Traffic Distribution of IM Services

Rui-Bing Li [1,2,a],Fang-Fang Sun[1,2]

[1]*State Key Lab. of Networking and Switching Technology,*
[2]*Key Lab. of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing, P.R.China*
[a]*muzi2014@bupt.edu.cn* , [b]*cuihy@bupt.edu.cn*

Abstract—With the proposition of new opinion that the distribution of many human activities follows heavy-tailed distribution rather than Poisson distribution, the research on human behaviors has received widely attention. As the rapidly development of Immediate Message (IM) services in China, it could exactly reflect the characters of user online behaviors. QQ and We Chat are two of the most popular IM services in China. In the paper, we analyze the records of all QQ and We Chat users in City-A. We find the distribution of traffic records produced by IM service users follows heavy-tailed distribution, but it can't be fitted by existing functions properly. Then we present a new distribution --Lognormal Exponent (LNE) distribution to approximate the tail of the statistics. Our research will benefit for the research on human dynamics and the traffic engineering.

1 Introduction

In recent years, researchers have found that traditional Poisson distribution fail to describe the statistical time properties of human activities by analyzing records in massive database. Since Bara básІ [1] found that interval time distribution of human behavior showed non-Poisson statistics, many researchers have tried to establish models to character the properties of human activity, including travels [2,3], task executions [4], physical contacts [5]and financial activities [6] etc. Recently, analyzing the multi-scale properties of Internet has drawn significant attention of researchers [7]. Papers prefer various models to explain the heavy-tailed properties of the Internet. In [8], the author use Zipf laws with a shape parameter less than one to fit the popularity curves of YouTube files. In [9], the distribution of the number of conversations of a user in a month is approximated by Lognormal. Distribution of time intervals between two consecutive messages of Cyworld is described by three sections of power-law distribution [10]. Also power-law distribution is applied to model the patterns of inter-event time between conversations in [11]. In [12] authors review power-law distribution and explore implications of power-law in computer networks.

According to Cisco [13], IP traffic will grow at a rate of 34% and the global IP traffic will reach 64EB per month till 2014. More than 97.5% share of Internet traffic will be content-related. ABI Research denotes that the mobile traffic has reached 13412 PB in 2012 with 69% growth every year. 3G data usage accounted for 46% of the total amount with 130% growth. Operator Verizon Wireless expects 50% of data usage is from 4G LTE network. Strategy Analytics [14] also points out that with the growing popularity of smart phones, the traffic of mobile data will grow three times by 2017. Smart phones and tablets PCs gradually replace the PCs becoming the main computing devices of consumers. So analysis on mobile service traffic is not only useful for operator to improve resource scheduling, solve traffic unbalanced and areas of congestion but also useful for the study of human behavior.

Mobile QQ and We Chat have the function of sending messages, voicing or videoing chat with friends, writing logs, and so on. Due to the convenience of the two applications, they have become two of the most popular IM services in China. There are tens of millions records produced by IM users one day in a developed city in China. Therefor in the paper we mainly analyze on the traffic distribution of QQ and We Chat.

2 Data Set and data processing

We use a data set which includes 940,000 IM users in City-A, 315511510 mobile users records from 00:00:00 17th November 2012 to 00:00:00 23th November 2013. The data set consists of various indexes like phone number of users, total traffic produced by users when they use services, traffic type, and so on. In the paper, we study the typical IM applications: QQ and WeChat.

For QQ and WeChat, there are many records with large size of traffic, so we group records by every 105 bytes. After grouping, we count the number of records in every group. Then we calculate the ration of the number of records belonging to any group and the number of total records to get the result of traffic distribution.

3 Lognormal Exponential Distribution

3.1 The Presentation of LNE Distribution

Studying the properties of the Internet has aroused the attention of many researchers with significant breakthroughs. Papers prefer to character the Internet by various distributions at different levels. Heavy-tailed distribution including Pareto and Lognormal is widely used in analyzing human behavior.

The probability density function of Pareto is

$$f(x) = ak^a x^{-a-1} \qquad (1)$$

where k is the scale parameter and the tail index α is the shape parameter. Its cumulative distribution function is a straight line with a slope of $-\alpha$ in log-log plot.

The probability density function of standard Lognormal distribution is

$$f(x) = \frac{e^{-((\ln x)^2/2\sigma^2)}}{x\sigma\sqrt{2\pi}} \quad x > 0; \sigma > 0 \qquad (2)$$

where the standard deviation σ is the scale parameter. A variable X is Lognormal distributed if $Y = \ln(X)$ is normally.

We find that the Lognormal and Pareto distribution can't well character the tail of the distribution of traffic from previous studies. Pareto tends to powerless when faced with nonlinear tail although it is widely used to research the character of Internet. While Lognormal fits well for nonlinear characteristics at the beginning, it decays faster or slower than the tail than the true distribution. So we present a new distribution - LNE to describe the temporal character of mobile Internet. The probability density function of LNE is

$$f(x) = \frac{1}{x}\exp\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right)\right) \quad x > 0 \qquad (3)$$

where mean μ and standard deviation σ are location parameter and the scale parameter respectively. We make the following transformations, assuming

$m = \ln x$, $n = \ln(f(x))$

then

$$n = g(m) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(m-\mu)^2}{2\sigma^2}\right) \qquad (4)$$

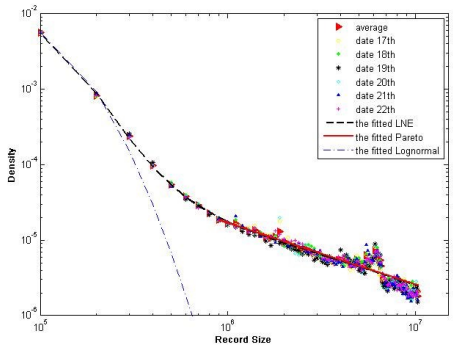
Variable m follows normally distribution with mean μ and variance σ^2. For a better fitter, in the paper we use the following function to character the distribution of m.

$$g(m) = a_1 e^{\frac{(m-b_1)^2}{c_1}} + a_2 e^{\frac{(m-b_2)^2}{c_2}} \qquad (5)$$

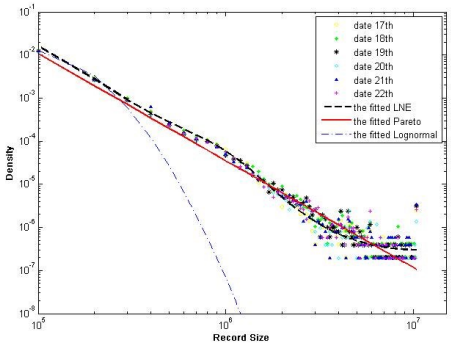
In log-log plot, the curve of LNE approximates normal distribution. We can derive the parameters of LNE by coordinate transformations from $g(m)$ to simplify calculation. In the next section, we would compare the LNE, Lognormal and Pareto in detail by the real data.

3.2 Distribution of Traffic

We use Pareto, Lognormal and LNE three types of distribution to fit the traffic distribution of the IM.



(a)Distribution of Traffic of QQ



(b)Distribution of Traffic of WeChat

Fig.1. Comparison the goodness of fit of three kinds of distribution. The x coordinate represents the size of traffic records (bytes), y coordinate is the probability density corresponding to a certain size of traffic. We use different types of points to describe different dates. Red full curve, blue dotted curve, and black broken curve respectively represent the pareto, lognormal and lne distribution.

Table 1 Goodness of Fitting

Function / R^2 / App	Pareto	Lognormal	LNE
QQ	0.7674	0.8307	0.8587
WeChat	0.9329	0.9576	0.9651

From Fig.1 and Table I we can see that the probability density of IM decrease with the traffic size increasing in log-log coordinates. Classical Pareto distribution only describes part features of the distribution of QQ. The LNE distribution fits better than Pareto distribution on We Chat application. In log-log coordinates, the rate of decay in LNE distribution slow down even to 0 with the increase of traffic size which is consistent with the real value. The rate of decay in Pareto distribution is constant. However, the rate of decay in Lognormal distribution increases with the increase of traffic size, and it only can fit well for

nonlinear characteristics at the beginning of traffic distribution.

By analyzing the distribution of the size of traffic records, we find the heavy tail characters of the internet traffic. Only a small number of traffic records is considerable. For example, the number of records whose size is smaller than 105 bytes in QQ application is 41,744,916, accounting 99.23% of total records. Only 77 records whose size between 102*105 bytes and 103*105 bytes. More than 98% of total records fall in 105 bytes in We Chat. Less than 2% of records fall in (105, 104*105) bytes.

Conclusion

In the paper, we mainly analyzed the traffic distribution of two IM services in City-A, a developed city in China. As a result of poor fitting effect of existing heavy-tailed function, we present a new distribution based on the properties of real traffic data. LNE combines the advantages of Pareto distribution and Lognormal distribution.

In log-log coordinate LNE can fit the distribution quite properly. With proper parameter LNE can fit nearly parallel tail which is impossible for Pareto and Lognormal distribution.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (61201153), the National 973 Program of China under Grant (2012CB315805), Prospective Research Project on Future Networks in Jiangsu Future Networks Innovation Institute (BY2013095-2-16), the National Basic Research Program 973 of China (2012CB315801), and the Fundamental Research Funds of China for the Central Universities (2013RC0113).

References

[1] BARABÁSI A L. The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005,435(7039): 207-211

[2] Brockmann Dirk, Lars Hufnagel, Theo Geisel. The scaling laws of human travel[J]. Nature,2006,439(7075): 462-465.

[3] Lu Xin, Linus Bengtsson, Petter Holme. Predictability of population displacement after the 2010 Haiti earthquake [J]. Proceedings of the National Academy of Sciences,2012,109(29): 11576-11581.

[4] Pickard, Galen, et al. "Time-critical social mobilization." Science, 2011, 334(6055):509-512.

[5] Zhao Kun, Juliette Stehlé, Ginestra Bianconi, Alain Barrat. Social network dynamics of face-to-face interactions [J]. Physical Review E,2011,83: 056109.

[6] Sun-Chong Wang, Jie-Jun Tseng, Chung-Ching Tai, Ke-Hung Lai, Wei-Shao Wu, Shu-Heng Chen. Network topology of an experimental futures exchange [J]. The European Physical Journal B, 2008, 62(1): 105-111.

[7] Vespignani, Alessandro. Predicting the behavior of techno-social systems [J].Science, 2009, 325(5939): 425-428.

[8] Guillemin, Fabrice, Thierry Houdoin, Stephanie Moteau. Volatility of YouTube content in Orange networks and consequences [C]. In Communications (ICC), 2013 IEEE International Conference on, pp. 2381-2385. IEEE, 2013.

[9] Leskovec, Jure, Eric Horvitz. Planetary-scale views on a large instant-messaging network [C]. In Proceedings of the 17th international conference on World Wide Web, pp. 915-924. ACM, 2008.

[10] Chun, Hyunwoo, Haewoon Kwak, Young-Ho Eom, Yong-Yeol Ahn, Sue Moon, Hawoong Jeong. Comparison of online social relations in volume vs interaction: a case study of cyworld [C]. In Proceedings of the 8th ACM SIGCOMM conference on Internet measurement, pp. 57-70. ACM, 2008.

[11] Wei HONG, HAN Xiao-Pu, ZHOU Tao, WANG Bing-Hong. Heavy-tailed statistics in short-message communication[J]. Chinese Physics Letters,2009,2009(2):297-299.

[12] Mahanti. A, Carlsson. N, Arlitt. M, Williamson. C. A tale of the tails: Power-laws in Internet measurements [J]. Network, 2013,27(1): 59-64.

[13] Forecast C V N I. Cisco Visual Networking Index: Global Mobile data Traffic Forecast Update 2009-2014[J]. Cisco Public Information, February, 2010, 9.

[14] Ingrid Lunden. Mobile Data Traffic To Grow 300% Globally By 2017 Led By Video, Web Use, Says Strategy Analytics.
http://techcrunch.com/2013/07/03/mobile-data-use-to-grow-300-globally-by-2017-led-by-video-web-traffic-says-strategy-analytics/