Human Behavior Dynamics in Online Social Media: A Time Sequential Perspective

Xiao Yang
School of Computer Science and Technology, Harbin Institute of Technology
Harbin, China
shawcsn@gmail.com

Zhaoxin Zhang
School of Computer Science and Technology, Harbin Institute of Technology
Harbin, China
heart@hit.edu.cn

Ke Wang
Department of Mathematics, College of Science, Harbin Institute of Technology
Harbin, China
w_k@hit.edu.cn

ABSTRACT
We present an empirical analysis of human behavior dynamics in online social media from a time sequential perspective. By analyzing the users' inter-event time distribution between two consecutive actions and fitting with accurate numerical method, we find that the users' behavior pattern follows the heavy tailed or power-law distribution both in collective and individual scale, not the traditional Poisson processes hypothesis. With the estimating of the personal actions' time sequential self-similarity, we find that the individual behavior can be predicted to some extent. We propose a more fine-grained method to eliminate the effect of individual non-regular behaviors on modeling in cascading non-homogeneous Poisson process model, and the simulated fitting results are more close to the actual human behaviors' characters.

Categories and Subject Descriptors
J.4 [**Computer Applications**]: Social and Behavior Sciences; H.1 [**Information Systems**]: Models and Principles.

General Terms
Human Factors, Measurement, Theory.

Keywords
human dynamics, online social media, power-law, heavy tails, self-similarity, predictability, behavior model

1. INTRODUCTION
The dynamics of many social, technological and economic phenomena are driven by individual human actions, turning the quantitative understanding of human behavior into a central question of modern science. Recently, understanding the regularity in complex human dynamics has attracted more and more attention in various fields. The classical view has assumed that human activities are homogeneous Poisson processes [1]. Such processes have a well-known statistical property: the time interval between two consecutive events, called the inter-event time $\bar{t}$, follows an exponential distribution: $P(\bar{t}) = \lambda e^{-\lambda \bar{t}}$. This means that the individual activity pattern is usually simplified as a completely random point-process, and the time difference between two consecutive events should be almost uniform, and the long gap is hardly to be observed. However, Barabási [2]reported his empirical studies on e-mail and surface mail communication at *Nature*, which showed a far different scenario: those communication patterns follow non-Poisson statistics, characterized by bursts of rapidly occurring events separated by long gaps, and the inter-event time $\bar{t}$ follows a power-law distribution: $P(\bar{t}) \sim \bar{t}^{-\alpha}$. Indeed, an increasing number of recent measurements indicate that the timing of many human actions systematically deviate from the Poisson prediction, the waiting or inter-event times being better approximated by a heavy tailed or Pareto distribution. The heavy tails have also been observed in many human behaviors [3], including market transaction, web browsing, movie watching, short message sending and so on. The difference between a Poisson and a heavy tailed behavior is striking: the exponential decay of a Poisson distribution forces the consecutive events to follow each other at relatively regular time intervals and forbids very long waiting times. In contrast, the slowly decaying heavy tailed processes allow for very long periods of inactivity that separate bursts of intensive activity.

The increasing evidence of non-Poisson statistics of human activity pattern highlights a question: what is the origin of those heavy tails? Based on the queuing theory, Barabási proposed a simple model [3] where the individual executes the highest priority task first, and they suggested the highest-priority-first (HPF) protocol a potential origin of those heavy tails. By analyzing the coherence of task processing and periodism, Malmgren [4, 5] proposed a cascading non-homogeneous Poisson process model. But this model cannot embody the unexpected cessation or dynamic evolution in individual activities. Oliveira [6] studied the influence on behavior pattern by human memorability, and proposed a memory-effect model to explain the heavy-tails phenomenon. Yet, the effect of memory is diversiform and hardly to measure under a uniform scale. Rybski [7] attributed this scaling law to long-term correlation patterns of human activity, and provide a mathematical framework that relates the generalized version of Gibrat's law to the long-term correlated dynamics.

The rapid development of social media websites has dramatically changed the way that people communicate with each other. Research human behavior dynamics in online social media has a significant improvement of understanding personal behavior patterns in collective network. Gómez et al.[9] measured community response time in terms of comment activity on

The 6th SNA-KDD Workshop '12 (SNA-KDD'12), August 12, 2012, Beijing, China. Copyright 2012 ACM 978-1-4503-1544-9 ...$15.00.

Slashdot stories, Choudhury et al.[11] characterize conversations through their interestingness, and finally, Kumar et al.[8] modeled the dynamics of conversations with a branching process incorporating recency. Some of existing studies demonstrate that the size distribution of posts and reviews follow a heavy-tailed distribution such as Zipf's law [8] or log-normal distribution [9], another portion of the literature suggest a light-tailed one, such as negative binomial distribution [10].

In this paper, we focus on finding the regular pattern of user's dynamics of online conversations and other activities. To explain the power-law distribution of human behavior dynamics in online social media, we propose a method to apply the time series analysis methods in the human dynamics modeling, producing the time series-analysis-based cascading non-homogeneous Poisson process model. This model divides a long time series into several relatively more stable short ones through cluster analysis. Therefore it could eliminate the effect of individual non-regular behaviors on modeling, improve the simulation accuracy and make the simulated results more close to the actual human behaviors' characters displayed in the real network.

2. STATISTICAL AND EMPIRICAL RESULTS

This part will show our statistical and empirical results of users' activities in online social media such as tweeting, following, posting, comments, and so on.

2.1 Data Description

Four datasets are used in our empirical measurements. The detailed properties of each dataset are shown in Table 1. The D1 dataset is collected by using the distributed crawler technology from *Twitter* which contains social network data and the tweets data including the posted user-id and posted timestamp. The D2 dataset is obtained from an open shared data (http://larica.uniurb.it/sigsna/data/) of the real-time feed aggregator that consolidates the updates from social media and social networking websites: *FriendFeed*, and it contains the users' stream of update information and the users' social network data. The D3 dataset is downloaded from the *WISE 2012 Challenge* (http://www.wise2012.cs.ucy.ac.cy) which contains tweets and followship network data, and the original data was crawled from *Sina Weibo* (http://weibo.com), a popular micro-blogging service in China. The D4 dataset is acquired from an online data mining competition that we joined: the Track 1(http://www.kddcup2012.org) of *KDD Cup 2012*, and the original data was provided by *Tencent Weibo* (http://t.qq.com), one of the largest micro-blogging websites in China. This dataset mainly records the users' time sequential following history. All of the datasets contain the users' records of online activities and corresponding time sequential timestamps.

Table 1. Information about the datasets

Source	Users	Actions	Data-ID
Twitter	118507	10000000	D1
FriendFeed	496389	12450658	D2
WISE 2012 challenge	58655849	369797723	D3
KDD Cup 2012	2320895	73209277	D4

2.2 Statistical properties

The inter-event time plays an important role in many human collective behaviors. Individual behavior pattern can be reflected by his activities' timeline. For the convenience of narrative, we describe the user's behavior from all the datasets as the user's action, which could be the tweeting on *Twitter* and *Weibo*, updating entries on *FriendFeed*, or following a followee on *Weibo*. From all the datasets, we compute the distribution of waiting times between two entities from the same user and treat it as statistical numerical value.

First, we put our attention to study the behavior of aggregated users on a whole, by treating users as identical. We empirically measure the distribution of waiting times by collecting the time series data of all entities from users. The density plot of waiting times between two consecutive comments in a log-log scale is shown in Figure 1-2. Second, we look at the distribution of two consecutive actions from single users. To reflect the more credible statistical law of individual behavior pattern in online social network, we choose the interested target users whose actions' number is above 500. The density plot of inter-event times between two consecutive actions of single users in a log-log scale is shown in Figure 3.

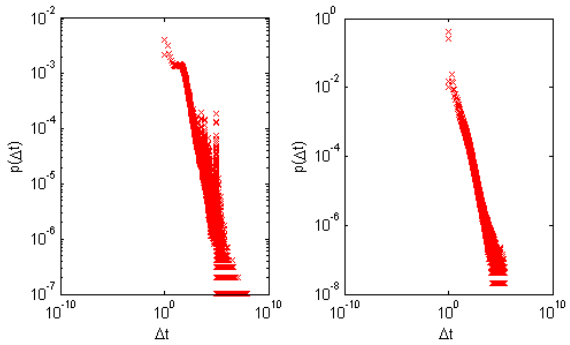

Figure 1. The total users' inter-event time distribution between two consecutive tweets at *Twitter* (left) and *Sina Weibo* (right).

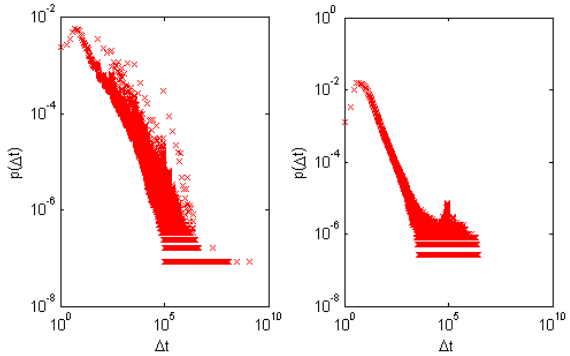

Figure 2. The total users' inter-event time distribution between two consecutive update entries at *FriendFeed* (left) and two consecutive followings at *Tencent Weibo* (right).

It is clearly seen that all the distributions follow the heavy tailed distribution both in collective and individual scale, not the traditional Poisson processes hypothesis. This result implies that, for each user, frequent actions may follow by a significantly long period of inactivity. This result reflects the heterogeneous action

and uniform law of human behavior dynamics in online social media.

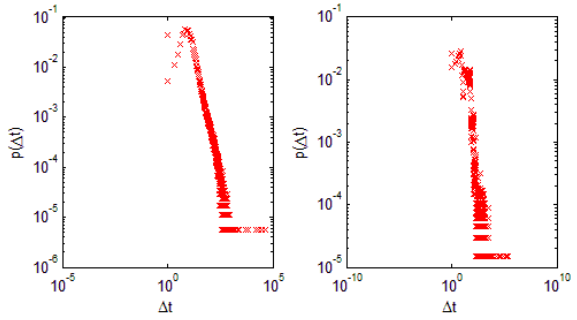

Figure 3. The two users' inter-event time distribution between two consecutive actions at *Twitter* (left) and *Sina Weibo* (right).

2.3 Fitting method

The heavy tailed distribution of human dynamics has been found in the physical world for a long time, and our empirical research finds that the human dynamics in virtual online social network also follow this heterogeneous law. The most widely used mathematical methods to quantify the heavy tailed data is the power-law distribution function. Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena. Unfortunately, the detection and characterization of power laws are complicated by the large fluctuations that occur in the tail of the distribution, the part of the distribution representing large but rare events, and by the difficulty of identifying the range over which power-law behavior holds. Newman et al.[12] found that commonly used methods for analyzing power-law data, such as least-squares fitting, can produce substantially inaccurate estimates of parameters for power-law distributions, and even in cases where such methods return accurate answers they are still unsatisfactory because they give no indication of whether the data obey a power law at all.

Mathematically, let x represent the inter-event time quantity whose distribution we are interested in. The power-law distribution is one described by a probability density $p(x)$ such that

$$p(x) = Cx^{-\alpha} \tag{1}$$

Under the standardized condition, we have

$$\int_{x_{\min}}^{\infty} Cx^{-\alpha} dx = 1 \qquad \alpha > 1 \tag{2}$$

We can find that

$$p(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha} \tag{3}$$

where α is the scaling parameter and $x_{\min}$ is the minimum value at which power-law behavior holds. The complementary cumulative distribution function (CDF) is

$$P(x) = \left(\frac{x}{x_{\min}} \right)^{-\alpha+1} \tag{4}$$

Then, let us consider the estimation of the scaling parameter α. The method of choice for fitting parametrized models such as power-law distributions to observed data is the method of maximum likelihood, which provably gives accurate parameter estimates in the limit of large sample size. Given a data set containing n observations $x_i \geq x_{\min}$, we would like to know the value of α for the power-law model that is most likely to have generated our data. The probability that the data were drawn from the model is proportional to

$$p(x \mid \alpha) = \prod_{i=1}^{n} \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha} \tag{5}$$

We can derive maximum likelihood estimators [12] (MLEs) of the scaling parameter for the discrete cases as

$$\alpha = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right]^{-1} \tag{6}$$

Based on the maximum-likelihood estimation (MLE) approach, we fit power laws to our empirical data. The scaling parameters of different datasets are shown in Table 2.

Table 2. The scaling parameters α of different datasets

Figure1	Figure2	Figure3	Figure4
0.8431	0.9079	1.0139	1.008

From the results of accurate numerical fitting method, we can find that the collective human behavior dynamics in different online social media follow the universal heavy tailed or power-law distribution only with the difference of scaling parameters. In the following, we will explore the model of this non-Poisson statistics of human behavior.

2.4 Predictability analysis

Another fundamental question is: is human behavior dynamics in online social media predictable? Song et al. [15] found that human mobility patterns are remarkably predictable and reported their findings at *Science*. Inspired by their work, in this part, we focus on addressing this problem by quantitating the self-similarity of human time sequential actions. If the occurring pattern of users' actions in online social media is similar in continuous time cycles and has long-term memorability, we can assume that the individual behavior can be predicted to some extent. The time sequential distribution of two users' inter-event time at *FriendFeed* and *Sina Weibo* is shown in Figure 4, and we can see the inter-event time sequence is self-similar in a period of time and have some repetitive pattern.

The Hurst exponent [13] is used as a measure of the long term memory of time series. It relates to the autocorrelations of the time series and the rate at which these decrease as the lag between pairs of values increases. The Hurst exponent H quantifies the relative tendency of a time series either to regress strongly to the mean or to cluster in a direction, and it has been extensively used in many fields such as the stock markets.

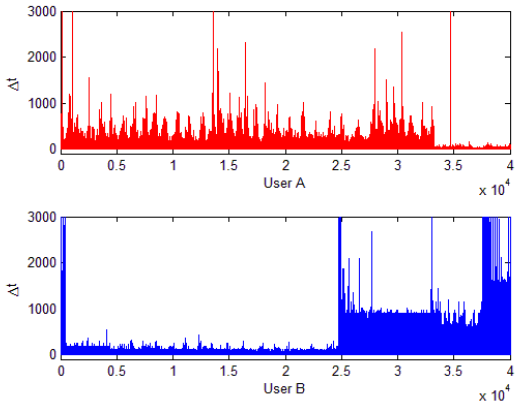

Figure 4. The time sequential distribution of two users' inter-event time at *FriendFeed* (red line) and *Sina Weibo* (blue line).

A value H in the range 0.5 < H < 1 indicates a time series with long-term positive autocorrelation. A value in the range 0 < H < 0.5 indicates a time series with long-term switching between high and low values in adjacent pairs. A value of H=0.5 can indicate a completely uncorrelated or a random series. The Hurst exponent can be calculated by rescaled range analysis [14] (R/S analysis).

We divide the user-actions time period into *M* contiguous sub-periods of length *n*, such that $M \times n = N$. Each sub-period is labeled I_a, with $a = 1, 2, ..., M$ and then each element in I_a is labeled $y_{k,a}$, with $k = 1, 2, ..., n$. For each sub-period I_a of length *n*, we calculate the time series of accumulated departures $X_{k,a}$, from the mean at a given time *k* for each sub-period I_a

$$X_{k,a} = \sum_{i=1}^{k} (y_{i,a} - \frac{1}{n}\sum_{k=1}^{n} y_{k,a}), k = 1, 2, ..., n \qquad (7)$$

Obviously, the last accumulated departure of every sub-period is 0. Thus, the average value of $(R/S)_n$ for sub-period length *n* is estimated by

$$(R/S)_n = \frac{1}{M}\sum_{a=1}^{M}\left(\frac{\max(X_{k,a}) - \min(X_{k,a})}{\sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_{k,a} - \frac{1}{n}\sum_{k=1}^{n} y_{k,a})^2}}\right) \qquad (8)$$

Where *R* represents the range that the time series covers from maximum values to minimum values within I_a, and *S* represents the standard deviation for each sub-period I_a. Repeat this process from Eq. (7) to Eq. (8) by increasing *n* over successive integer values until N/ 2. The Hurst exponent may be estimated using least square regression of the following form

$$\log(R/S)_n = H * \log n + \log c + \varepsilon_t \qquad (9)$$

Where H is the Hurst exponent and can be estimated as the slope of the equation. We calculated the Hurst exponent of time series shown in Figure 4. The Hurst exponent of user A is 0.774, and the Hurst exponent of user B is 0.698. Thus the behaviors of users in online social media don't obey random walk but exhibit long-term stability and correlation. So maybe we can predict the user's online social behavior by using the predicting and analysis tools from stock market based on time series analysis technology.

3. MODEL

Currently the research of human dynamics is not mature yet, and different human dynamic models give un-unified explanations for the non-Poisson attributions of human behaviors. The hypothesis in the priority-based queuing model [3] that users share a universal behavior pattern is unreasonable; besides the cascading non-homogeneous Poisson process [4] cannot embody the unexpected cessation or dynamic evolution in individual activities. Homogeneous Poisson processes have two well-known statistical properties: the time between consecutive events, the inter-event time τ, follows an exponential distribution, $p(\tau) = \rho e^{-\rho\tau}$, and the number of events N_T during a time interval of duration *T* time units follows a Poisson distribution with mean ρT. Malmgren et al. [4] proposed a model of human dynamics that incorporates the hypothesized periodic and cascading features of human activity. They accounted for periodic activity with a primary process, which their model as a nonhomogeneous Poisson process. Whereas a homogeneous Poisson process has a constant rate ρ, a nonhomogeneous Poisson process has a rate $\rho(t)$ that depends on time. In their model, the rate $\rho(t)$ depends on time in a periodic manner; that is, $\rho(t) = \rho(t+W)$, where *W* is the period of the process. They related the rate of the nonhomogeneous Poisson process to the daily and weekly distributions of active interval initiation

$$\rho(t) = N_w p_d(t) p_w(t) \qquad (10)$$

To apply this model in the empirical data, we first need to estimate the parameters of the model from the data. By using this double-layer cascading Poisson process can generate corresponding power-law distribution of the specific person. While in the practical experiment we found that this model failed to fit the user's activities when user's behavior pattern has some sudden burst or changes over time. So this coarse-grained time series partition method cannot cover the individual irregular actions and diversity of behavior. We propose a method to apply the cluster-analysis methods in the human dynamics modeling, producing the clustering-analysis-based cascading non-homogeneous Poisson process model. This model divides a long time series into several relatively more stable short ones through cluster analysis. Therefore it could eliminate the effect of individual non-regular behaviors on modeling. First, divide and gather the user's inter-event time distribution as list of vectors, then calculate the Euclidean distance between each time vectors pair. Then by using the hierarchical clustering algorithm, we seek to build a hierarchy of time interval clusters. In each clustering group, use the cascading non-homogeneous Poisson process to fit the internal user's behavior pattern. By using the between-groups linkage clustering method, we can get the hierarchical partition of periodicity in original data. Most users' list of vectors, or the

behavior pattern, can be divided into 2 to 5 parts. The simulated fitting results of two users' inter-event time cumulative distribution between two consecutive actions from Figure 3 is shown in Figure 5, and it's obvious that this method can improve the simulation accuracy and make the simulated results more close to the actual human behaviors' characters displayed in the real network.

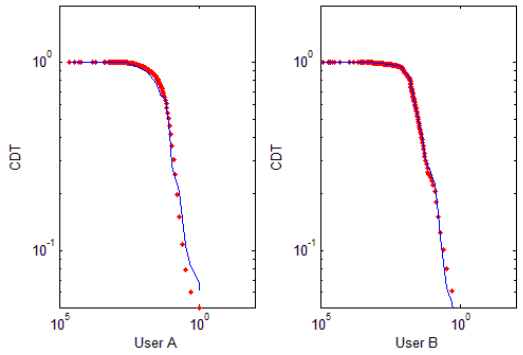

Figure 5. The simulated fitting results of two users' inter-event time cumulative distribution.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we investigated statistical properties of human behavior dynamics in online social media. By analyzing the users' inter-event time distribution between two consecutive actions, we find the heavy tailed phenomenon of users' behavior pattern in different online social network. To more exactly quantify the heavy tailed data as the power-law distribution function, we introduce the maximum-likelihood estimation (MLE) approach to estimate the accuracy of fitting and get the corresponding scaling parameters. We use the Hurst exponent to represent the self-similarity of user's actions, and the result shows that the behaviors of users in online social media don't obey random walk but exhibit long-term stability and correlation, which means that the individual behavior pattern is predictable. Then, we analyzed the disadvantage and shortage of the cascading non-homogeneous Poisson process model. To more comprehensively cover the individual irregular actions and diversity of behavior, we propose a method to apply the cluster-analysis methods in the human dynamics modeling, producing the clustering-analysis-based cascading non-homogeneous Poisson process model. The simulated fitting results show that this model can improve the simulation accuracy. For future work, we have made some empirical analysis of human behavior dynamics under the influence of social network. We will try to build the predicting model of users' action and topic evolution under the time series analysis framework, build convincing models to cover human behavior dynamics both in collective and individual scale of online social media under a universal framework, and answer the question about how social networks shape human behavior and how human behaviors shape the social network.

5. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grant No.61100189, 61003261). We also appreciate the reviewers' precious comments.

6. REFERENCES

[1] Reynolds, P. Call Center Staffing (The Call Center School Press, Lebanon, Tennessee, 2003).

[2] Barabasi, A.L. 2005. The origin of bursts and heavy tails in human dynamics. Nature. 435, 7039 (2005), 207–211.

[3] Vázquez, A., Oliveira, J.G., Dezsö, Z., Goh, K.I., Kondor, I. and Barab'asi, A.L. 2006. Modeling bursts and heavy tails in human dynamics. Physical Review E. 73, 3 (2006), 036127.

[4] Malmgren, R.D., Stouffer, D.B., Motter, A.E. and Amaral, L.A.. 2008. A poissonian explanation for heavy tails in e-mail communication. Proceedings of the National Academy of Sciences. 105, 47 (2008), 18153.

[5] Malmgren, R.D., Stouffer, D.B., Campanharo, A.S.L.. and Amaral, L.A.. 2009. On universality in human correspondence activity. Science. 325, 5948 (2009), 1696–1700.

[6] Oliveira, J.G. and Vazquez, A. 2009. Impact of interactions on human dynamics. Physica A: Statistical Mechanics and its Applications. 388, 2-3 (2009), 187–192.

[7] Rybski, D., Buldyrev, S.V., Havlin, S., Liljeros, F. and Makse, H.A. 2009. Scaling laws of human interaction activity. Proceedings of the National Academy of Sciences. 106, 31 (2009), 12640–12645.

[8] Kumar, R., Mahdian, M. and McGlohon, M. 2010. Dynamics of conversations. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (2010), 553–562.

[9] Gómez, V., Kaltenbrunner, A. and L'opez, V. 2008. Statistical analysis of the social network and discussion threads in slashdot. Proceedings of the 17th international conference on World Wide Web (2008), 645–654.

[10] Tsagkias, M., Weerkamp, W. and De Rijke, M. 2009. Predicting the volume of comments on online news stories. Proceedings of the 18th ACM conference on Information and knowledge management (2009), 1765–1768.

[11] De Choudhury, M., Sundaram, H., John, A. and Seligmann, D.D. 2009. What makes conversations interesting?: themes, participants and consequences of conversations in online social media. Proceedings of the 18th international conference on World wide web (2009), 331–340.

[12] Clauset, A., Shalizi, C.R. and Newman, M.E.J. 2009. Power-Law Distributions in Empirical Data. SIAM Rev. 51, 4 (Nov. 2009), 661–703.

[13] H. E. Hurst, R. O. Black, and Y. M. Simaika, Long Term Storage: An Experimental Study (Constable, London, 1965).

[14] Qian, B. and Rasheed, K. 2004. Hurst exponent and financial market predictability. Proceedings of The 2nd IASTED international conference on financial engineering and applications (2004), 203–209.

[15] Song, C., Qu, Z., Blumm, N. and Barabási, A.L. 2010. Limits of predictability in human mobility. Science. 327, 5968 (2010), 1018–1021.