The Pattern of Information Diffusion in Microblog

Dong Wang[†¶], Zhenyu Li[†], Kave Salamatian[§], Gaogang Xie[†]
[†]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[¶]Graduate School of Chinese Academy of Sciences, Beijing, China
[§]Universite de Savoie, France
{wangdong01, zyli, xie}@ict.ac.cn, kave.salamatian@univ-savoie.fr

ABSTRACT

Microblog (*e.g.* Twitter) is fast emerging social medium for information diffusion. This paper presents an analysis of the pattern of information diffusion in microblog. We have collected the data of 1,798,901 tweets from Sina microblog service- the Chinese equivalent of Twitter. Our analysis unveils that the distribution of tweets' popularity follow the stretched exponential (SE) model. Based on this observation we analyse a multiplicative cascade model for describing tweets popularity.

1. INTRODUCTION

Microblog services, such as Twitter and Sina microblog, have greatly changed the way of information dissemination. The speed and convenience of microblogs make them competitive services with classical media. With the increase of importance of microblog as a social medium for information sharing, understanding mechanisms describing how information diffuses over microblogs, and explaining how some tweets become popular, are meaningful in order to predict the evolution of this new social medium in the future.

In this paper, we make an analysis of tweet's popularity. We investigate a cascade model [1] of information propagation in microblog services. In this model that is particularly suited to describing web 2.0 services, we assume that information diffusion proceeds in an eventually random number of successive stages. Our aim is to validate the usage of such a model for describing information diffusion in microblog services.

We collected the data of 1,798,901 tweets of Sina microblog from Dec. 6th, 2010 to Jan. 1st, 2011, containing all the publicly available details such as the re-tweet number and the information of the users' participating in the re-tweeting process. Our analysis unveils that the distribution of tweets' popularity follows the stretched exponential (SE) model, instead of the expected power-law model. We show that the social cascade processes of tweets naturally converge to the SE model and the parameters can be used to estimated crucial properties of cascade. Moreover, the re-tweet number decreases exponentially with the growth of re-tweeting hop giving preliminary evidence for a simple multiplicative model.

2. THE MULTIPLICATIVE CASCADE MODEL FOR TWEET POPULARITY

A microblog user might follow another user, *i.e.*, he will receive all messages (called tweets) sent by the followed person. Followers might re-tweet some of the messages they receive to their own followers. The distance between re-tweeters and the tweet's publisher is called *hops*. The re-tweeting mechanism enables users to spread information to users that could not normally access it. Through re-tweeting hot messages can be received by tens of thousands of users. We measure popularity of a tweet by the number of people that have re-tweeted it.

The re-tweeting pattern in microblog can be described as a random multiplicative cascade process. Formally, a random multiplicative cascade process $X(.)$ can be described at each point k as a multiplication of n random variables $m_1, \ldots, m_n$, *i.e.*, $X(k) = m_1 \times m_2 \times \ldots \times m_n$. To relate this model to the propagation of information, we define that the i-th stage begins at the time for generation of the first re-tweet at hop i and ends until the first re-tweet at hop $i+1$ appears,that is, the number of cascade stages is equivalent to the maximum re-tweeting hop. The number of new users that will re-tweet a tweet at i-th $(1 < i \leq n)$ stage is a coefficient α_i of the overall number of users that have re-tweeted the tweet up to the $(i-1)$-th stage. Then the overall number of users re-tweet a particular tweet t after n stages, denoted as $N^n(t)$, is given by $N^n(t) = (1+\alpha_1)(1+\alpha_2)\ldots(1+\alpha_n)$, where the expansion coefficient α_i is in fact related to two main factors: the proportion of followers that will re-tweet at i-th stage, and the outdegree (i.e. the number of followers) of the re-tweeters.

Similarly to Central limit theorem that applies to sum of random variables, we can derive an asymptotic limit theorem for multiplicative processes where all multiplied random variables are i.i.d [2]. Such processes

ACM CoNEXT Student Workshop, December 6, 2011, Tokyo, Japan.

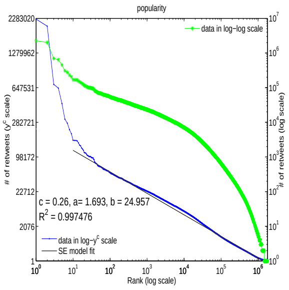

Figure 1: Stretched exponential distr. Fitting

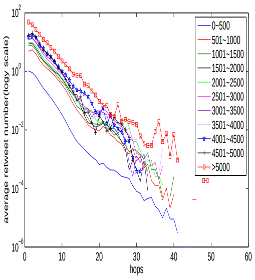

Figure 2: Distribution of avg. re-tweet number in each hop

converge to a stretched exponential (SE) distribution defined as:

$$P(X \geq x) = e^{-(\frac{x}{x_0})^c} \tag{1}$$

where the stretched factor is related to the number of multiplied random variables m, *i.e.* the number of cascade stages, through a simple relation $c = \frac{1}{m}$, and x_0 is a constant parameter that is related to ranking scale. Because of its particular shape a stretched exponential distribution can be easily mistaken with a power law [3]. However processes following a stretched exponential distribution will have a particular rank ordering statistic that will be different from the one of a power law . Let i be the rank of an observation from a stretched exponentially distributed process and y_i its observed value. It can be shown theoretically that this relation is valid

$$y_i^c = -a \log i + b \tag{2}$$

where $a = x_0^c$ and $b = y_1^c$, meaning that the modified ranking diagram, showing y_i^c the observed values with exponent c *vs.* the log of its rank, follows in a straight line with slope $a = x_0^c$. This analysis suggest that if an empirical distribution follows a streched exponential it can be meaningful to search for a multiplicative cascade that could explain the emergence of this global distribution. In order to check this, we fitted SE models to the observed tweet popularity rank-ordering distribution using the matlab toolbox provided by authors of [4]. We show in Figure 1 the popularity distribution for all collected tweets in both log-log scale and log-y^c scale. The parameters of the SE model along with the R^2 statistic of the fitting are marked in the figure. The figure shows that the SE model fits the distribution very well, except the first several points that are due to the "King effect" [3] (this resulting from the fact that popular topics reduce the attractiveness of other topic because of their high popularity). We have fitted SE for different days and list in Table 1 the obtained parameters showing the relative consistency of the c parameter in close dates.

In particular the SE model predict that we can expect a number of maximum re-tweeting hop (a number of cascade stage) around $m = \frac{1}{c}$. We also show in Table 1, the number of maximum re-tweeting hop h_e derived using SE model: $h_e = \frac{1}{c}$ and the empirically observed value over the dataset h_a. As can be observed these values are very close. These results give more rationals for a multiplicative cascade model of tweet popularity.

Table 1: Parameters of different days of tweets

Time	c	a	R^2	h_e	h_r
06/12/10	0.38	8.617	0.9980	2.63	2.69
07/12/10	0.36	6.523	0.9987	2.78	2.73
08/12/10	0.37	7.979	0.9987	2.70	2.72

To model the multiplicative cascade, we have analyzed how the number of re-tweets relates to the number of re-tweeting hops from the tweet generator. For this purpose we have stratified the data into 11 strates according to the popularity of the tweets inside it, *i.e.* for $1 \leq i \leq 10$, the i-th strate contains the tweets with re-tweet numbers between $500i$ to $500(i+1)$; the last set contains all remaining tweets. We show in Figure 2 in a semi-log scale the evolution of the average number of re-tweets as a function of its hop distance from tweet source. Interestingly all curves seems to be almost parallel and to be finely fitted to a straight lines. This indicates that the average re-tweet number decreases exponentially with the hop distance. This is compatible with a cascade model with a constant value of $E\{\alpha_i\} = \hat{\alpha}$ for all stages. Interestingly these figures mean that the tweet's popularity mainly depends on the re-tweet number at the first hop (or the two first hops), *i.e.*, the number of followers of the originator that forward the tweet. However this conclusion needs more analysis and larger space to be developed.

3. CONCLUSIONS AND FUTURE WORKS

In this paper, we gave evidence that diffusion in a microblog can be explained with a very simple multiplicative cascade model.We showed that the overall distribution of tweet popularity is compatible with this model.We even give some indication that cascade coefficient $E\{\alpha\}$ might be constant. However, we need more precise information about the diffusion process over the microblog to validate this multiplicative cascade model both microscopically (at the scale of a single tweet) and macroscopically (at the scale of global statistics). This is the aim of our forthcoming research.

4. REFERENCES

[1] M. Cha, A. Mislove, B. Adams and K. P. Gummadi. Characterizing Social Cascades in Flickr. In *Proceedings of the WOSN'08*,2008.

[2] U. Frisch and D. Sornette. Extreme deviations and applications. *Journal of Physics I France*, 1997.

[3] J. Laherrere, D. Sornette. Stretched exponential distributions in Nature and Economy: Fat tails with characteristic scales. In *European Physical Journal B 2*, 2008.

[4] A. Clauset, C. R. Shalizi, M. E.J. Newman. Power-law distributions in empirical data. In *SIAM Review 51(4)*, 2009.