Interaction Network Properties of Twitter

ABSTRACT

We present results of network analyses of Twitter that address two related issues: social reciprocity and information diffusion. Analyses of user interactions as denoted by "@username" mentions show that rather than being reciprocal, Twitter behavior is dominated by one-way social connections and that the majority of mentions flow through a small percentage of very active Twitter users. We constructed and analyzed an influence network of Twitter users for three 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.

Author Keywords

Twitter, information diffusion, social network analysis.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

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 status 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.

Given that interaction behaviors are better known in social and blogging networks, a useful starting point is to consider Twitter with respect to each of these more established forms of social media. Indeed Twitter has properties that would seem to place it somewhere between a social network and a blogging platform. For example, a user can follow another user, similar to friending another user in a social network. However, following is unidirectional: I can follow you without you following me. This is different than dominant social networks like Facebook and may change the interaction dynamics among users. As one example impact of unidirectional relationships, numerous celebrities have very large followings of other users who they themselves are not following. The celebrity phenomenon raises the additional point that following a person is not the same as actually interacting with that person. We may follow many people, but reply to or retweet only a few.

Further, especially compared to a social network, Twitter is showing signs of becoming more of a source for information dispersion than for maintaining social contacts. Write-ups such as the cover article in Time magazine (June 15, 2009) call out link and information sharing generally as the reason why Twitter is powerful. To the extent that Twitter is in fact made up of unidirectional user connections and has a focus on information sharing, it becomes more of an information source than a social network.

Arguably blogging, particularly with the popularity of professional blogs like TechCrunch and Gizmodo, also transitioned or at least expanded from a personal and social sharing medium to an information sharing medium. However, Twitter also differs from a blogging platform in important ways that may impact interaction dynamics. For example, Twitter is fairly impoverished with respect to ways users can interact. There is no commenting on or threading of tweets, nor is there a simple way to link to a specific tweet as one would link to a blog post.

Given our positioning of Twitter as somewhat of a hybrid of social networking and blogging, we examined its properties for spreading of information (loosely similar to present day blogging) within a network of interacting users (social networking). What are the characteristics of the interaction network in Twitter and how does information flow through that network? We break our analyses into two main sections. First, we examine patterns of ongoing social interaction such as whether Twitter is more conversion or broadcast. Second, we look at the speed, scale, and range of the propagation of information.

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

BACKGROUND

Large scale network analyses of Twitter have been relatively scarce. In one of the earliest studies, Java et al. [9] examined the follower network, including 1,348,543 posts and 829,053 user profiles crawled through the follower network sampled during 2 months in 2007. The study reported high degree correlation and reciprocity in the follower network and revealed there is great variety in users' intentions and usages on Twitter. Although this work is only two years old, Twitter has evolved considerably. As mentioned, Twitter has grown tremendously over that time period[1], which means the network structure and dynamics may have changed due to a broader user population. Also, from a data accessibility standpoint, Twitter has now made available a sample of the public data stream via API (that we used in the current work), which greatly reduces any possible sampling bias in the network due to the previously necessary method of crawling follower networks.

More recently, Huberman et al. [8] examine 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.

Of note, the Huberman et al. work also report a very high level of reciprocity in the friendship network, with fully 90% of friends reciprocating by being friends of the user. We investigate this notion of reciprocity in greater detail and as we will see below, our results paint a somewhat different picture at least with respect to Twitter looking more like a broadcast than conversation medium.

While studies of Twitter have been sparse, similar analyses are much more prevalent in the blog arena, where social interactions have been more comprehensively explored and design implications generated. Most relevant to the present work are network structure and information diffusion analyses. With respect to link structure in blogs, links have been used to detect communities [1, 16], and a variety of factors such as geography [15], age and common interests [11], and existing friends [4] have been correlated with link formation. Kumar et al. characterized the link-network structure of the blogosphere and examined the bursty pattern in the linking activity and how link distribution varies by time effect [10, 12, 11]. In [12] they highlight the importance of including time-based topic analyses that reflect not just bursts of content, but bursts of connections between users. Building on this notion, we analyze properties of information diffusion in Twitter that incorporate time. For example, given a topic, how does that topic spread over time within connected networks of users versus in Twitter as a whole?

Information diffusion studies of the blogosphere have tracked information flow through links over time to identify the factors and patterns crucial to diffusion efficiency. The efforts include but are not limited to: identification of topics over time [5, 7], tracking topic flows [2, 3, 13], and modeling the dynamics of adoption cascades [6, 13, 14]. In [6], Gruhl et al. discuss information cascades, or how information travels from one person to another, in blogspace. We spend considerable attention on this topic in Twitter. We examine specific properties of Twitter users, such as the amount they post, for their ability to predict the extent to which their content will spread in the network.

All network analyses require the identification of linkages between users. In blogs, Adar and Adamic [2] highlight the benefits of "via" links that contain clear references to topic sources when trying to track topic spread. The @username convention adopted by Twitter users is employed expressly for the purposes of content attribution and thus we utilize it in our analyses. That is, within a given topic (e.g., the subset of tweets with matching keywords), tweets with @username mentions are very strong indicators of information spreading across the network.

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 and influence networks based on @username mentions to extract network structural properties and attributes of users and content that predict information propagation within these structures.

ANALYSES

Data and Methods

Data

Our primary data source is one month of the Twitter 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[2]. One month is a snapshot of Twitter but should be sufficient for our analyses, as most social interactions and many topic propagation life spans are shorter than one month in Twitter (as we show below). Additionally, as mentioned, prior to Twitter releasing its public API, researchers could only

[1] 1382% from 02/08-02/09 according to Nielsen: http://blog.nielsen.com/nielsenwire/online_mobile/twitters-tweet-smell-of-success/

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

crawl follower networks and were likely to miss large numbers of users and posts, and thus our dataset provides a denser and more complete record of the Twitter user population and activities.

Method

In terms of methods, we utilized a variety of analysis techniques largely related to social network analysis. At the heart of these analyses are links between users. In Twitter there are two primary forms for inter-user links: following and what we call "mentioning." Following happens at a single point in time when one user forms a unidirectional link with another user. Behaviorally, other than the initial act of following a user, following is passive. We focus our analyses on mentioning, the practice of referring to another user in a tweet via the "@username" convention. This practice is used for a variety of purposes such as replying and "retweeting", but in all cases indicates some form of interaction relationship between the user tweeting and the user being mentioned. In contrast to following, mentioning is active and ongoing. Our definition of mentioning is nearly identical to that of friends in Huberman et al. [8], but requires only a single use of "@username" to create a link between users.

In order to investigate people's interaction dynamics, we tackle the question through two interlinked perspectives: patterns of dyadic interaction behaviors via mentioning (e.g., how bidirectional is the Twitter mention network), and properties of information diffusion within the interaction network. In particular, we innovatively employed survival analysis [3] to assess many temporal patterns in micro-interactions and the diffusion network. Thus, we extend previous diffusion studies with an additional dimension: time. For example, when we measure reciprocity in mentioning, we not only measure the overall probability of replying, but also the varying probability over time. In discovering many of these temporal patterns, we used Cox proportional hazards regression model to assess the prediction power of multiple factors. Based on its log-linear relationship assumption between the independent variables and the underlying hazard function, the significance for predictors is presented by the degree to which the coefficient differs from 1.0, with the sign of the deviation from 1.0 describing either positive or negative effects. Finally, as a general note about sample size and statistical significance, we recognize that some findings are significant partially due to the large sample size. Thus, when possible we report measures of effect size, such as correlation.

Interaction Behavior: Mentioning

Basic characteristics

As indicated, Twitter users can target a tweet to another users by using the "@username" convention. Any message containing "@username" will be directed to that user, as well as posted to the public timeline unless it is a direct message (predicated by "d") or the sender's tweets are protected. This convention gets used in a variety of ways, such as when retweeting (reposting another user's tweet) and mentioning another user in a tweet. In our analysis we take any use of @username to indicate a behavioral connection between the poster and that user. Arguably the different uses of @username indicate different types of connections (e.g., retweeting is not the same as replying to a tweet), though these distinctions are difficult to evaluate objectively, and thus simply including all uses of @username avoids issues with such distinctions. Also, including all @username instances in our network analyses should generate a more comprehensive picture of the Twitter interaction network. For terminology purposes, we refer to any use of the @username convention as a "mention." In network terms, being mentioned refers to a user's inbound link, while mentioning others refers to a user's outbound link.

Of the more than 3 million users in our dataset, 47.61% mentioned another user, 31.23% were mentioned, and these users are highly concentrated within the group of high Twitter participation. That is, the correlation between the log of the number of posts (log(nPost)) and the log of the number of mentions (log(nMention)) is 0.62 ($p<2.2e-16$), and the correlation between log(nPost) and the log of the number of times mentioned (log(nMentioned)) is 0.54 ($p<2.2e-16$). For historical comparison, we note the expansion of mentioning in the two years since the Java et al. work that reported only one in eight tweets contained @username and only 21% of users used the convention.

In addition, activity levels in mentioning and being mentioned are highly correlated (0.63, $p<2.2e-16$), although more than 65% of users have more instances of mentioning other people than being mentioned themselves (versus about 25% of users with the opposite pattern and about 10% with equal numbers of mentions and being mentioned). This uneven distribution between in-links and out-links is also revealed in the low correlation between in degree (number of people who mention you) and out degree (number people you mention; $r=0.08$, $p<2.2e-16$). This indicates that @username mention links converge to a relatively small set of users, and raises the issue of the degree of reciprocity in the Twitter mention network, which we address next.

Reciprocity

If mentioning is an indicator of a social network tie between two individuals, to what extent is that relationship reciprocal? To measure the degree to which mentioning is uni- or bi-directional, we examined the likelihood a person will mention a person who has mentioned them, within a reasonable time period. To do so we employed survival analysis to quantify the probability that a mention of person A by person B would be followed by a mention of person B by person A within 300 hours. If I mention you, do you mention me within 12 days or so? We purposefully scoped

[3] http://en.wikipedia.org/wiki/Survival_analysis

this analysis to this time period first to establish a somewhat strict definition of reciprocity and second to get a sense of how conversational is Twitter.

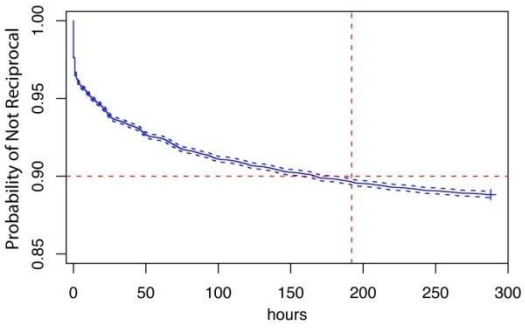


Figure 1. Survival curve of non-reciprocal mentions

Figure 1 presents the survival curve of non-reciprocally mentioned mentions over our 300 hour observation period. Almost 90% of mentions do not generate a corresponding mention in the reverse direction during this time. Further, except for the sharp drop in the very earliest hours following a mention where almost 5% generated a reverse mention, the trend is rather flat and the probability of reverse mentioning approaches zero after 4 days. This indicates that reciprocal mentioning is not common, especially compared to the relative prominence of mentioning generally (more than one third of tweets in our sample).

It is possible that users more likely to reciprocally mention and to do so promptly simply are very active Twitter users who post and mention with high frequency. To account for this, we used the Cox proportional-hazards regression model to quantify the degree to which various user activity measures can predict reciprocal mentioning (recall that degree of deviation from 1.0 indicates the strength of the predictors). Table 1 shows results of the regression predicting the time at which person B will mention person A after person A has mentioned person B. Indeed, the overall mention rate of person B (B-mentionRate) is a strong predictor of a faster reciprocal mention. Conversely person A's mention rate (A-mentionRate) and the frequency with which person B is mentioned (B-MentionedRate) are predictors of slower reciprocal mentioning. This indicates that those who are more frequently mentioned tend to not show reciprocal

	exp(coef)	z	Pr(>|z|)
A-nPost	1.0005	7.54	4.65e-14 ***
B-nPost	0.9998	-3.07	0.00211 **
A-nMention	0.9995	-7.59	3.23e-14 ***
A-nMentioned	1.0002	17.76	< 2e-16 ***
B-nMention	1.0009	17.04	< 2e-16 ***
B-nMentioned	0.9992	-12.36	< 2e-16 ***
A-MentionRate	0.8372	-18.67	< 2e-16 ***
A-MentionedRate	1.0030	12.76	< 2e-16 ***
B-MentionRate	1.9909	156.66	< 2e-16 ***
B-MentionedRate	0.7245	-66.29	< 2e-16 ***
R^2= 0.052 (max possible= 0.858)			

Table 1: Regression analysis predicting reciprocal mentioning of person A by person B.

mentioning behaviors, and that for the original person mentioning (person A), when she mentions frequently (with a higher mentionRate), she has less chance of being reciprocally mentioned by person B. Figures 2a and 2b show the comparison grouped by two of above predictors.

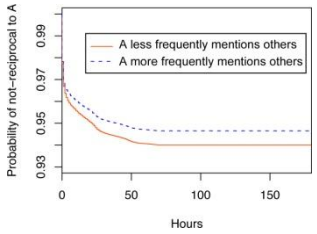


Figure 2a. Survival curve of non-reciprocal mention by A-Mentions-rate

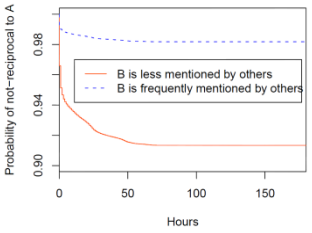


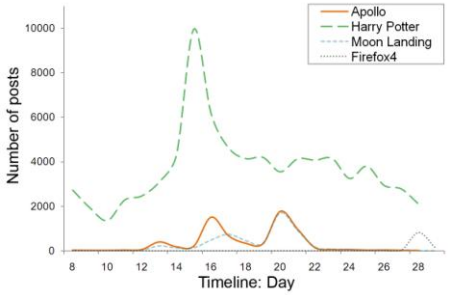
Figure 2b. Survival curve of non-reciprocal mention by B-isMentioned-rate

Containing Links as Indicator of Being Mentioned

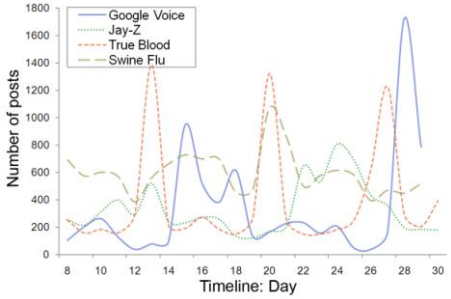
Finally, we detected a small but significant correlation between indegree (normalized by total degree) and log(nPost) ($r=0.126$, $p<2.2e-16$) and the ratio of containing link(s) in posts ($r=0.171$, $p<2.2e-16$). This result suggests that a user can significantly add more value to a tweet by providing external information sources.

Information Diffusion

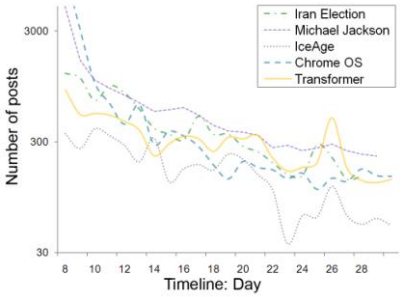
Our major focus was how information is diffused within the social network structures of Twitter. To start we present a few examples of temporal aspects of how information propagates in Twitter globally, without regard to social network structure. We then describe construction of an



Complete cycle



Undetermined trends



Declining trending topics

Figure 3: Sample trending topics

influence network that we use to measure the speed, scale and range of information propagation through Twitter's social network structures. We note here that for these analyses we used a 20 day (July 8 – 28, 2009) subset of our dataset, as that was the largest dataset we could use while maintaining reasonable compute time for our statistics.

Global temporal patterns

In order to get a sense about how information propagates in Twitter, we first investigated the general temporal patterns for different trend topics. Because we are primarily interested in information that has reached a considerable scale of influence, we only observed popular topics that many people mentioned in their posts. Figure 3 presents several trending topics during the time of our dataset. These topics are at different stages of their lifecycles, with some topics having gone through an almost full cycle during this period (e.g., Moon Landing), and others, such as Iran Election and Michael Jackson in the later stages of their lifecycles. We see that many topics have posting spikes and that even for these relatively prominent topics, many had completed their life cycle or had tailed off considerably during our observation period.

Influence Network

To measure how information propagates through network structures in Twitter we constructed an influence network that is similar to the mention network used earlier, but with an additional constraint of topical similarity in the tweet. That is we first consider that B mentioning A is an indicator that B was influenced by A in some way. However, because there is no direct link from B's tweet to A's tweet we do not know which tweet of A that B is mentioning. For better accuracy of measuring an influential relationship we use topics to confine the probability. As shown in Figure 4, we define B being influenced by A when B also talks about the same topic and mentions A in a single tweet at some point after A's tweet.

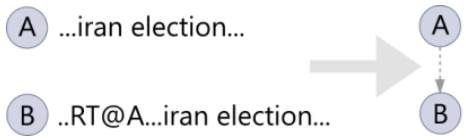


Figure 4: Topic-constrained influence of one Twitter user on another.

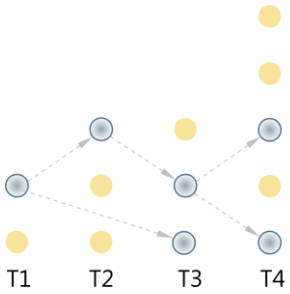


Figure 5: Building an influence network.

Figure 5 demonstrates how we then built our network of influence. All posts that contain the topic keywords (e.g., "Iran Election") are labeled with timestamps. From these, we track those that include mentions and the same topic keywords as earlier tweets, including obvious keyword

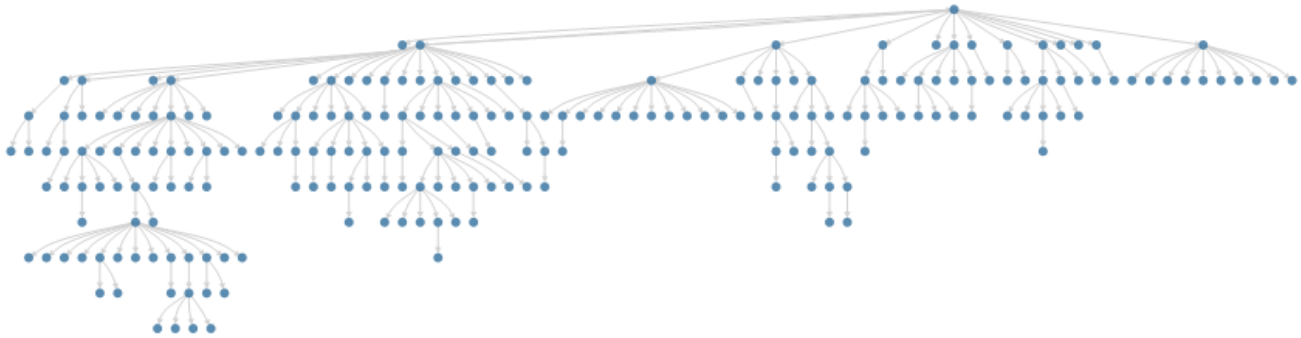


Figure 6: An example influence network about "Iran Election"

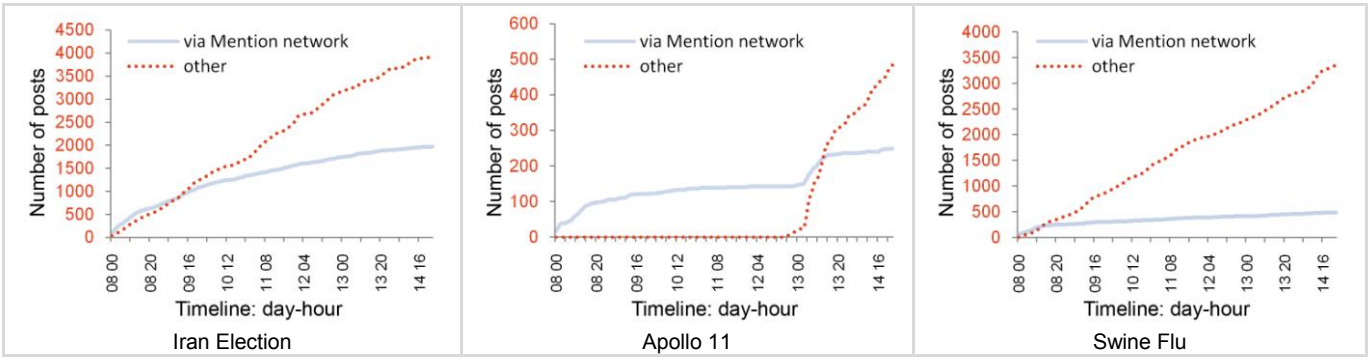


Figure 7: Comparing cumulative curves of number of nodes in and not in influence networks

variants like changes in case. We consider the current node and its parent to both belong to the network diffusing the topic. Figure 6 presents an example of an influence network around the topic of the Iran Election. Note that there will be many such influence networks in Twitter for a given topic, most of which will be operating at the same time. Also note that this is a particularly influential example that produced many offspring nodes, and this served as a good illustration of the flat pattern of the influence tree. In fact most nodes did not even generate offspring of the first degree.

Any given influence network will contain only a small subset of Twitter users. That is, the number of users tweeting about a topic within an influence network is likely to be far smaller than the number of users tweeting on that topic globally, with growth curves likely to flatten out before the growth across all of Twitter flattens out. Figure 7 shows cumulative curves for tweets on three different topics both within influence networks and for all other tweets containing the same topic keywords (that is, tweets that were not influenced by others according to our definition). As expected, the growth of tweets in influence networks flatten out well below the general growth of tweets about the same topic. However, worth calling out is that topic growth within mention networks can happen faster, even if just by a few hours as in "Iran Election" and "Apollo 11". Thus, even in cases where the absolute volume of tweets about a topic is smaller in influence networks than in Twitter as a whole, these networks may still be of value.

Local dynamics: speed, scale, and range

To further investigate the local dynamics of information diffusion in Twitter, we developed measures for three dimensions of influence networks in Twitter: speed, whether and when an influence action will take place; scale, how many others would be influenced at the first degree; and range, how far the influence can continue on in depth. We discuss each in turn.

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. Again employing survival analysis, both questions can be addressed in a single model. Similar to the earlier analysis predicting when a mention is going to be reciprocated, we predict when a tweet containing a topic is likely to be mentioned by another tweet also containing the topic.

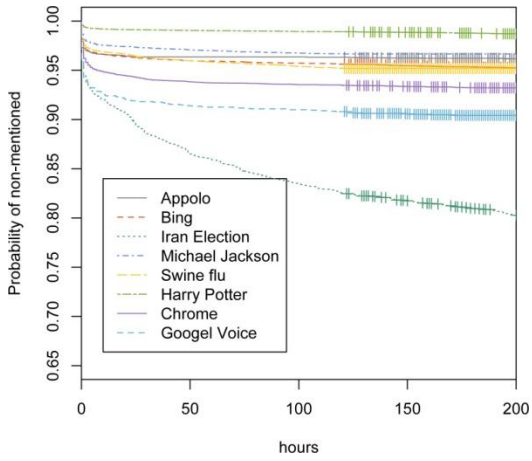


Figure 8: Survival curve of non-reciprocal tweets

Figure 8 displays the survival curves of eight sample topics. The y-axis is the portion of posts that have not obtained an offspring at the time point. First we see significant topic differences. For example, tweets about "Iran Election" are much more likely to be mentioned than others. Second, influenced nodes are mostly likely to appear in the first couple of hours after the tweet. In general, the majority of tweets did not produce any offspring at all.

Beyond gross level differences in speed across topics, we assessed the degree to which a number of features of both users and tweets themselves predict the speed of influence. For instance, aspects of each individual author, such as their activity level in tweeting and mentioning and being mentioned may also predict influence. 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:

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 2: 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.

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 2. We see that when the author is more active in posting (nPost) and has a higher rate of being mentioned (MentionedRate), the present tweet will gain influenced offspring in a shorter time. When the post is a mention per se (isMention), it has a higher chance to continue the influence. 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 influential in terms of producing 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 2, 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 influenced 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 influential 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 are influenced as first degree child nodes in the mention 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 and we have already looked at the likelihood of being mentioned in

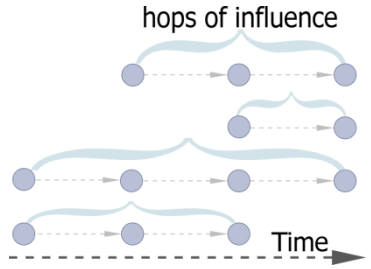


Figure 9: A generic influence tree

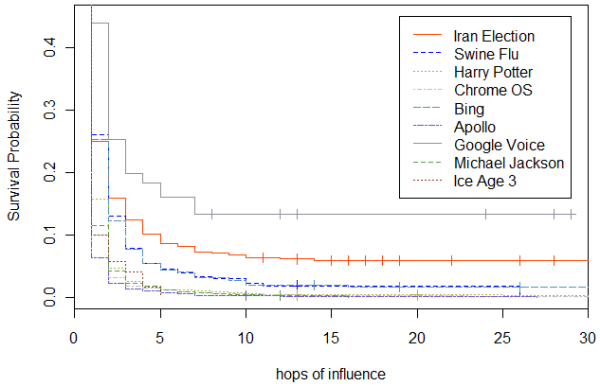


Figure 10: Influential life curve of ancestor nodes

earlier sections, 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 3 presents regression results on our sample trending topics. R-square 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. This is consistent with earlier analysis: those frequently mentioned users are good at producing child nodes both in terms of speed and quantity. In addition, including links in tweets often generates more child nodes.

Range As a final metric of information diffusion, we measure the range of influence as indicated by the number

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 3: Predicting number of child nodes one can produce

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)
Reporting exp(coef) with p-value. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1								

Table 4: Predicting length of influence chain of ancestor nodes

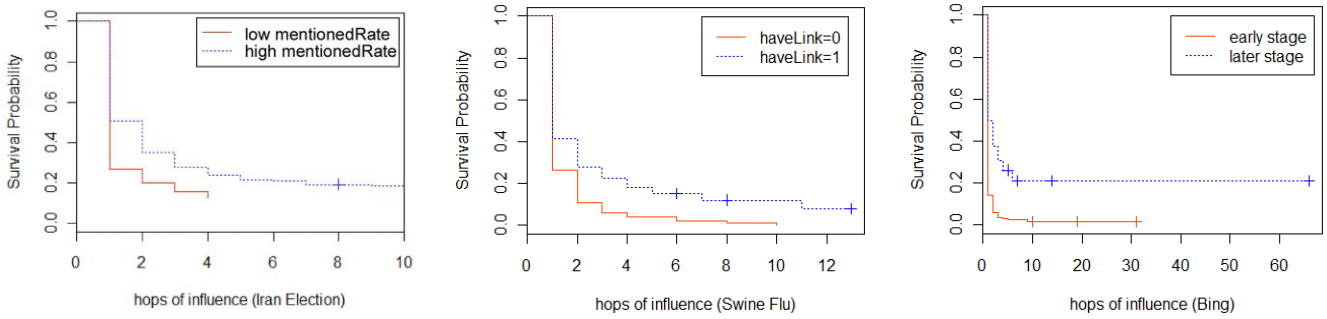


Figure 11: Key predictors of range of influence for three example topics

of hops in an influence 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 9, the length of the chain indicates how far the original node diffuses its influence in depth.

First we investigated general patterns of these influence chains. Figure 10 presents the survival curves of influence chains for our sample trending topics. For most of these topics, more than half the ancestor nodes fail to produce their offspring of 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 "Google Voice" tend to have longer chain life than topics like "Ice Age 3".

Similar to our analyses for speed and scale, we again examined aspects of users and tweets that may predict greater range of influence. Table 4 presents the significant results of a regression analysis predicting the length of a topic chain within an influence network. Figure 11 visualizes three of the better predictors: whether the poster has a high or low rate of being mentioned, whether the tweet contains a link, and whether the tweet occurred at the early or late stages of our observation. From these we see that the probability of survival is higher at increasing hops of influence when the poster has a high mentioned rate, when the tweet contains a link (for some topics), and when the tweet comes at a later stage.

DISCUSSION

Our analyses focused on @username mentions in order to utilize the "hidden" network of actual user interactions in Twitter rather than the potentially very passive follower network. We first examined basic properties of the mention network and then used that network by scoping it to specific topics to measure aspects of how the network impact information diffusion in Twitter. The initial analysis revealed that mentioning is very skewed in Twitter, with a small number of users getting the majority of the mentions. Continuing the lack of reciprocity, Twitter does not appear to be a conversation: mentioning a person does not predict that the person will in turn mention you within 300 hours. In fact, as shown in Table 1, the more a person mentions, the less likely they are to be mentioned. If we assume that mentioning is an indicator of value, this general skewness suggests a fairly small percentage of users contributing high value content. From a system design perspective, such as designing an archive and search system for Twitter, this implies that prioritizing content from user's who are mentioned frequently would be a good strategy for obtaining and surfacing higher quality content.

For a more nuanced picture of the role mentioning plays in information diffusion in Twitter, we constructed an influence network, which looks at the "flow" of @username mention in tweets over time and within specific topics. At a high level, we saw that topics (see Apollo 11 and Iran Election in Figure 7) can propagate through mention networks earlier than through Twitter more broadly. Thus, for "real time search", closely following mentions,

especially those from users with certain influence characteristics (discussed below) is a good way to identify burgeoning topics in their early stages. This could happen in conjunction with monitoring trending keywords, such as by weighting keywords that are both trending upward and were tweeted by particular users.

What are the characteristics of users and content with the most influence in the influence network? Here were ran three analyses to attempt to extract these characteristics with respect to the speed, scale, and range of information propagating through Twitter. First, for speed (how quickly will a tweet produce an offspring tweet), again 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 3) 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 influence network), a few predictors stand out. As we have seen consistently, the mention 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 influence 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. This is particularly salient given our interaction analysis results that showed that when information propagates (at least as denoted by the @username convention) it often happens through a small core of users and is very much unidirectional. In short, the social component that makes Twitter a social medium really appears to play a key role with respect to Twitter as an information medium.

Limitations and Future Work

Our dataset was reasonably comprehensive, but nonetheless poses issues related to sampling. For example, our sample covered only a short period of time and thus did not allow us to observe and quantify complete life cycles of all the trending topics. Thus we could not specifically differentiate stages and paint the full trajectory of the change in diffusion efficiency. Also, since the primary goal of this study was to understand how macro trending tropics evolve and accumulate influence, we did not look at those smaller topics taking places within small communities. In general, as with any social network analysis, we cannot draw the complete representation of the real structure. In this way, our analysis sought to generate aggregate and statistical understanding of the structure. We hope this work complements qualitative investigation that better define individual differences.

As Twitter continues to evolve, these types of analyses should be repeated and expanded. One area for expansion might be to exploit some of the user conventions seen in Twitter. For example, ever Friday Twitter users engage in the "follow Friday" practice in which they suggest other users to follow. This may be another mechanism for uncovering high value users and could be another variable to factor into a ranking algorithm. It would interesting to measure the correlation between the frequency of being suggested on follow Friday and some of the metrics (such as posting and mention rate) we saw as important in the current analyses.

As second area worthy of future work and one potentially applicable beyond Twitter is to better understand how diffusion efficiency varies at different points in the life cycle of a topic. Any systematic differences here might predict different types of topics or suggest latent topics that are about to become major news stories.

CONCLUSION

Twitter is the currently dominant system for microblogging, a new form of social media that has attributes of both social networking and blogging. We analyzed a large sample of one month of Twitter content with the specific goals of understanding the basic properties of the network of interactions in Twitter and then to quantify the ways that network properties influence information propagation. The primary take-away points are that Twitter interactions are

very unidirectional, flowing through a small number of critical users, and that predicting attributes of information flow (i.e., our speed, scale, and range analyses) requires analysis of both the tweets themselves, but also the social aspects of Twitter.

ACKNOWLEDGMENTS

Anonymous for blind review

REFERENCES

1. Adamic, L.A. and Glance, N. *The political blogosphere and the 2004 U.S. election: divided they blog*. in *Proceedings of the 3rd international workshop on Link discovery*. 2005. Chicago, Illinois ACM.
2. Adar, E. and Adamic, L.A. *Tracking Information Epidemics in Blogspace*. in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. 2005.
3. Adar, E., Zhang, L., Adamic, L., and Lukose, R. *Implicit Structure and the Dynamics of Blogspace*. in *Workshop on the Weblogging Ecosystem*. 2004.
4. Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. *Group Formation in Large Social Networks: Membership, Growth, and Evolution*. in *Proceedings of KDD*. 2006.
5. Gabrilovich, E., Susan Dumais, and Horvitz, E. *Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty*. in *Proceedings of the Thirteenth International World Wide Web Conference (WWW 2004)*. 2004. New York.
6. Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A., *Information diffusion through blogspace*, in *Proceedings of the 13th international conference on World Wide Web*. 2004, ACM: New York, NY, USA.
7. Havre, S., Hetzler, B., and Nowell, L. *ThemeRiver: Visualizing Theme Changes over Time*. in *Proceedings of the IEEE Symposium on Information Vizualization*. 2000.
8. Huberman, B., Romero, D.M., and Wu, F., *Social networks that matter: Twitter under the microscope*. First Monday, 2009. 14.
9. Java, A., Song, X., Finin, T., and Tseng, B. *Why We Twitter: Understanding Microblogging Usage and Communities*. in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. 2007.
10. Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. *On the bursty evolution of blogspace*. in *Proceedings of the 12th international conference on World Wide Web* 2003: ACM.
11. Kumar, R., Novak, J., Raghavan, P., and Tomkins, A., *Structure and evolution of blogspace*. Communications of the ACM, 2004. 47(12): p. 35 - 39.
12. Kumar, R., Novak, J., Raghavan, P., and Tomkins, A., *On the Bursty Evolution of Blogspace*. World Wide Web, 2005. 8(2): p. 159-178.
13. Leskovec, J., Backstrom, L., and Kleinberg, J. *Meme-tracking and the dynamics of the news cycle*. in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009.
14. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., Vanbriesen, J., and Glance, N. *Cost-effective outbreak detection in networks*. in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2007: ACM.
15. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., and Tomkins, A., *Geographic Routing in Social Networks*. Proceedings of the National Academy of Sciences, 2005. 103(33): p. 11623-11628.
16. Tseng, B.L., Tatemura, J., and Wu, Y. *Tomographic Clustering To Visualize Blog Communities as Mountain Views* in *Proceedings of 2nd Annual Workshop on the Weblogging Ecosystem*. 2005.