Trends in Social Media: Persistence and Decay

Sitaram Asur* and **Bernardo A. Huberman*** and **Gabor Szabo*** and **Chunyan Wang†**

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 find that the resonance of the content with the users of the social network plays a major role in causing trends. Also, we observe that a majority of the content propagated to cause trends arise from traditional media sources with social media acting as a selective amplifier for them.

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 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 (Wu and Huberman 2007). Another factor responsible for their decay is the competitive nature of the medium. As content starts propagating through 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 reason 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 (Romero et al. 2011). 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 (McCombs and Shaw 1993).

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. We first analyze the distribution of the number of tweets across trending topics. We observe that they are characterized by a clear log-normal distribution, similar to that found in other networks such as Digg and which is generated by a stochastic multiplicative process (Wu and Huberman 2007). 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 long trends are characterized by the resonating nature of the content, which is found to arise mainly from traditional media sources, and subsequently amplified by chains of retweets from many users in social media.

Related work

There has been prior work on analyzing connections on Twitter. Huberman et al. (2008) studied social interactions on Twitter to reveal that the driving process for usage is

*Social Computing Lab, HP Labs, Palo Alto, California, USA
†Dept of Applied Physics, Stanford University, USA

a sparse hidden network underlying the friends and followers, while most links represent meaningless interactions. Jansen et al. (2009) 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. (2010) proposed a propagation model that predicts which users will tweet which URL based on the history of past activity.

Yang and Leskovec (2011) 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. (2008) 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, (2009) have distinguished the effects of homophily from influence as motivators for propagation. Recently, Romero and others (2011) 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.

Twitter Trends

Twitter is an extremely popular online microblogging service consisting of close to 200 million users. Each user submits periodic status updates, known as *tweets*, that consist of short messages limited in size to 140 characters. 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. *Trending topics* are presented as a list by Twitter on their main Twitter.com site, and are those keywords 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. To obtain the dataset of trends for this study, we repeatedly used the Twitter Search 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, we obtained 16.32 million tweets on 3361 different topics over 40 days in Sep-Oct 2010.

Growth of Trends

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).

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 (Mitzenmacher 2003). 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 and i.i.d. as a function of t with mean 1 and variance σ^2. $\gamma(t)$ is introduced to account for the novelty decay (Wu and Huberman 2007). 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, then take the logarithm of both sides:

$$\ln N_q(t) - \ln N_q(0) = \sum_{s=1}^{t} \ln\left[1 + \gamma(s)\xi(s)\right] \tag{3}$$

Here $N_q(0)$ is the initial number of tweets in the earliest time step. The RHS of Eq. (3) 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.

In other words, we expect from this model that $\ln\left[N_q(t)/N_q(0)\right]$ 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 considered is big enough, the log-normal property is observed.

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.

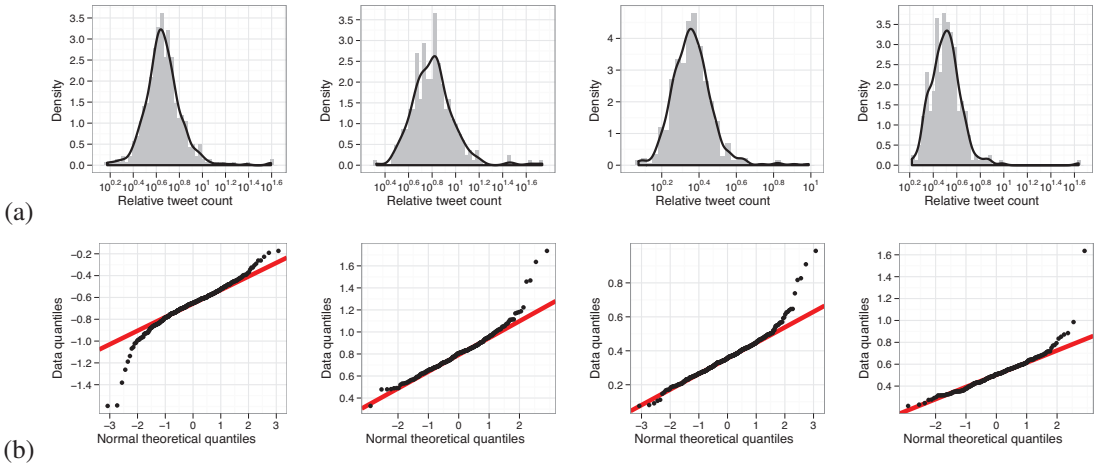

Figure 1: **(a) The densities of the ratios between cumulative tweet counts measured in two respective time frames. From left to right in the figure, the indices of the time frames between which the ratios were taken are: (2, 10), (2, 14), (4, 10), and (4, 14), respectively. The horizontal axis has been 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.**

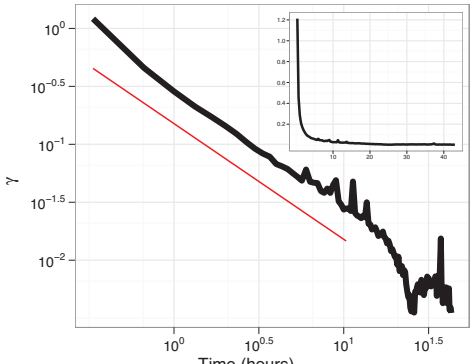

Figure 2: **The decay factor $\gamma(t)$ in time as measured using Eq. (4). The log-log plot exhibits that it decreases in a power-law fashion, with an exponent that is measured to be exactly -1 (the linear regression on the logarithmically transformed data fits with $R^2 = 0.98$). The inset displays the same $\gamma(t)$ function on standard linear scales.**

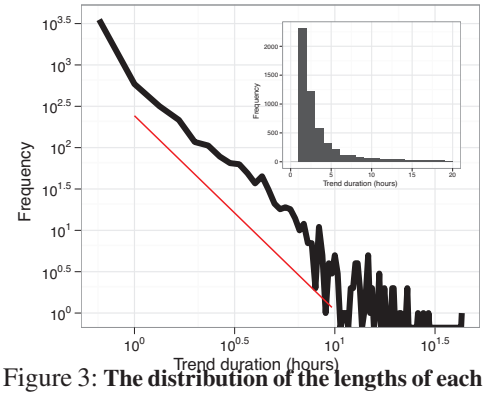

Figure 3: **The distribution of the lengths of each sequence. Both graphs are shown in the log-log scale with the inset giving the actual histograms in the linear scale.**

Thus averaging over the fractions between consecutive tweet counts yields $\gamma(t)$:

$$\gamma(t) = \left\langle \frac{N_q(t)}{N_q(t-1)} \right\rangle_q - 1. \quad (4)$$

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 $\gamma(t) \sim 1/t$.

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. 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. This could be because time zones 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 3(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.

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 ($\geq$ 50 trending topics)**

Relation to authors and activity

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. The main way to propagate information on Twitter is by retweeting. We found that 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.

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. 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 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.

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 log-normal distribution, which is validated by our empirical results. We also have found that most topics do not trend for long on Twitter. When we considered the impact of the users of the network, we discovered that 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.

References

Agarwal, N.; Liu, H.; Tang, L.; and Yu, P. S. 2008. Identifying the influential bloggers in a community. *Proceedings of the international conference on Web search and web data mining* 207–218.

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

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

Huberman, B. A.; Romero, D. M.; and Wu, F. 2008. Social networks that matter: Twitter under the microscope. *CoRR* abs/0812.1045.

Jansen, B.; Zhang, M.; Sobel, K.; and Chowdury, A. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*.

McCombs, M. E., and Shaw, D. L. 1993. The evolution of agenda-setting research: Twenty five years in the marketplace of ideas. *Proc. Journal of Communication* (43):68–84.

Mitzenmacher, M. 2003. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* 1(2).

Romero, D. M.; Galuba, W.; Asur, S.; and Huberman, B. A. 2011. Influence and passivity in social media. *Proceedings of the 20th international conference companion on World wide web* 113–114.

Wu, F., and Huberman, B. A. 2007. Novelty and collective attention. *Proc. Natl. Acad. Sci. USA* 104(45):17599–17601.

Yang, J., and Leskovec, J. 2011. Patterns of temporal variation in online media. In *ACM International Conference on Web Search and Data Minig (WSDM)*. Stanford InfoLab.