://") for end users and system designers looking for tweets that touch lots of users.

Taken together, we see a clear theme that the mention rate of the person tweeting is a strong predictor of all aspects of information diffusion through social networks in Twitter. Other attributes of the tweets themselves, such as whether it includes a link or comes at the early or late stages of a topic also are important, but based on our analysis we suggest utilizing these in conjunction with properties of the user for any type of network ranking algorithm.

References

Adamic, L. A., & Glance, N. (2005). *The political blogosphere and the 2004 U.S. election: divided they blog.* international workshop on Link discovery.

Adar, E., & Adamic, L. A. (2005). *Tracking Information Epidemics in Blogspace.* Web Intelligence.

Adar, E., Zhang, L., Adamic, L., & Lukose, R. (2004). *Implicit Structure and the Dynamics of Blogspace.* Workshop on the Weblogging Ecosystem.

Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). *Group Formation in Large Social Networks: Membership, Growth, and Evolution.* KDD.

boyd, d., Golder, S., & Lotan, G. (2010). *Tweet Tweet Retweet: Conversational Aspects of Retweeting on Twitter.* Proceedings of HICSS-43, Kauai, HI.

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. London: Chapman & Hall.

Gabrilovich, E., Susan Dumais, & Horvitz, E. (2004). *Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty.* WWW 2004, New York.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). *Information diffusion through blogspace*. WWW.

Havre, S., Hetzler, B., & Nowell, L. (2000). *ThemeRiver: Visualizing Theme Changes over Time.* Proceedings of the IEEE Symposium on Information Vizualization.

Honeycutt, C., & Herring, S. C. (2009). *Beyond microblogging: Conversation and collaboration via Twitter.* Proceedings of HICSS-42, Los Alamitos, CA.

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.* 9th WebKDD and 1st SNA-KDD workshop on Web mining and social network analysis.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2003). *On the bursty evolution of blogspace.* WWW.

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

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). *Meme tracking and the dynamics of the news cycle.* KDD.

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. *PNAS, 103*(33), 11623-11628.

Tseng, B. L., Tatemura, J., & Wu, Y. (2005). *Tomographic Clustering To Visualize Blog Communities as Mountain View.s* 2nd Workshop on the Weblogging Ecosystem.