Trends in Social Media : Persistence and Decay

Sitaram Asur
Social Computing Lab
HP Labs
Palo Alto, California, USA
sitaram.asur@hp.com

Bernardo A. Huberman
Social Computing Lab
HP Labs
Palo Alto, California, USA
bernardo.huberman@hp.com

Gabor Szabo
Social Computing Lab
HP Labs
Palo Alto, California, USA
gabors@hp.com

Chunyan Wang
Dept. of Applied Physics
Stanford University
California, USA
chunyan@stanford.edu

ABSTRACT

Social media generates a prodigious wealth of real-time content at an incessant rate. From all the content that people create and share, only a few topics manage to attract enough attention to rise to the top and become temporal trends which are displayed to users. The question of what factors cause the formation and persistence of trends is an important one that has not been answered yet. In this paper, we conduct an intensive study of trending topics on Twitter and provide a theoretical basis for the formation, persistence and decay of trends. We also demonstrate empirically how factors such as user activity and number of followers do not contribute strongly to trend creation and its propagation. In fact, we find that the resonance of the content with the users of the social network plays a major role in causing trends.

1. INTRODUCTION

Social media is growing at an explosive rate, with millions of people all over the world generating and sharing content on a scale barely imaginable a few years ago. This has resulted in massive participation with countless number of updates, opinions, news, comments and product reviews being constantly posted and discussed in social web sites such as Facebook, Digg and Twitter, to name a few.

This widespread generation and consumption of content has created an extremely competitive online environment where different types of content vie with each other for the scarce attention of the user community. In spite of the seemingly chaotic fashion with which all these interactions take place, certain topics manage to attract an inordinate amount of attention, thus bubbling to the top in terms of popularity. Through their visibility, this popular topics contribute to the collective awareness of what is trending and at times can also affect the public agenda of the community.

At present there is no clear picture of what causes these topics to become extremely popular, nor how some persist in the public eye longer than others. There is considerable evidence that one aspect that causes topics to decay over time is their novelty [11]. Another factor responsible for their decay is the competitive nature of the medium. As content starts propagating throught a social network it can usurp the positions of earlier topics of interest, and due to the limited attention of users it is soon rendered invisible by newer content. Yet another aspect responsible for the popularity of certain topics is the influence of members of the network on the propagation of content. Some users generate content that resonates very strongly with their followers thus causing the content to propagate and gain popularity [9].

The source of that content can originate in standard media outlets or from users who generate topics that eventually become part of the trends and capture the attention of large communities. In either case the fact that a small set of topics become part of the trending set means that they will capture the attention of a large audience for a short time, thus contributing in some measure to the public agenda. When topics originate in media outlets, the social medium acts as filter and amplifier of what the standard media produces and thus contributes to the agenda setting mechanisms that have been thoroughly studied for more than three decades [7] .

In this paper, we study trending topics on Twitter, an immensely popular microblogging network on which millions of users create and propagate enormous content via a steady stream on a daily basis. The trending topics, which are shown on the main website, represent those pieces of content that bubble to the surface on Twitter owing to frequent mentions by the community. Thus they can be equated to crowdsourced popularity. We then determine the factors that contribute to the creation and evolution of these trends, as they provide insight into the complex interactions that lead to the popularity and persistence of certain topics on Twitter, while most others fail to catch on and are lost in the flow.

We first analyze the distribution of the number of tweets across trending topics. We observe that they are characterized by a strong log-normal distribution, similar to that found in other networks such as Digg and which is generated by a stochastic multiplicative process [11]. We also find that the decay function for the tweets is mostly linear. Subsequently we study the persistence of the trends

to determine which topics last long at the top. Our analysis reveals that there are few topics that last for long times, while most topics break fairly quickly, in the order of 20-40 minutes. Finally, we look at the impact of users on trend persistence times within Twitter. We find that traditional notions of user influence such as the frequency of posting and the number of followers are not the main drivers of trends, as previously thought. Rather, long trends are characterized by the resonating nature of the content, which is found to arise mainly from traditional media sources. We observe that social media behaves as a selective amplifier for the content generated by traditional media, with chains of retweets by many users leading to the observed trends.

2. RELATED WORK

There has been some prior work on analyzing connections on Twitter. Huberman et al. [5] studied social interactions on Twitter to reveal that the driving process for usage is a sparse hidden network underlying the friends and followers, while most of the links represent meaningless interactions. Jansen et al. [6] have examined Twitter as a mechanism for word-of-mouth advertising. They considered particular brands and products and examined the structure of the postings and the change in sentiments. Galuba et al. [4] proposed a propagation model that predicts which users will tweet about which URL based on the history of past user activity.

Yang and Leskovec [12] examined patterns of temporal behavior for hashtags in Twitter. They presented a stable time series clustering algorithm and demonstrate the common temporal patterns that tweets containing hashtags follow. There have also been earlier studies focused on social influence and propagation. Agarwal et al. [1] studied the problem of identifying influential bloggers in the blogosphere. They discovered that the most influential bloggers were not necessarily the most active. Aral et al, [2] have distinguished the effects of homophily from influence as motivators for propagation. As to the study of influence within Twitter, Cha et al. [3] performed a comparison of three different measures of influence - indegree, retweets, and user mentions. They discovered that while retweets and mentions correlated well with each other, the indegree of users did not correlate well with the other two measures. Based on this, they hypothesized that the number of followers may not a good measure of influence. Recently, Romero and others [9] introduced a novel influence measure that takes into account the passivity of the audience in the social network. They developed an iterative algorithm to compute influence in the style of the HITS algorithm and empirically demonstrated that the number of followers is a poor measure of influence.

3. TWITTER

Twitter is an extremely popular online microblogging service, that has gained a very large user following, consisting of close to 200 million users. The Twitter graph is a directed social network, where each user chooses to follow certain other users. Each user submits periodic status updates, known as *tweets*, that consist of short messages limited in size to 140 characters. These updates typically consist of personal information about the users, news or links to content such as images, video and articles. The posts made by a user are automatically displayed on the user's profile page, as well as shown to his followers. A *retweet* is a post originally made by one user that is forwarded by another user. Retweets are useful for propagating interesting posts and links through the Twitter community.

Twitter has attracted lots of attention from corporations due to the immense potential it provides for viral marketing. Due to its huge reach, Twitter is increasingly used by news organizations to disseminate news updates, which are then filtered and commented on by the Twitter community. A number of businesses and organizations are using Twitter or similar micro-blogging services to advertise products and disseminate information to stockholders.

4. TWITTER TRENDS DATA

Trending topics are presented as a list by Twitter on their main Twitter.com site, and are selected by an algorithm proprietary to the service. They mostly consist of two to three word expressions, and we can assume with a high confidence that they are snippets that appear more frequently in the most recent stream of tweets than one would expect from a document term frequency analysis such as TFIDF. The list of trending topics is updated every few minutes as new topics become popular.

Twitter provides a Search API for extracting tweets containing particular keywords. To obtain the dataset of trends for this study, we repeatedly used the API in two stages. First, we collected the trending topics by doing an API query every 20 minutes. Second, for each trending topic, we used the Search API to collect all the tweets mentioning this topic over the past 20 minutes. For each tweet, we collected the author, the text of the tweet and the time it was posted. Using this procedure for data collection, we obtained 16.32 million tweets on 3361 different topics over a course of 40 days in Sep-Oct 2010.

We picked 20 minutes as the duration of a timestamp after evaluating different time lengths, to optimize the discovery of new trends while still capturing all trends. This is due to the fact that Twitter only allows 1500 tweets per search query. We found that with 20 minute intervals, we were able to capture all the tweets for the trending topics efficiently.

We noticed that many topics become trends again after they stop trending according to the Twitter trend algorithm. We therefore considered these trends as separate sequences: it is very likely that the spreading mechanism of trends has a strong time component with an initial increase and a trailing decline, and once a topic stops trending, it should be considered as new when it reappears among the users that become aware of it later. This procedure split the 3468 originally collected trend titles into 6084 individual trend sequences.

5. DISTRIBUTION OF TWEETS

We measured the number of tweets that each topic gets in 20 minute intervals, from the time the topic starts trending until it stops, as described earlier. From this we can sum up the tweet counts over time to obtain the cumulative number of tweets $N_q(t_i)$ of topic q for any time frame t_i,

$$N_q(t_i) = \sum_{\tau=1}^{i} n_q(t_\tau), \tag{1}$$

where $n_q(t)$ is the number of tweets on topic q in time interval t. Since it is plausible to assume that initially popular topics will stay popular later on in time as well, we can calculate the ratios $C_q(t_i, t_j) = N_q(t_i)/N_q(t_j)$ for topic q for time frames t_i and t_j. Figure 1(a) shows the distribution of $C_q(t_i, t_j)$'s over all topics for four arbitrarily chosen pairs of time frames (nevertheless such that $t_i > t_j$, and t_i is relatively large, and t_j is small).

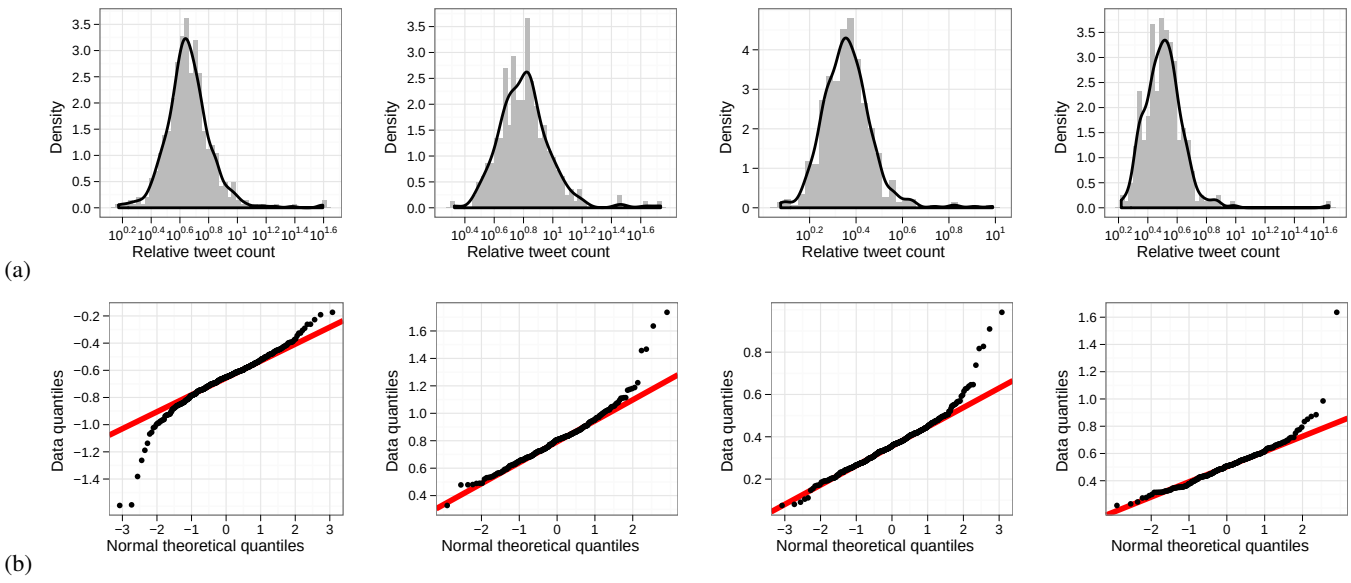

Figure 1: (a) The densities of the ratios between cumulative tweet counts measured in two respective time frames. From left to right in the figure, the indices of the time frames between which the ratios were taken are: (2, 10), (2, 14), (4, 10), and (4, 14), respectively. The horizontal axis has been rescaled logarithmically, and the solid line in the plots shows the density estimates using a kernel smoother. *(b)* The Q-Q plots of the cumulative tweet distributions with respect to normal distributions. If the random variables of the data were a linear transformation of normal variates, the points would line up on the straight lines shown in the plots. The tails of the empirical distributions are apparently heavier than in the normal case.

These figures immediately suggest that the ratios $C_q(t_i, t_j)$ are distributed according to log-normal distributions, since the horizontal axes are logarithmically rescaled, and the histograms appear to be Gaussian functions. To check if this assumption holds, consider Fig. 1(b), where we show the Q-Q plots of the distributions of Fig. 1(a) in comparison to normal distributions. We can observe that the (logarithmically rescaled) empirical distributions exhibit normality to a high degree for later time frames, with the exception of the high end of the distributions. These 10-15 outliers occur more frequently than could be expected for a normal distribution.

Log-normals arise as a result of multiplicative growth processes with noise [8]. In our case, if $N_q(t)$ is the number of tweets for a given topic q at time t, then the dynamics that leads to a log-normally distributed $N_q(t)$ over q can be written as:

$$N_q(t) = [1 + \gamma(t)\xi(t)]\, N_q(t-1), \tag{2}$$

where the random variables $\xi(t)$ are positive, independent and identically distributed as a function of t with mean 1 and variance σ^2. Note that time here is measured in discrete steps ($t-1$ expresses the previous time step with respect to t), in accordance with our measurement setup. $\gamma(t)$ is introduced to account for the novelty decay [11]. We would expect topics to initially increase in popularity but to slow down their activity as they become obsolete or known to most users. Since $\gamma(t)$ is made up of decreasing positive numbers, the growth of N_t slows with time.

To see that Eq. (2) leads to a log-normal distribution of $N_q(t)$, we first expand the recursion relation:

$$N_q(t) = \prod_{s=1}^{t} [1 + \gamma(s)\xi(s)]\, N_q(0). \tag{3}$$

Here $N_q(0)$ is the initial number of tweets in the earliest time step. Taking the logarithm of both sides of Eq. (3),

$$\ln N_q(t) - \ln N_q(0) = \sum_{s=1}^{t} \ln [1 + \gamma(s)\xi(s)] \tag{4}$$

The RHS of Eq. (4) is the sum of a large number of random variables. The central limit theorem states thus that if the random variables are independent and identically distributed, then the sum asymptotically approximates a normal distribution. The i.i.d condition would hold exactly for the $\xi(s)$ term, and it can be shown that in the presence of the discounting factors (if the rate of decline is not too fast), the resulting distribution is still normal [11].

In other words, we expect from this model that $\ln [N_q(t)/N_q(0)]$ will be distributed normally over q when fixing t. These quantities were shown in Fig. 1 above. Essentially, if the difference between the two times where we take the ratio is big enough, the log-normal property is observed.

The intuitive explanation for the multiplicative model of Eq. (2) is that at each time step the number of *new* tweets on a topic is a multiple of the tweets that we already have. The number of past tweets, in turn, is a proxy for the number of users that are aware of the topic up to that point. These users discuss the topic on different forums, including Twitter, essentially creating an effective network through which the topic spreads. As more users talk about a particular topic, many others are likely to learn about it, thus giving the multiplicative nature of the spreading. The noise term is necessary to account for the stochasticity of this process. On the other hand, the monotically decreasing $\gamma(t)$ characterizes the decay in timeliness and novelty of the topic as it slowly becomes obsolete and known to most users, and guarantees that $N_q(t)$ does not grow

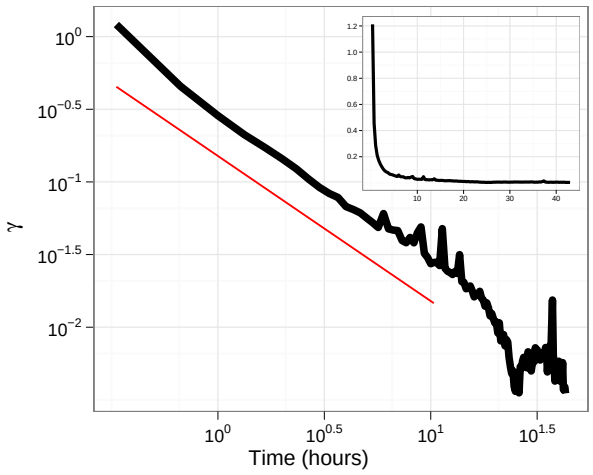

Figure 2: The decay factor $\gamma(t)$ in time as measured using Eq. (5). The log-log plot exhibits that it decreases in a power-law fashion, with an exponent that is measured to be exactly -1 (the linear regression on the logarithmically transformed data fits with $R^2 = 0.98$). The fit to determine the exponent was performed in the range of the solid line next to the function, which also shows the result of the fit while being shifted lower for easy comparison. The inset displays the same $\gamma(t)$ function on standard linear scales.

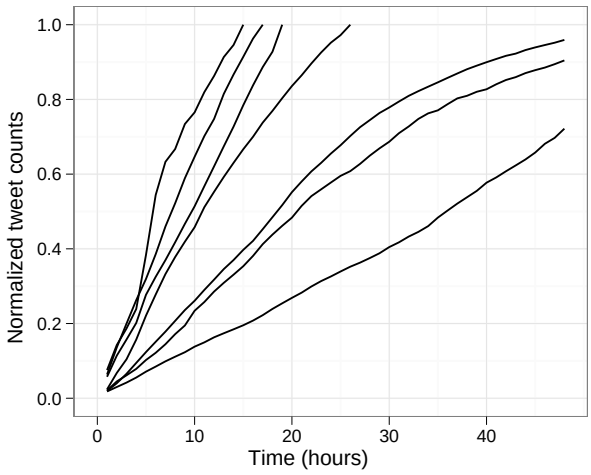

Figure 3: The number of total tweets on topics in the first 48 hours, normalized to 1 so that they can be shown on the same plot. The randomly selected topics were (from left to right): "Earnings", "#pulpopaul", "Sheen", "Deuces Remix", "Isaacs", "#gmp24", and "Mac App".

unbounded [11].

To measure the functional form of $\gamma(t)$, we observe that the expected value of the noise term $\xi(t)$ in Eq. (2) is 1. Thus averaging over the fractions between consecutive tweet counts yields $\gamma(t)$:

$$\gamma(t) = \left\langle \frac{N_q(t)}{N_q(t-1)} \right\rangle_q - 1. \quad (5)$$

The experimental values of $\gamma(t)$ in time are shown in Fig. 2. It is interesting to notice that $\gamma(t)$ follows a power-law decay very precisely with an exponent of -1, which means that $\gamma(t) \sim 1/t$.

6. THE GROWTH OF TWEETS OVER TIME

The interesting fact about the decay function $\gamma(t) = 1/t$ is that it results in a *linear increase* in the total number of tweets for a topic over time. To see this, we can again consider Eq. (4), and approximate the discrete sum of random variables with an integral of the operand of the sum, and substitute the noise term with its expectation value, $\langle \xi(t) \rangle = 1$ as defined earlier (this is valid if $\gamma(t)$ is changing slowly). These approximations yield the following:

$$\ln \frac{N_q(t)}{N_q(0)} \approx \int_{\tau=0}^{t} \ln\left[1 + \gamma(\tau)\right] d\tau \approx \int_{\tau=0}^{t} \frac{1}{\tau} d\tau = \ln t. \quad (6)$$

In simplifying the logarithm above, we used the Taylor expansion of $\ln(1+x) \approx x$, for small x, and also used the fact that $\gamma(\tau) = 1/\tau$ as we found experimentally earlier.

It can be immediately seen then that $N_q(t) \approx N_q(0)\, t$ for the range of t where $\gamma(t)$ is inversely proportional to t. In fact, it can be easily proven that no functional form for $\gamma(t)$ would yield a linear increase in $N_q(t)$ other than $\gamma(t) \sim 1/t$ (assuming that the above approximations are valid for the stochastic discrete case). This suggests that the trending topics featured on Twitter increase their tweet counts linearly in time, and their dynamics is captured by the multiplicative noise model we discussed above.

To check this, we first plotted a few representative examples of the cumulative number of tweets for a few topics in Fig. 3. It is apparent that all the topics (selected randomly) show an approximate initial linear growth in the number of tweets.We also checked if this is true in general. Figure 4 shows the second discrete derivative of the total number of tweets, which we expect to be 0 if the trend lines are linear on average. A positive second derivative would mean that the growth is superlinear, while a negative one suggests that it is sublinear. We point out that before taking the average of all second derivatives over the different topics in time, we divided the derivatives by the average of the total number of tweets of the given topics. We did this so as to account for the large difference between the ranges of the number of tweets across topics, since a simple averaging without prior normalization would likely bias the results towards topics with large tweet counts and their fluctuations. The averages are shown in Fig. 4.

We observe from the figure that when we consider all topics there is a very slight sublinear growth regime right after the topic starts trending, which then becomes mostly linear, as the derivatives data is distributed around 0. If we consider only very popular topics (that were on the trends site for more than 4 hours), we observe an even better linear trend. One reason for this may be that topics that trend only for short periods exhibit a concave curvature, since they lose popularity quickly, and are removed from among the Twitter trends by the system early on.

These results suggest that once a topic is highlighted as a trend on a very visible website, its growth becomes linear in time. The reason for this may be that as more and more visitors come to the site and see the trending topics there is a constant probability that they will also talk and tweet about it. This is in contrast to scenarios where the primary channel of information flow is more informal.

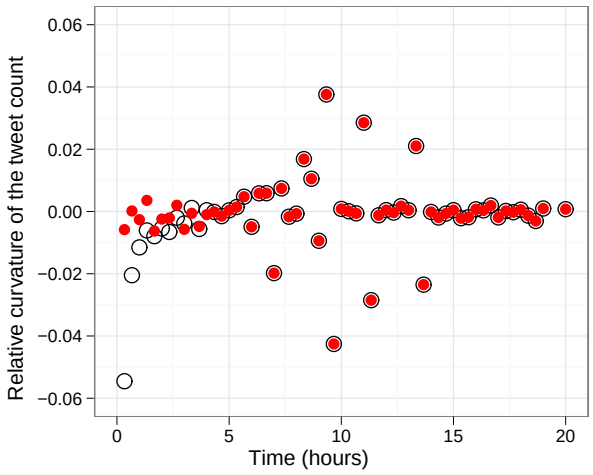

Figure 4: The average of the second derivative of the total number of tweets over all topics. For one topic, we first divided the derivative values by the mean of the tweet counts so as to minimize the differences between the wide range of topic popularities. The open circles show the derivatives obtained with this procedure for all topics, while the smaller red dots represent only topics that trended for longer than 4 hours.

In that case we expect that the growth will exhibit first a phase with accelerated growth and then slow down to a point when no one talks about the topic any more. Content that spreads through a social network or without external "driving" will follow such a course, as has been showed elsewhere [10, 12].

7. PERSISTENCE OF TRENDS

An important reason to study trending topics on Twitter is to understand why some of them remain at the top while others dissipate quickly. To see the general pattern of behavior on Twitter, we examined the lifetimes of the topics that trended in our study. From Fig 5(a) we can see that while most topics occur continuously, around 34% of topics appear in more than one sequence. This means that they stop trending for a certain period of time before beginning to trend again.

A reason for this behavior may be the time zones that are involved. For instance, if a topic is a piece of news relevant to North American readers, a trend may first appear in the Eastern time zone, and 3 hours later in the Pacific time zone. Likewise, a trend may return the next morning if it was trending the previous evening, when more users check their accounts again after the night.

Given that many topics do not occur continuously, we examined the distribution of the lengths sequences for all topics. In Fig 5(b) we show the length of the topic sequences. It can be observed that this is a power-law which means that most topic sequences are short and a few topics last for a very long time. This could be due to the fact that there are many topics competing for attention. Thus, the topics that make it to the top (the trend list) last for a short time. However, in many cases, the topics return to trend for more time, which is captured by the number of sequences shown in Fig 5(a), as mentioned.

7.1 Relation to authors and activity

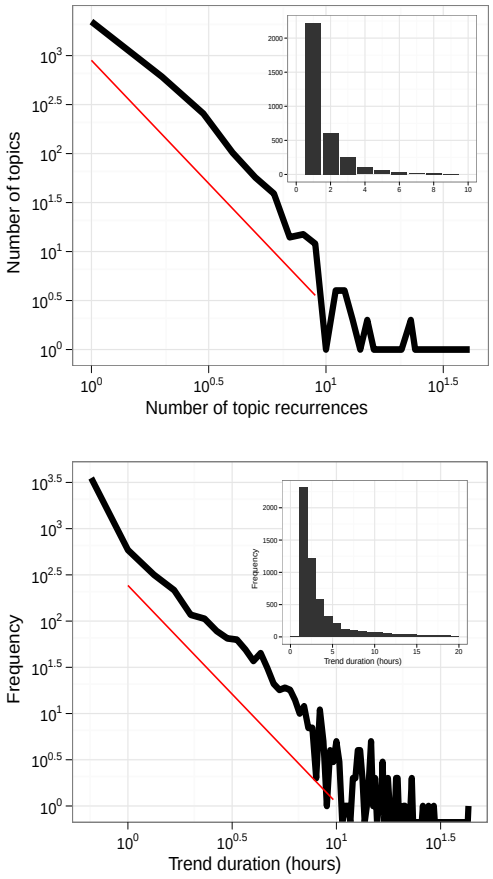

Figure 5: *(a)* The distribution of the number of sequences a trending topic comprises of *(b)* The distribution of the lengths of each sequence. Both graphs are shown in the log-log scale with the inset giving the actual histograms in the linear scale.

We first examine the authors who tweet about given trending topics to see if the authors change over time or if it is the same people who keep tweeting to cause trends. When we computed the correlation in the number of unique authors for a topic with the duration (number of timestamps) that the topic trends we noticed that correlation is very strong (0.80). This indicates that as the number of authors increases so does the lifetime, suggesting that the propagation through the network causes the topic to trend.

To measure the impact of authors we compute for each topic the active-ratio a_q as:

$$a_q = \frac{Number\ of\ Tweets}{Number\ of\ Unique\ Authors} \tag{7}$$

The correlation of active-ratio with trending duration is as shown in Fig 6. We observe that the active-ratio quickly saturates and varies little with time for any given topic. Since the authors change over time with the topic propagation, the correlation between number of tweets and authors is high (0.83).

7.2 Persistence of long trending topics

On Twitter each topic competes with the others to survive on the trending page. As we now show, for the long trending ones we can derive an expression for the distribution of their average length.

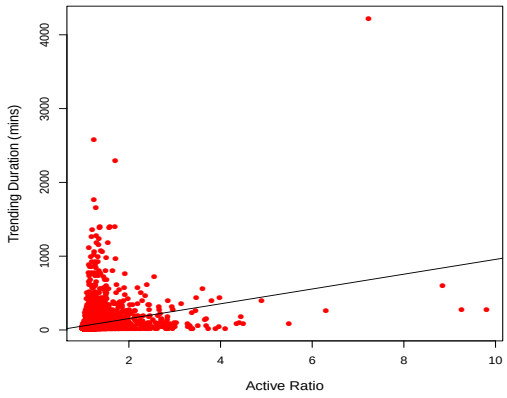

Figure 6: Relation between the active-ratio and the length of the trend across all topics, showing that the active-ratio does not vary significantly with time.

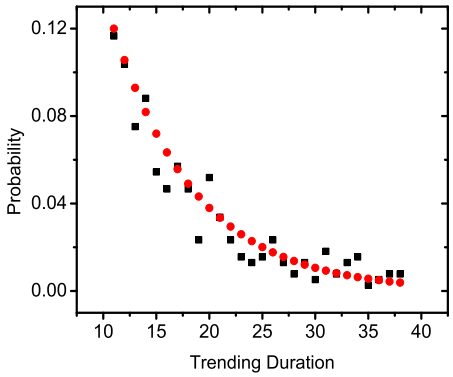

Figure 7: Distribution of trending times. The black dots represents actual trending data pulled from Twitter, and the red dots are the predictions from a geometric distribution with p=0.12.

We assume that, if the relative growth rate of tweets, denoted by $\phi_t = \frac{N_t}{N_{t-1}}$, falls below a certain threshold θ, the topic would stop trending. When we consider long-trending topics, as they grow in time, they overcome the initial novelty decay, and the γ term in equation (3) becomes fairly constant. So we can measure the change over time using only the random variable ξ as :

$$\log \phi_t = \log \frac{N_t}{N_{t-1}} = \log \frac{N_t}{N_0} - \log \frac{N_{t-1}}{N_0} \simeq \xi_t \tag{8}$$

Since the ξ_s are independent and identical distributed random variables, $\phi_1, \phi_2, \cdots \phi_t$ would be independent with each other. Thus the probability that a topic stops trending in a time interval s, where s is large, is equal to the probability that ϕ_s is lower than the threshold θ, which can be written as:

$$\begin{aligned} p = \Pr(\phi_s < \theta) &= \Pr(\log \phi_s < \log(\theta)) \\ &= \Pr(\xi_s < \log(\theta)) = F(\log \theta) \end{aligned} \tag{9}$$

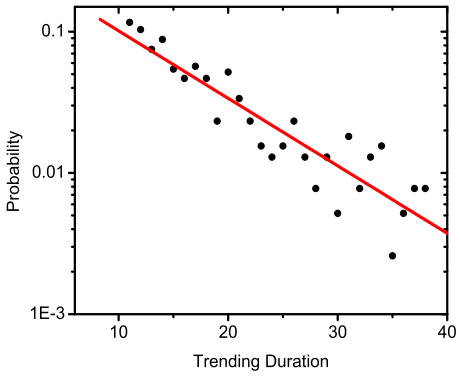

Figure 8: Fit of trending duration to density in log scale. The straight line suggests an exponential family of the trending time distribution. The red line gives a fit with an R^2 of 0.9112.

$F(x)$ is the cumulative distribution function of the random variable χ. Given that distribution we can actually determine the threshold for survival as:

$$\theta = e^{F^{-1}(p)} \tag{10}$$

From the independence property of the ϕ, the duration or life time of a trending topic, denoted by L, follows a geometric distribution, which in the continuum case becomes the exponential distribution. Thus, the probability that a topic survives in the first k time intervals and fails in the $k+1$ time interval, given that k is large, can be written as:

$$\Pr(L = k) = (1-p)^k p \tag{11}$$

The expected length of trending duration L would thus be:

$$< L >= \sum_0^\infty (1-p)^k p \cdot k = \frac{1}{p} - 1 = \frac{1}{F(\log \theta)} - 1 \tag{12}$$

We considered trending durations for topics that trended for more than 10 timestamps on Twitter. The comparison between the geometric distribution and the trending duration is shown in Fig 7. In Fig 8 the fit of the trending duration to density in a logarithmic scale suggests an exponential function for the trending time. The R-square of the fitting is 0.9112.

8. TREND-SETTERS

We consider two types of people who contribute to trending topics - the sources who begin trends, and the propagators who are responsible for those trends propagating through the network due to the nature of the content they share.

8.1 Sources

We examined the users who initiate the most trending topics. First, for each topic we extracted the first 100 users who tweeted about it prior to its trending. The distribution of these authors and the topics is a power-law, as shown in Fig 9. This shows that there are few authors who contribute to the creation of many different topics. To focus on these multi-tasking users, we considered only the authors who contributed to at least five trending topics.

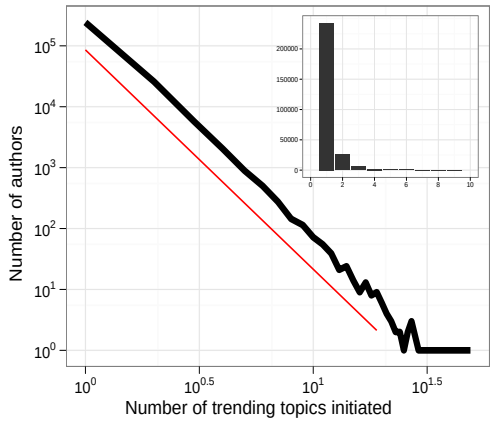

Figure 9: Distribution of the first 100 authors for each trending topic. The log-log plot shows a power-law distribution. The inset graph gives the actual histogram in the linear scale.

When we consider people who are influential in starting trends on Twitter, we can hypothesize two attributes - a high frequency of activity for these users, as well as a large follower network. To evaluate these hypotheses we measured these two attributes for these authors over these months.

Frequency: The tweet-rate can effectively measure the frequency of participation of a Twitter user. The mean tweet-rate for these users was 26.38 tweets per day, indicating that these authors tweeted fairly regularly. However, when we computed the correlation of the tweet-rate with the number of trending topics that they contributed to, the result was a weak positive correlation of 0.22. This indicates that although people who tweet a lot do tend to contribute to the trending topics, the rate by itself does not strongly determine the popularity of the topic. In fact, they happen to tweet on a variety of topics, many of which do not become trends. We found that a large number of them tended to tweet frequently about sporting events and players and teams involved. When some sports-related topics begin to trend, these users are among the early initiators of them, by virtue of their high tweet-rate. This suggests that the nature of the content plays a strong role in determining if a topic trends, rather than the users who initate it.

Audience: When we looked at the number of followers for these authors, we were surprised to find that they were almost completely uncorrelated (correlation of 0.01) with the number of trending topics, although the mean is fairly high (2481) [1]. The absence of correlation indicates that the number of followers is not an indication of influence, similar to observations in earlier work [9].

8.2 Propagators

We have observed previously that topics trend on Twitter mainly due to the propagation through the network. The main way to propagate information on Twitter is by retweeting. 31% of the tweets of trending topics are retweets. This reflects a high volume of propagation that garner popularity for these topics. Further, the number of retweets for a topic correlates very strongly (0.96) with the trend duration, indicating that a topic is of interest as long as there are people retweeting it.

Each retweet credits the original poster of the tweet. Hence, to identify the authors who are retweeted the most in the trending topics, we counted the number of retweets for each author on each topic.

[1]This is due to the fact that one of these authors has more than a million followers

Author	Retweets	Topics	Retweet-Ratio
vovo_panico	11688	65	179.81
cnnbrk	8444	84	100.52
keshasuja	5110	51	100.19
LadyGonga	4580	54	84.81
BreakingNews	8406	100	84.06
MLB	3866	62	62.35
nytimes	2960	59	50.17
HerbertFromFG	2693	58	46.43
espn	2371	66	35.92
globovision	2668	75	35.57
huffingtonpost	2135	63	33.88
skynewsbreak	1664	52	32
el_pais	1623	52	31.21
stcom	1255	51	24.60
la_patilla	1273	65	19.58
reuters	957	57	16.78
WashingtonPost	929	60	15.48
bbcworld	832	59	14.10
CBSnews	547	56	9.76
TelegraphNews	464	79	5.87
tweetmeme	342	97	3.52
nydailynews	173	51	3.39

Table 1: Top 22 Retweeted Users in at least 50 trending topics each

Domination: We found that in some cases, almost all the retweets for a topic are credited to one single user. These are topics that are entirely based on the comments by that user. They can thus be said to be dominating the topic. The *domination-ratio* for a topic can be defined as the fraction of the retweets of that topic that can be attributed to the largest contributing user for that topic. However, we observed a negative correlation of -0.19 between the domination-ratio of a topic to its trending duration. This means that topics revolving around a particular author's tweets do not typically last long. This is consistent with the earlier observed strong correlation between number of authors and the trend duration. Hence, for a topic to trend for a long time, it requires many people to contribute actively to it.

Influence: On the other hand, we observed that there were authors who contributed actively to many topics and were retweeted significantly in many of them. For each author, we computed the ratio of retweets to topics which we call the *retweet-ratio*. The list of influential authors who are retweeted in at least 50 trending topics is shown in Table 1. We find that a large portion of these authors are popular news sources such as CNN, the New York Times and ESPN. This illustrates that social media, far from being an alternate source of news, functions more as a filter and an amplifier for interesting news from traditional media.

9. CONCLUSIONS

To study the dynamics of trends in social media, we have conducted a comprehensive study on trending topics on Twitter. We first derived a stochastic model to explain the growth of trending topics and showed that it leads to a lognormal distribution, which is validated by our empirical results. We also have found that most topics do not trend for long, and for those that are long-trending, their persistence obeys a geometric distribution.

When we considered the impact of the users of the network, we discovered that the number of followers and tweet-rate of users are not the attributes that cause trends. What proves to be more important in determining trends is the retweets by other users, which is more related to the content that is being shared than the attributes of the users. Furthermore, we found that the content that trended was largely news from traditional media sources, which are then amplified by repeated retweets on Twitter to generate trends.

10. REFERENCES

[1] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the Influential Bloggers in a Community. *WSDM'08*, 2008.

[2] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, December 2009.

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Fourth International AAAI Conference on Weblogs and Social Media*, May 2010.

[4] W. Galuba, D. Chakraborty, K. Aberer, Z. Despotovic, and W. Kellerer. Outtweeting the Twitterers - Predicting Information Cascades in Microblogs. In *3rd Workshop on Online Social Networks (WOSN 2010)*, 2010.

[5] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *ArXiv e-prints*, December 2008, 0812.1045.

[6] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci.*, 60(11):2169–2188, 2009.

[7] M. E. McCombs and D. L. Shaw. The Evolution of Agenda-Setting Research: Twenty Five Years in the Marketplace of Ideas. *Journal of Communication*, (43 (2)):68–84, 1993.

[8] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1:226–251, 2004.

[9] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. Influence and passivity in social media. In *20th International World Wide Web Conference (WWW'11)*, 2011.

[10] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, 2010.

[11] F. Wu and B. A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences of the United States of America*, 104(45):17599–17601, November 2007.

[12] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 177–186, 2011.