Characterizing and Modeling Social Mobile Data Traffic in Cellular Networks

Chen Qi*†, Zhifeng Zhao*†, Rongpeng Li‡, and Honggang Zhang*†
*York-Zhejiang Lab for Cognitive Radio and Green Communications
†College of Information Science and Electronic Engineering
Zhejiang University, Zheda Road 38, Hangzhou 310027, China
‡Huawei Technologies Co., Ltd., Shanghai 210206, China
Email: {qichen7c, zhaozf, lirongpeng, honggangzhang}@zju.edu.cn

***Abstract*—Understanding traffic characteristics in cellular networks is of great significance for better network design and performance optimization. The rapid development of various social networking applications for smart devices makes it an imperative to carry out cellular data traffic analysis further into the application level. In this paper, based on a plenty of practical mobile data traffic records, we focus on three typical application types and draw conclusions in terms of statistical characteristics and appropriate distribution model for social mobile data traffic. Firstly, the universal existence of burstiness and self-similarity is demonstrated by testing traffic series at different time scales. Afterwards, α-stable distributions are used to model traffic series benefiting from their internal burstiness and self-similarity. The minor fitting errors verify the validity of α-stable model and a preliminary traffic prediction shows the usefulness of α-stable model for further traffic analysis.**

I. Introduction

With the continuous development of wireless communication technologies and the increasing popularity of smart portable devices, mobile data service is continually gaining its dominance within cellular networks. Meanwhile, the volume of cellular data traffic has experienced sustainably significant growth, and is expected to increase dramatically in the next 5 years [1]. Such an unprecedented surge of mobile traffic poses a severe requirement for future cellular communication systems. To address the challenge, understanding the traffic characteristic and building precise models to capture the characteristics are of vital importance, taking advantage of which better optimization and management of cellular networks like opportunistic scheduling [2], and energy saving [3] can be designed and put into effect.

There have been a number of works dealing with cellular data traffic analysis. Some study data traffic characteristics from the network's perspective, namely, the aggregated traffic transmitted through base stations (BSs), and examine the basic characteristics like the temporal periodicity [4], spatial traffic density distribution and corresponding models [5], and the correlation [6] for BSs in the entire network of interest. Others pay attention to the traffic at the level of subscribers, analyzing the traffic distributions [4] while extracting human mobility and activity patterns, such as the heavy-tailed property from different subscribers to find the influence of these patterns on data service in cellular networks [7].

However, general characteristics like periodicity cannot precisely capture the varying bursts in traffic time series, which however have significant impact on the network performance. Besides, traffic models in cellular networks seldom take the human behavior patterns into consideration, for example, the heavy-tailed property and its induced long-range dependence (LRD). These shortages have drawn wide attentions in wired networks [8]–[10], but there is no related work focusing on emerging wireless service types such as mobile instant messaging (IM) and mobile video applications for cellular networks. Thus in this paper, for the social mobile cellular data traffic, we investigate the phenomena of universal bursts in traffic time series as a starting point, and characterize traffic series by the burstiness on different time scales and the self-similarity. Afterwards, we examine the preciseness of α-stable distribution to model the social mobile data traffic series and explain the differences within the estimated parameters of α-stable model among different types of service. Moreover, further discussions are conducted on explanation and application on α-stable distribution based traffic model.

The remainder of the paper is organized as follows. Section II introduces the information of datasets under study. In Section III, we characterize significant statistical characteristics of social mobile cellular data traffic. In Section IV, we model the data traffic series of different service types based on α-stable distribution, and demonstrate its validity and feasibility. Finally, we conclude this paper in Section V.

II. Dataset Description

Our dataset is based on a significant number of practical traffic records from one of the biggest cellular operators in Hangzhou, an eastern provincial capital in China. The records in dataset are originated from nearly 5000 BSs with more than 10 million subscribers involved. Each traffic record has a resolution of 5 minutes, including timestamps, location area code (LAC), cell ID, application name and the corresponding volume of data traffic.

In this study, WeChat/Weixin, HTTP web browsing and QQLive Video are selected as the representatives of the three typical types of mobile service, IM, web browsing and video for discuss, respectively. Particularly, WeChat/Weixin is a widely booming social IM service which allows over 6

TABLE I
DATASET UNDER STUDY

Service Type	IM (WeChat/Weixin)	Web Browsing (HTTP)	Video (QQLive)
Traffic Resolution	5 min	5 min	5 min
Duration	1 day	1 day	1 day
No. of Active BSs	2292	4507	4472

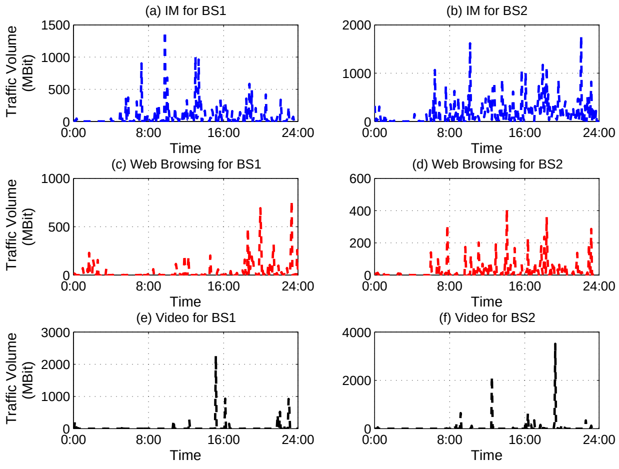

Fig. 1. Traffic time series of different mobile service types during one day.

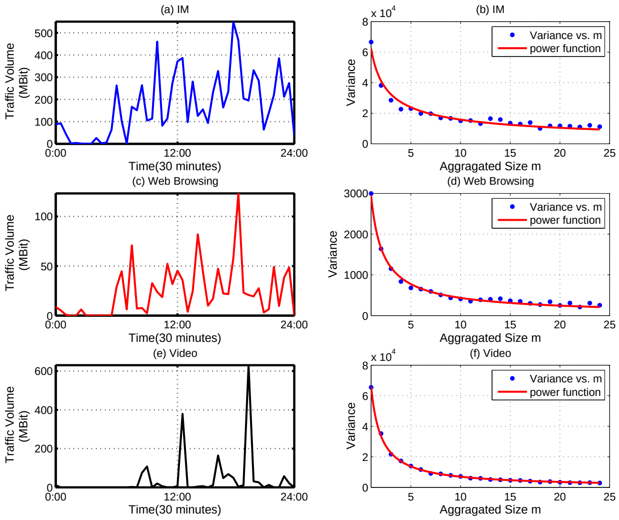

Fig. 2. Left: 6-Aggregated traffic series of different mobile service types. Right: The fitting result of the variances of aggregated series with respect to Power-law.

hundred million mobile users to exchange text messages and multimedia files like voices, pictures and videos with each other via smart phones [11], in China as well as around the world. The summary information on the mobile traffic dataset under study is listed in Table I .

III. CHARACTERIZING SOCIAL MOBILE DATA TRAFFIC

In this section, we mainly focus on characterizing application-level social mobile data traffic in cellular networks. In the first place, the burstiness in the traffic series viewed at different time scales is examined by observing the variances of related traffic series. The corresponding result could imply the possibility of self-similarity in some sense and thus a further examination on the existence of self-similarity is performed.

A. The Burstiness at Different Time Scales

Burst commonly implies sharp increase in volume of information interaction in seconds, which is potentially accompanied with the emergence of unexpected events or centralized activities of human beings. It is generally believed that bursty phenomena appears apparently and enormously in cellular data traffic series [12] which is closely related to people's daily life. In this section, we have a brief look at the burstiness of application-level cellular data traffic at different time scales and validate this intrinsic characteristics.

Fig. 1 illustrates traffic time series of different service types of two random selected BSs during one day. As Fig. 1 depicts, there certainly exists plenty of bursts for all the three types of services. Particularly, IM and HTTP web browsing services frequently produce traffic bursts; while distinct from them, video service with more sporadic activities generates more significant traffic bursts.

Afterwards, *m-aggregated* series is used to convert the original traffic series to the time series at different time scales:

Definition 1. Given a discrete time series $X = \{X_1, X_2, ..\}$, *m-aggregated* time series is defined according to the average of the original series X over non-overlapping, adjacent blocks of size m [13]:

$$X_n^{(m)} = \frac{1}{m} \sum\nolimits_{i=nm-(m-1)}^{nm} X_i. \tag{1}$$

Taking BS2 in Fig. 1 as an example, several aggregated series are computed with the increasing aggregated size m. The left column in Fig. 2 shows 6-aggregated traffic series (i.e., viewing the traffic series at a scale of 30 minutes) of different types of services. Obviously, the curves maintain bursty characteristics although they become smooth to some degree. Furthermore, the numeral characteristic, variance is calculated in order to examine the degree of burstiness at different time scales quantificationally. Fig. 2 (b), (d), (f) plot variances of different aggregated traffic series versus the corresponding aggregated size m with blue dots and the fitting results according to power functions in red curves. As the figure depicts, there exists such relationship as $Var(X^{(m)}) \sim am^{-b}$ with $0 < b < 1$ for all the three types of services. That is to say, the variance of the sampled traffic series decreases more slowly than the reciprocal of the sample size [10], which indicates that the burstiness of application-level data traffic series remains significant as the time scale increases.

Remark 1. *Application-level cellular data traffic series for IM, Web Browsing and video service appear bursty across a long range of time scales. The burstiness remains significant as the time scale increases.*

B. The Self-Similarity

In the last subsection, the universal existence of burstiness in application-level data traffic series has been illustrated. Recalling Fig. 1 (b), (d), (f) and Fig. 2 (a), (c), (e), we can see the 6-aggregated traffic series resembles the original ones on shapes. In this regard, self-similarity, a concept from fractal theory [9] comes naturally which reflects that objects' appearance remains unchanged regardless of the scale of viewpoint. For time series, it can be defined as follows:

Definition 2. Given a zero-mean, stationary time series $X = \{X_1, X_2, ...\}$, we say that X is *H-self-similarity*, if for any positive $m \in N$, the sum of the original series X over nonoverlapping blocks of size m, (i.e., the form in Eq. (1)), has the same distribution as X rescaled by m^H. That is,

$$X_n \overset{\mathrm{d}}{=} m^{-H} \sum_{i=nm-(m-1)}^{nm} X_i = m^{-H} m X^{(m)}. \tag{2}$$

where the notion $\overset{\mathrm{d}}{=}$ denotes equality in the sense of distribution [10].

The parameter H is known as the Hurst parameter with the value ranging from 0.5 to 1.0 and has a positive correlation with the degree of self-similarity. That is to say, H =0.5 indicates the lack of self-similarity whereas large value for H (i.e., close to 1.0) indicates a large degree of self-similarity [9].

Generally, graphical methods such as variance-time plot, R/S plot are used to test for self-similarity [13]. In Section III-A, the right column in Fig. 2 has shown that the variance of aggregated series decreases slowly according to power function, which is indicative of self-similarity. Moreover, the linear trend of the log-log plot of R/S statistic against m proves self-similarity and the slope of the fitting line is an estimate of H parameter [13]. Fig. 3 (a) $\sim$ (c) shows the R/S log-log plots of different types of services for one random selected BS, where the appropriate fitting lines suggest the certain existence of self-similarity of the sampled traffic series in terms of all the three service types under study.

Moreover, in order to gauge self-similar property for the whole dataset, a deep look at the estimated H parameter and R-square are taken, as the H parameter indicates the degree of self-similarity and R-square examines the accuracy of the linear trend of R/S log-log plot. According to Fig. 3 (d) $\sim$ (f), for the three types of services, almost all the R-square is larger than 0.96, which indicates a pretty precise linear fit. The distributions of H parameter exhibit some differences: IM and Web browsing have similar results with the values of H parameter mainly ranging from 0.6 to 0.8 while video service shows weaker self-similarity as more than 95% of H parameter is smaller than 0.7.

Previous works focusing on self-similarity in wired networks [9], [13] explains the self-similarity of traffic series in terms of user behavior, network evolution, file system characteristics and traffic aggregation, which always result in significant bursts in traffic series. Such explanations also apply to application-level cellular data traffic.

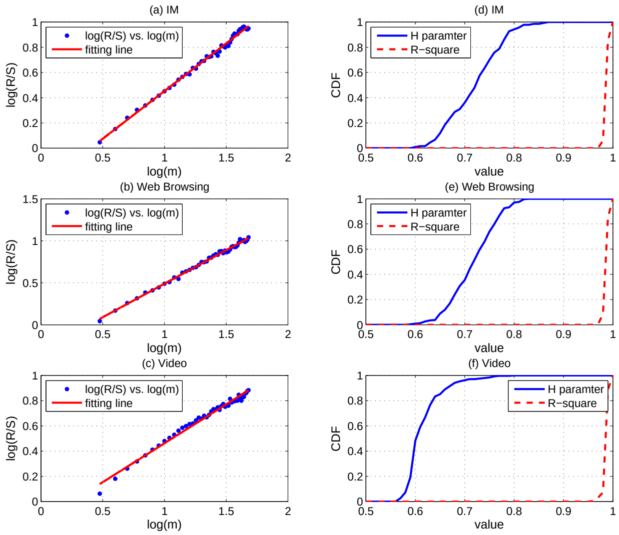

Fig. 3. Left: R/S plot and fitting result. Right: The CDF of H parameter and R-square.

Remark 2. *There widely exists self-similarity in application-level cellular data traffic in terms of IM, Web browsing and video services. Specifically, for IM and web browsing service, most traffic series exhibit a moderate degree of self-similarity while video service shows weaker self-similarity compared with the other two services under study.*

IV. Modelling Traffic Dynamics with α-Stable Distribution

In last section, statistical characteristics of application-level cellular data traffic like burstiness and self-similarity are demonstrated, which have significant effects on network performance [12] nevertheless are often ignored in traffic modeling. Here a class of distribution known as α-stable distribution which is always producing strong bursts as well as closely related to self-similarity is used to model traffic series purposely on the aforementioned characteristics.

A. Definition of α-Stable Distribution

α-Stable models, with few exceptions, lack a closed-form expression of the probability density function (PDF), and are generally specified by their characteristic functions.

Definition 3. A random variable X is said to obey α-stable models if there are parameters $0 < \alpha \leq 2$, $\sigma \geq 0$, $-1 \leq \beta \leq 1$, and $\mu \in \mathcal{R}$ such that its characteristic function is of the following form:

$$\begin{aligned}\Phi(\omega) &= E(\exp j\omega X) \\ &= \begin{cases} \exp\left\{-\sigma^{\alpha}|\omega|^{\alpha}\left(1 - j\beta(\operatorname{sgn}(\omega))\tan\dfrac{\pi\alpha}{2}\right) + j\mu\omega\right\}, \\ \hfill \alpha \neq 1; \\ \exp\left\{-\sigma|\omega|\left(1 + j\beta(\operatorname{sgn}(\omega))\ln|\omega|\right) + j\mu\omega\right\}, \alpha = 1. \end{cases}\end{aligned} \tag{3}$$

TABLE II
THE PARAMETER FITTING RESULTS IN THE α-STABLE MODELS

Service Type	**Parameters**			
	α (Stability)	β (Skewness)	σ (Scale)	μ (Shift)
IM	1.5658	1	180.7507	250.4082
Web Browsing	1.6018	1	32.3341	42.7453
Video	0.5130	1	1×10^{-10}	0

Here, α is called the characteristic exponent and indicates the index of stability, while β is identified as the skewness parameter. α and β together determine the shape of the models. Moreover, σ and μ are called scale and shift parameters, respectively. Specifically, if $\alpha = 2$, α-Stable models reduce to Gaussian distributions.

Furthermore, for an α-stable modeled random variable X, there exists a linear relationship between the parameter α and the function $\Psi(\omega) = \ln\left\{-\mathrm{Re}\left[\ln\left(\hat{\Phi}(\omega)\right)\right]\right\}$ as

$$\Psi(\omega) = \ln\left\{-\mathrm{Re}\left[\ln\left(\Phi(\omega)\right)\right]\right\} = \alpha\ln(\omega) + \alpha\ln(\sigma). \quad (4)$$

Usually, it is challenging to prove whether a dataset follows a specific distribution, especially for α-stable models without a closed-form expression for the PDF. Therefore, when a dataset is said to satisfy α-stable models, it usually means the dataset is consistent with the hypothetical distribution and the corresponding properties. In other words, the validation needs to firstly estimate parameters of α-stable models based on the given dataset, and then compare the real distribution of the dataset with the estimated α-stable model [14]. Specifically, the corresponding parameters in α-stable models can be determined by maximum likelihood methods, quantile methods, or sample characteristic function methods [14], [15].

B. Fitting Results with Respect to α-Stable Models

In this subsection, the results of fitting application-level cellular data traffic to α-stable models are examined. Firstly, for traffic series of different service types in a random selected BS, the parameters of α-stable models are estimated based on quantile methods [16] and the results are listed in Table II.

Afterwards, we use the α-stable models, produced by the aforementioned estimated parameters, to generate some random variable, and compare the induced cumulative distribution function (CDF) with the exact (empirical) one. Fig. 4 presents the corresponding comparison between the simulated results and the real ones. Recalling the statement in Section IV-A, if the simulated dataset has the same or approximately same distribution as the real one, the empirical dataset could be deemed as α-stable modeled. Therefore, Fig. 4 indicates the traffic records in these selected areas could be simulated by α-stable models. On the other hand, Fig. 5 shows the preciseness error CDF for all the BSs after fitting $\Psi(\omega)$ with respect to ω to a linear function, and implies that there merely exists minor fitting errors for all the BSs of different service types. This phenomenon further verify the validity of α-stable model.

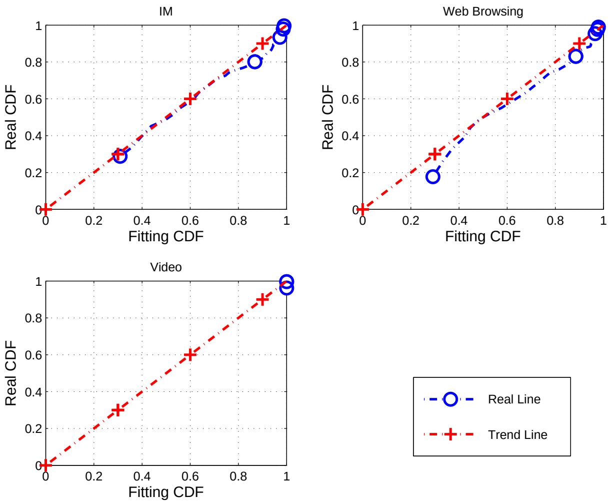

Fig. 4. For different service types, α-stable model fitting results versus the real (empirical) ones in terms of the CDF.

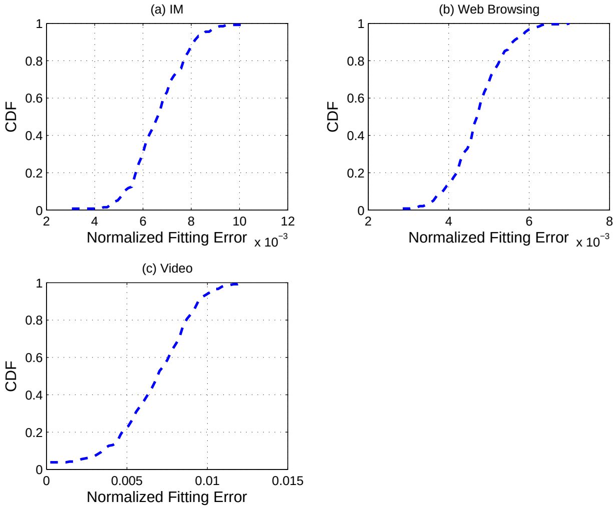

Fig. 5. The preciseness error CDF for all the BSs after fitting $\Psi(\omega)$ with respect to $\ln(\omega)$ to a linear function.

At the same time, the distinct parameters in α-stable models reflect different characteristics for traffic series of different service types. According to Table II, the IM service has a larger μ. Since the expectation of an α-stable modeled variable equals μ when $1 < \alpha \leq 2$ [17], the popularity of social networks makes IM service usually generate more traffic than Web browsing service. On the other hand, due to traffic-consuming characteristics of video service, people usually prefer to enjoy online videos via wireless networks. Hence, video service has sporadic traffic activities during one day. Correspondingly, for video service, the fitting value of α becomes much less than the other service types.

Remark 3. *According to the minor fitting errors, α-stable models are suitable to characterize the application-level data traffic in cellular networks.*

The reasons that IM service traffic universally obeys α-stable models could be explained as follows. [18] unveiled

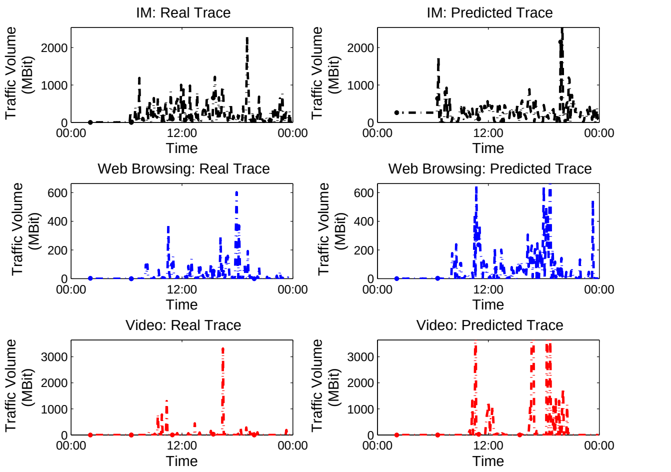

Fig. 6. For different service types, the prediction performance by an α-stable model-based $(36, 10, 1)$-linear prediction method.

that the message length of one individual IM activity follows a power-law distribution. Meanwhile, the traffic distribution within one BS can be regarded as the accumulation of lots of IM activities. Moreover, according to the generalized central limit theorem, the sum of a number of random variables with power-law distributions decreasing as $|x|^{-\alpha-1}$ where $0 < \alpha < 2$ (and therefore having infinite variance) will tend to an α-stable model as the number of summands grows. Hence, the traffic series of IM service within one BS follows α-stable models and such explanation can be generalized to other service types.

C. A Preliminary Traffic Prediction Analysis Based on α-stable Model

The aforementioned results have demonstrated the validity for α-stable models to describe application-level social mobile data traffic with universal burstiness and self-similarity. Moreover, methods based on α-stable model can be utilized to deal with further application-level traffic processing. For example, [19] proposes a linear prediction method based on α-stable model for self-similar traffic which can be used to avoid network congestion and improve network performance. Applying the method to our application-level data traffic, we utilize 36 consecutive traffic records to predict the traffic value at the next moment. Fig. 6 depicts the corresponding prediction performance for the traffic series of three service types. According to Fig. 6, the linear prediction method could well predict the traffic trends. However, there still exists some gap between the real traffic trace and the predicted one, which need to be improved in future works.

V. Conclusion

In this paper, we carry out cellular data traffic analysis further into application level in terms of statistical characteristics and suitable model. Taking three typical types of services as examples, we demonstrate the universal existence of burstiness and self-similarity and their great significance in social mobile data traffic series. To capture these characteristics, α-stable distribution is taken to model traffic series. The minor fitting errors for different service types verify the validity of α-stable models and the estimated parameter can reflect the characteristics of traffic series well. Besides, a linear prediction method based on α-stable model is implemented as an application and can predict the trends of application-level traffic series, which demonstrates the feasibility of the analytic results.

References

[1] Cisco, “Cisco visual networking index: Global mobile data traffic forecast update, 2014–2019,” Feb. 2015. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html

[2] R. Li, Z. Zhao, X. Zhou, and H. Zhang, “Energy savings scheme in radio access networks via compressive sensing-based traffic load prediction,” *Trans. Emerg. Telecommun. Technol. (ETT)*, vol. 25, no. 4, pp. 468–478, Apr. 2014.

[3] U. Paul, L. Ortiz, S. R. Das, G. Fusco, and M. M. Buddhikot, “Learning probabilistic models of cellular network traffic with applications to resource management,” in *Proc. IEEE DySPAN 2014*, McLean, VA, USA, Apr. 2014, pp. 82–91.

[4] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, “Understanding traffic dynamics in cellular data networks,” in *INFOCOM, 2011 Proceedings IEEE*. Shanghai China: IEEE, Apr. 2011, pp. 882–890.

[5] H. Klessig, V. Suryaprakash, O. Blume, A. Fehske, and G. Fettweis, “A framework enabling spatial analysis of mobile traffic hot spots,” *IEEE Wireless Communications Letters*, vol. 3, no. 5, pp. 537–540, Oct. 2014.

[6] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang, “Spatial modeling of the traffic density in cellular networks,” *IEEE Wireless Communications*, vol. 21, no. 1, pp. 80–88, Feb. 2014.

[7] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang, “Human mobility patterns in cellular networks,” *IEEE Communications Letters*, vol. 17, no. 10, pp. 1877–1880, Oct. 2013.

[8] O. Cappé, E. Moulines, J.-C. Pesquet, A. Petropulu, and X. Yang, “Long-range dependence and heavy-tail modeling for teletraffic data,” *IEEE Signal Processing Magazine*, vol. 19, no. 3, pp. 14–27, Oct. 2002.

[9] A. Popescu, “Traffic self-similarity,” in *IEEE International Conference on Telecommunications, ICT2001, Bucharest, Romania*. Citeseer, 2001, pp. 20–24.

[10] M. E. Crovella and A. Bestavros, “Self-similarity in world wide web traffic: evidence and possible causes,” *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, Dec. 1997.

[11] Tencent, Inc., “Wechat - the new way to connect,” 2011. [Online]. Available: http://www.wechat.com/en/

[12] X. Zhou, Z. Zhao, R. Li, Y. Zhou, T. Chen, Z. Niu, and H. Zhang, “Toward 5g: when explosive bursts meet soft cloud,” *IEEE Network*, vol. 28, no. 6, pp. 12–17, Nov. 2014.

[13] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, “Self-similarity through high-variability: statistical analysis of ethernet lan traffic at the source level,” *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71–86, Feb. 1997.

[14] G. Xiaohu, Z. Guangxi, and Z. Yaoting, “On the testing for alpha-stable distributions of network traffic,” *Computer Communications*, vol. 27, no. 5, pp. 447–457, Mar. 2004.

[15] J. R. Gallardo, D. Makrakis, and L. Orozco-Barbosa, “Use of α-stable self-similar stochastic processes for modeling traffic in broadband networks,” vol. 40, no. 1. Elsevier, Mar. 2000, pp. 71–98.

[16] J. H. McCulloch, “Simple consistent estimators of stable distribution parameters,” *Communications in Statistics-Simulation and Computation*, vol. 15, no. 4, pp. 1109–1136, Jan. 1986.

[17] G. Samoradnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. New York: CRC Press, Jun. 1994.

[18] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang, “Understanding the nature of social mobile instant messaging in cellular networks,” *IEEE Communications Letters*, vol. 18, no. 3, pp. 389–392, Mar. 2014.

[19] L. Xiang, X.-H. Ge, C. Liu, L. Shu, and C.-X. Wang, “A new hybrid network traffic prediction method,” in *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*. Miami, FL: IEEE, Dec. 2010, pp. 1–5.