Wang and Huberman *EPJ Data Science* 2012, **1**:2
http://www.epjdatascience.com/content/1/1/2

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Long trend dynamics in social media

Chunyan Wang[1*] and Bernardo A Huberman[2*]

*Correspondence:
chunyan@stanford.edu;
bernardo.huberman@hp.com
[1]Department of Applied Physics, Stanford University, Stanford, CA, USA
[2]Social Computing Lab, HP Labs, Palo Alto, California, USA

Abstract

A main characteristic of social media is that its diverse content, copiously generated by both standard outlets and general users, constantly competes for the scarce attention of large audiences. Out of this flood of information some topics manage to get enough attention to become the most popular ones and thus to be prominently displayed as trends. Equally important, some of these trends persist long enough so as to shape part of the social agenda. How this happens is the focus of this paper. By introducing a stochastic dynamical model that takes into account the user's repeated involvement with given topics, we can predict the distribution of trend durations as well as the thresholds in popularity that lead to their emergence within social media. Detailed measurements of datasets from Twitter confirm the validity of the model and its predictions.

1 Introduction

The past decade has witnessed an explosive growth of social media, creating a competitive environment where topics compete for the attention of users [1, 2]. A main characteristic of social media is that both users and standard media outlets generate content at the same time in the form of news, videos and stories, leading to a flood of information from which it is hard for users to sort out the relevant pieces to concentrate on [3, 4]. User attention is critical for the understand of how problems in culture, decision making and opinion formation evolve [5–7]. Several studies have shown that attention allocated to on-line content is distributed in a highly skewed fashion [8–11]. While most documents receive a negligible amount of attention, a few items become extremely popular and persist as public trends for long a period of time [12–14]. Recent studies have focused on the dynamical growth of attention on different kinds of social media, including Digg [15–17], Youtube [18], Wikipedia [19–21] and Twitter [14, 22–24]. The time-scale over which content persists as a topic in these media also varies on a scale from hours to years. In the case of news and stories, content spreads on the social network until its novelty decays [15]. In information networks like Wikipedia, where a document remains alive for months and even years, popularity is governed by bursts of sudden events and is explained by the rank shift model [19].

While previous work has successfully addressed the growth and decay of news and topics in general, a remaining problem is why some of the topics stay popular for longer periods of time than others and thus contribute to the social agenda. In this paper, we focus on the dynamics of long trends and their persistence within social media. We first introduce a dynamic model of attention growth and derive the distribution of trend durations for

all topics. By analyzing the resonating nature of the content within the community, we provide a threshold criterion that successfully predicts the long term persistence of social trends. The predictions of the model are then compared with measurements taken from Twitter, which as we show provides a validation of the proposed dynamics.

This paper is structured as follows. In Section 2 we describe our model for attention growth and the persistence of trends. Section 3 describes the data-set and the collection strategies used in the study, whereas Section 4 discusses the measurements made on data-sets from Twitter and compares them with the predictions of the model. Section 5 concludes with a summary of our findings and future directions.

2 Model

On-line micro-blogging and social service websites enable users to read and send text-based messages to certain topics of interest. The popularity of these topics is commonly measured by the number of postings about these topics [15, 19]. For instance on Twitter, Digg and Youtube, users post their thoughts on topics of interest in the form of tweets and comments. One special characteristic of social media that has been ignored so far is that users can contribute to the popularity of a topic more than once. We take this into account by denoting first posts on a certain topic from a certain user by the variable First Time Post (*FTP*). If the same user posts on the topic more than once, we call it a Repeated Post (*RP*). In what follows, we first look at the growth dynamics of *FTP*.

When a topic first catches people's attention, a few people may further pass it on to others in the community. If we denote the cumulative number of *FTP* mentioning the topic at time t by N_t, the growth of attention can be described by $N_t = (1 + \chi_t)N_{t-1}$, where the χ_t are assumed to be small, positive, independent and identically distributed random variables with mean μ and variance σ^2. For small χ_s, the equation can be approximated as:

$$N_t \simeq \prod_{s=1}^{t} e^{\chi_s} N_0 = e^{\sum_{s=1}^{t} \chi_s} N_0. \tag{1}$$

Taking logarithms on both sides, we obtain $\log \frac{N_t}{N_0} = \sum_{s=1}^{t} \chi_s$. Applying the central limit theorem to the sum, it follows that the cumulative count of *FTP* should obey a log-normal distribution.

We now consider the persistence of social trends. We use the variable vitality, $\phi_t = \frac{N_t}{N_{t-1}}$, as a measurement of popularity, and assume that if the vitality of a topic falls below a certain threshold θ_1, the topic stops trending. Thus

$$\log \phi_t = \log \frac{N_t}{N_{t-1}} = \log \frac{N_t}{N_0} - \log \frac{N_{t-1}}{N_0} \simeq \chi_t. \tag{2}$$

The probability of ceasing to trend at the time interval s is equal to the probability that ϕ_s is lower than a threshold value θ_1, which can be written as:

$$\begin{aligned} p &= \Pr(\phi_s < \theta_1) = \Pr(\log \phi_s < \log(\theta_1)) \\ &= \Pr(\chi_s < \log(\theta_1)) = F(\log(\theta_1)), \end{aligned} \tag{3}$$

where $F(x)$ is the cumulative distribution function of the random variable χ. We are thus able to determine the threshold value from $\theta_1 = e^{F^{-1}(p)}$ if we know the distribution of the

random variable χ. Notice that if χ is independent and identically distributed, it follows that the distribution of trending durations is given by a geometric distribution with $\Pr(L = k) = (1-p)^k p$. The expected trending duration of a topic, $E(L)$, is therefore given by

$$E(L) = \sum_{0}^{\infty} (1-p)^k p \cdot k = \frac{1}{p} - 1 = \frac{1}{F(\log(\theta_1))} - 1. \tag{4}$$

Thus far we have only considered the impact of *FTP* on social trends by treating all topics as identical to each other. To account for the resonance between users and specific topics we now include the *RP* into the dynamics. We define the instantaneous number of *FTP* posted in the time interval t as FTP_t, and the repeated posts, *RP*, in the time interval t as RP_t. Similarly we denote the cumulative number of all posts-including both *FTP* and *RP*-as S_t. The resonance level of fans with a given topic is measured by $\mu_t = \frac{FTP_t + RP_t}{FTP_t}$, and we define the expected value of μ_t, $E(\mu_t)$ as the active-ratio a_q.

We can simplify the dynamics by assuming that μ_t is independent and uniformly distributed on the interval $[1, 2a_q - 1]$. It then follows that the increment of S_t is given by the sum of FTP_t and RP_t. We thus have

$$S_t - S_{t-1} = FTP_t + RP_t = \mu_t FTP_t = \mu_t (N_t - N_{t-1}) = \mu_t \chi_t N_{t-1}. \tag{5}$$

And also

$$\begin{aligned} E_\mu(S_t) &= E_\mu(S_{t-1}) + a_q(N_t - N_{t-1}) \\ &= E_\mu(S_{t-2}) + a_q(N_t - N_{t-2}) = \cdots \\ &= E_\mu(S_0) + a_q(N_t - N_0) = a_q N_t. \end{aligned} \tag{6}$$

We approximate S_{t-1} by $\mu_t N_{t-1}$. Going back to Equation 5, we have

$$S_t \simeq \mu_t(\chi_t + 1)N_{t-1} \simeq \mu_t e^{\chi_t} N_{t-1}. \tag{7}$$

From this, it follows that the dynamics of the full attention process is determined by the two independent random variables, μ and χ. Similarly to the derivation of Equation 3, the topic is assumed to stop trending if the value of either one of the random variables governing the process falls below the thresholds θ_1 and θ_2, respectively. One point worth mentioning here is that, θ_1 and θ_2 are system parameters, i.e. not dependent on the topic, but only on the studied medium. The probability of ceasing to trend, defined as $p^\star$, is now given by

$$p^\star = \Pr(\chi_t < \log(\theta_1)) \Pr(\mu_t < \theta_2) = \frac{\theta_2 - 1}{2(a_q - 1)} p, \tag{8}$$

$p = F(\log(\theta_1))$. The expected value of L_q for any topic q is given by

$$E(L_q) = \frac{2(a_q - 1)}{F(\log\theta_1)(\theta_2 - 1)} - 1. \tag{9}$$

Which states that the persistent duration of trends associated with given topics is expected to scale linearly with the topic users' active-ratio. From this result it follows that one can

predict the trend duration for any topic by measuring its user active-ratio after the values of θ_1 and θ_2 are determined from empirical observations.

3 Data

To test the predictions of our dynamic model, we analyzed data from Twitter, an extremely popular social network website used by over 200 million users around the world. Its interface of allows users to post short messages, known as tweets, that can be read and retweeted by other Twitter users. Users declare the people they follow, and they get notified when there is a new post from any of these people. A user can also forward the original post of another user to his followers by the re-tweet mechanism.

In our study, the cumulative count of tweets and re-tweets that are related to a certain topic was used as a proxy for the popularity of the topic. On the front page of Twitter there is also a column named trends that presents the few keywords or sentences that are most frequently mentioned in Twitter at a given moment. The list of popular topics in the trends column is updated every few minutes as new topics become popular. We collected the topics in the trends column by performing an API query every 20 minutes. For each of the topics in the trending column, we used the Search API function to collect the full list of tweets and re-tweets related to the topic over the past 20 minutes. We also collected information about the author of the post, identified by a unique user-id, the text of the post and the time of its posting. We thus obtained a dataset of 16.32 million posts on 3361 different topics. The longest trending topic we observed had a length of 14.7 days. We found that of all the posts in our dataset, 17% belonged to the *RP* category.

4 Results

We start by analyzing the distribution of N from our data-set. We found out that N_{10} follows a log-normal distribution, as can be seen from Figure 1. The Q-Q plot in Figure 1 follows a straight line. Different values of t yield similar results. The Kolmogorov-Smirnov normality test of $\log(N_{10})$ with mean 3.5577 and standard deviation 0.3266 yields a p-value of 0.0838. At a significance level of 0.05, the test fails to reject the null hypothesis that $\log(N_{10})$ follows normal distribution, a result which is consistent with Equation 1.

We also measured the distribution of χ from $\chi_t = \frac{N_t}{N_{t-1}} - 1$. We found that $\log(\chi)$ follows a normal distribution with mean equal to –1.4522 and a standard deviation value of 0.6715,

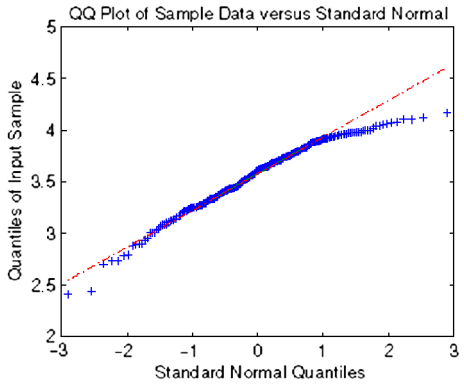

Figure 1 Q-Q plot of $\log(N_{10})$. The straight line shows that the data follows a lognormal distribution with a slightly shorter tail.

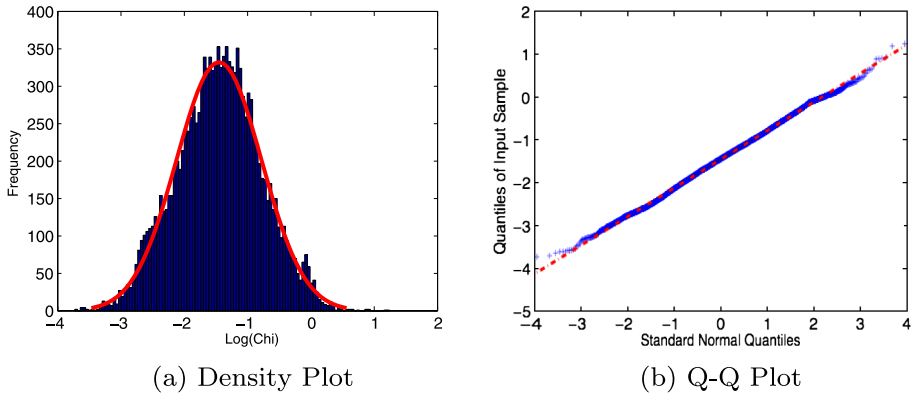

Figure 2 Density plot and Q-Q plot of log(χ). (a) Density plot of log(χ) over different t and social trends. **(b)** Q-Q plot of log(χ).

as shown in Figure 2. The Kolmogorov-Smirnov normality test statistic gives a high p-value of 0.5346. The mean value of χ is 0.0353, which is small for the approximations in Equation 1 and Equation 7 to be valid. We also examined the record breaking values of vitality, $\phi_t = \chi_t + 1$, which signal the behavior of the longest lasting trends. From the theory of records, if the values ϕ_t come from an independent and identical distribution, the number of records that have occurred up to time t, defined as $R_n(t)$, should scale linearly with $\log(t)$ [26, 27]. As is customary, we say that a new record has been established if the vitality of the trend at the moment is longer than all of the previous observations. As shown in Figure 3, there is a linear scaling relationship between $\log(t)$ and $R_n(t)$ for a sample topic "Kim Chul Hee". The topic kept trending for 14 days on Twitter in September 2010. Similar observations are repeated for other different topics on Twitter. One implication of this observation is that confirms the validity of our assumption that the values of $\chi_1, \chi_2, \ldots, \chi_t$ are independent and identically distributed.

Next we turn our attention to the distribution of durations of long trends. As shown in Figure 4 and Figure 5, a linear fit of trend duration as a function of density in a logarithmic scale suggests an exponential family, which is consistent with Equation 4. The red line in Figure 4 gives a linear fitting with R-square 0.9112. From the log-log scale plot in Figure 5, we observe that the distribution deviates from a power law, which is a characteristic of so-

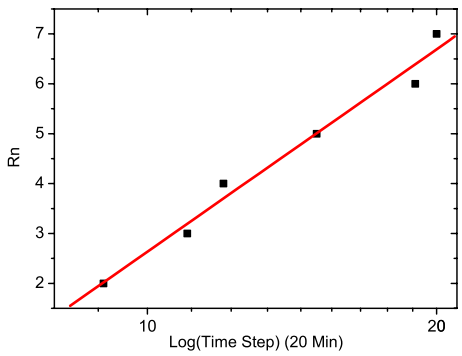

Figure 3 The linear scaling relationship between $R_n(t)$ and $\log(t)$ of topic 'Kim Chul Hee', a Korean pop star. The number of records that have occurred up to time t scales linearly with $\log(t)$.

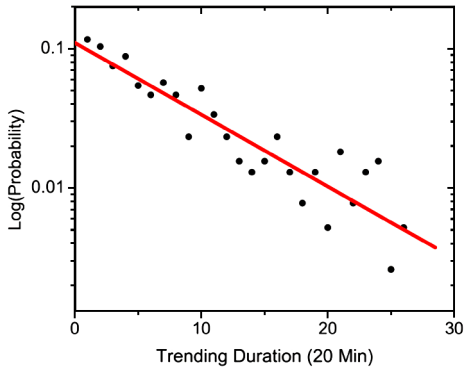

Figure 4 Semi-log plot of trending duration density. The straight line suggests an exponential family of the trending time distribution.

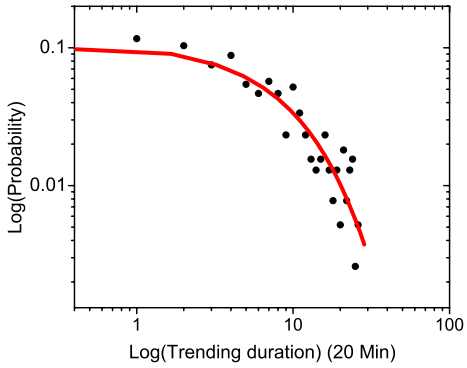

Figure 5 Density plot of trending duration in log-log scale. The distribution of duration deviates from a power law.

cial trends that originate from news on social media [25]. From the distribution of trending times, p is estimated to have a value of 0.12. Together with the measured distribution of χ and Equation 3, we can estimate the value of θ to be 1.0132.

We can also determine the expected duration of trend times stemming from the impact of active-ratio. The frequency count of active-ratios over different topics is shown in Figure 6, with a peak at $a_q = 1.2$. This observation suggests that while the ratio is centered around 1.2 for the majority of topics, there are a few topics obtain large amount of repeated attention. This observation may shadow light on existing observations about the highly skewed distribution in attention dynamic studies. As can be seen in Figure 7, the trend duration of different topics scales linearly with the active-ratio, which is consistent with the prediction of Equation 9. The R-square of the linear fitting has a value of 0.98664. From the slope of the linear fit and $\theta_1 = 1.0132$, and Equation 9 we obtain a value for $\theta_2 = 1.153$. With the value of θ_1 and θ_2, we are able to predict the expected trend duration of any given topic based on measurements of its active-ratio.

5 Discussion and conclusion

In this paper we investigated the persistence dynamics of trends in social media. By introducing a stochastic dynamic model that takes into account the user's repeated involve-

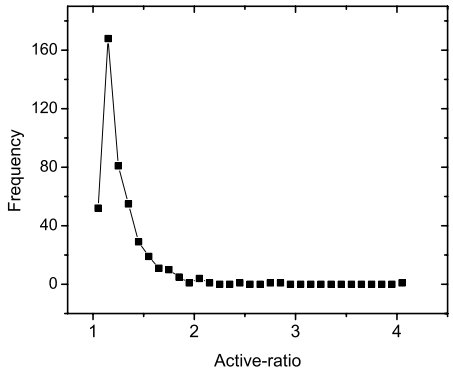

Figure 6 Frequency count of active-ratio over all topics. The maximum ratio is 1.2 among all topics.

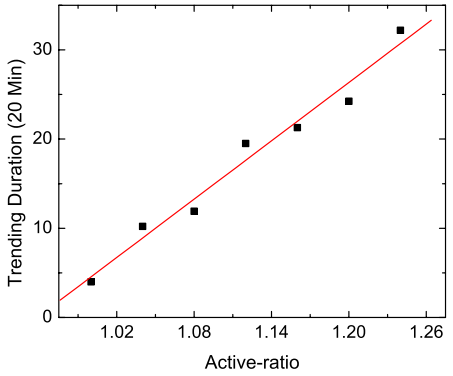

Figure 7 Linear relationship between trending duration and active-ratio in good agreement with the predictions of model.

ment with given topics, we are able to predict the distribution of trend durations as well as the thresholds in popularity that lead to the emergence of given topics as trends within social media. The predictions of our mode were confirmed by a careful analysis of a data from Twitter. Furthermore, a linear relationship between the resonance level of users with given topics, and the trending duration of a topic was derived. The proposed model provides a deeper understanding the popularity of on-line contents. Parameters θ_1 and θ_2 in our model are system specific and could be calculated from hidden algorithms when applying our model to other on-line social media websites. Possible refinements may include the effect of competition between topics, sudden burst of events, the effect of marketing campaigns and the actively censoring of specific topics [28]. In closing, we note that although the focus in this paper has been on trend dynamics that are featured on social media websites, the framework and model may be suitable to other types of content and off-line trends. The issue raised - that is, trending phenomenon under the impact of user's repeated involvement - is therefore a general one and should provide ample opportunities for future work.

Competing interests
The authors declare that they have no competing interests.

Author contributions
B.H. and C.W. designed the study and performed research. C.W. and B.H. wrote the paper. All authors read and approved the final manuscript.

Acknowledgements
We acknowledge useful discussions with S. Asur and G. Szabo. C.W. would like to thank HP Labs for financial support.

Received: 23 January 2012 Accepted: 18 May 2012 Published: 18 May 2012

References

1. McCombs ME, Shaw DL (1993) The evolution of agenda setting research: twenty five years in the marketplace of ideas. Journal of Communication 43(2):68-84
2. Falkinger J (2008) Limited attention as a scarce resource in information-rich economies. Econ J (Lond) 118(532):1596-1620
3. Agichtein E, Castillo C, Donato D, Gionis A, Mishne G (2008) Finding high-quality content in social media. In: Proceedings of the international conference on Web search and web data mining (WSDM)
4. Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of Social Media. Bus Horiz 53(1):59-68
5. Zhu J-H (1992) Issue competition and attention distraction: a zero-sum theory of agenda setting. Journal Q 69:825-836
6. Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. Science 316(5827):1036-1039
7. Guimerà R, Uzzi B, Spiro J, Amaral LAN (2005) Team assembly mechanisms determine collaboration network structure and team performance. Science 308(5722):697-702
8. Huberman BA, Pirolli PLT, Pitkow JE, Lukose RM (1998) Strong regularities in world wide web surfing. Science 280(5360):95-97
9. Johansen A, Sornette D (2000) Download relaxation dynamics on the WWW following newspaper publication of URL. Physica A 276(1-2):338-345
10. Huberman BA (2001) The laws of the web: patterns in the ecology of information. MIT Press, Massachusetts
11. Vázquez A et al (2006) Modeling bursts and heavy tails in human dynamics. Phys Rev E 73:036127
12. Neuman WR (1990) The threshold of public attention. Public Opin Q 54:159-176
13. Klamer A, Van Dalen HP (2002) Attention and the art of scientific publishing. J Econ Methodol 9(3):289-315
14. Becker H, Naaman M, Gravano L (2011) Beyond trending topics: real-world event identification on Twitter. In: Proceedings of 15th international conference on Weblogs and Social Media (ICWSM).
15. Wu F, Huberman BA (2007) Novelty and collective attention. Proc Natl Acad Sci USA 105:17599
16. Leskovec J, Backstrom L, Kleinberg J (2009) Meme-tracking and the dynamics of the news cycle. International conference on knowledge discovery and data mining (KDD)
17. Lerman K, Hogg T (2010) Using a model of social dynamics to predict popularity of news. In: Proceedings of 19th international World Wide Web conference (WWW)
18. Crane R, Sornette D (2008) Robust dynamic classes revealed by measuring the response function of a social system. Proc Natl Acad Sci USA 105:15649
19. Ratkiewicz J, Fortunato S, Flammini A, Menczer F, Vespignani A (2010) Characterizing and modeling the dynamics of online popularity. Phys Rev Lett 105:158701
20. Capocci A, Servedio VDP, Colaiori F, Buriol LS, Donato D, Leonardi S, Caldarelli G (2006) Preferential attachment in the growth of social networks: the Internet encyclopedia Wikipedia. Phys Rev E 74:036116
21. Zlatic V, Bozicevic M, Stefancic H, Domazetl M (2006) Wikipedias: collaborative web-based encyclopedias as complex networks. Phys Rev E 74:016115
22. Jansen BJ, Zhang M, Sobel K, Chowdhury A (2009) Twitter power: tweets as electronic word of mouth. J Am Soc Inf Sci 60(11):2169-2188
23. Lee K, Palsetia D, Narayanan R, Patwary MMA, Agrawal A, Choudhary A (2011) Twitter trending topic classification. 11th IEEE international conference on data mining workshops (ICDMW)
24. Gonçalves B, Perra N, Vespignani A (2011) Modeling users' activity on Twitter networks: validation of Dunbar's number. PLoS ONE 6(8):e22656
25. Sitaram A, Huberman BA, Szabo G, Wang C (2011) Trends in Social Media: persistence and decay. In: Proceedings of 15th international conference on Weblogs and Social Media (ICWSM)
26. Redner S, Petersen MR (2006) Role of global warming on the statistics of record-breaking temperatures. Phys Rev E 74:061114
27. Krug J (2007) Records in a changing world. J Stat Mech. doi:10.1088/1742-5468/2007/07/P07001 07001
28. Sydell L (2011) How Twitter's trending algorithm picks its topics. http://www.npr.org/2011/12/07/143013503/how-twitters-trending-algorithm-picks-its-topics

doi:10.1140/epjds2
Cite this article as: Wang and Huberman: **Long trend dynamics in social media.** *EPJ Data Science* 2012 **1**:2.