Understanding the Nature of Social Mobile Instant Messaging in Cellular Networks

Xuan Zhou, Zhifeng Zhao, *Member, IEEE*, Rongpeng Li, Yifan Zhou, Jacques Palicot, and Honggang Zhang, *Senior Member, IEEE*

***Abstract*—Social mobile instant messaging (MIM) applications, running on portable devices such as smartphones and tablets, have become increasingly popular around the world, and have generated significant traffic demands on cellular networks. Despite its huge user base and popularity, little work has been done to characterize the traffic patterns of MIM. Compared with traditional cellular network service, MIM traffic embodies several specific attributes such as non-Poisson arrivals, keep-alive (KA) mechanism and heavy-tailed message length, which consumes even small amount of core network bandwidth but considerable radio resources of mobile access network. This letter investigates user behavior patterns and traffic characteristics of MIM applications, based on real traffic measurements within a large-scale cellular network covering 7 million subscribers. Moreover, we propose a joint ON/OFF model to describe the traffic characteristics of MIM, and evaluate the performance of cellular network running MIM service in various scenarios. Comparing with the MIM service models of 3GPP, our results are more realistic to estimate the networks performance.**

***Index Terms*—Mobile instant messaging, inter-arrival time, message length, ON/OFF model, blocking rate, power-law, log-normal, cellular networks.**

I. Introduction

INSTANT messaging service has been widely adopted both on PC platforms and mobile devices in personal and business communications. However, newly emerging social applications such as mobile instant messaging (MIM) on smartphones and tablets have generated explosive traffics, which fundamentally affects the stabilization and reliability of on-operating cellular networks [1][2]. In this letter, we take into account a widely booming MIM application, "WeChat/Weixin", which allows about 600 million mobile users to exchange text messages, voices, pictures and videos with each other via smartphones [3], around the world and especially in China.

Recently, [4] and [5] have investigated the traffic characteristics of AIM (AOL Instant Messenger) and Windows Live Messenger on wired Internet, respectively. Meanwhile, several heavy-tailed distribution phenomena have been reported in wired Internet traffic, for which Weibull, lognormal and power-law distribution models were employed separately [4]-[6]. On the other hand, compared with traditional voice and messaging services within the cellular networks, the emerging social MIM application distinguishes itself with non-Poisson arrivals, keep-alive (KA) mechanism and heavy-tailed message length, which consumes even small amount of core network bandwidth but considerable radio resources of mobile access networks. Specifically, due to the KA mechanism to keep the mobile users in touch with the servers, the MIM service would become a huge burden on the network operators. Therefore, it's meaningful to investigate the basic characteristics of MIM applications traffic. However, to our best knowledge, few researchers have ever addressed the MIM traffic characteristics except the 3GPP Technical Report [7], which adopts an exponential model for the messages intervals and a geometric model for the message length. Interestingly, our recent measurement results are not compatible with these models proposed by 3GPP. In this letter, we make an intensive study on the fundamental characteristics of MIM through the observations from practical cellular networks. Based on real measurements of "WeChat/Weixin", we investigate the operation method and the service model of the MIM, and evaluate its influence on performance of cellular networks. Although [8] has introduced the impact of inactivity timer on the performance of cellular networks, satisfactory investigations have been seldom carried out to unveil the relationship between the way the MIM works and the performance of cellular networks.

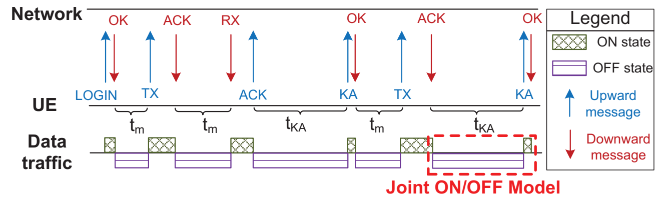

Fig. 1. MIM interactions and data traffic between UE and network.

II. MIM Traffic Measurements and Modeling

Fig. 1 illustrates how a MIM application works on the user equipment (UE). After login to the server, the MIM begins to transmit (TX) and receive (RX) messages via the cellular network, and the receiver would acknowledge (ACK) after receiving a message. Some mobile applications such as MIM and social network services (SNS) require long-lived

Manuscript received November 22, 2013. The associate editor coordinating the review of this letter and approving it for publication was D. Qiao.

X. Zhou, Z. Zhao, R. Li, Y. Zhou, and H. Zhang are with Zhejiang University, China (e-mail: {zhouxuan, zhaozf, lirongpeng, zhouyftt, honggangzhang}@zju.edu.cn). H. Zhang is also with the Université Européenne de Bretagne & Supélec, France.

J. Palicot is with Supélec, France (e-mail: jacques.palicot@supelec.fr).

This letter is supported by the National Basic Research Program of China (973Green, No. 2012CB316000) and the grant of "Investing for the Future" program of France ANR to the CominLabs Excellence Laboratory (ANR-10-LABX-07-01).

Digital Object Identifier 10.1109/LCOMM.2014.012014.132592

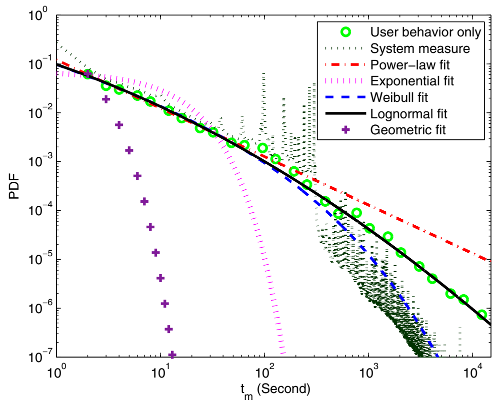

Fig. 2. The fitting results of MIM's inter-arrival time by candidate distributions.

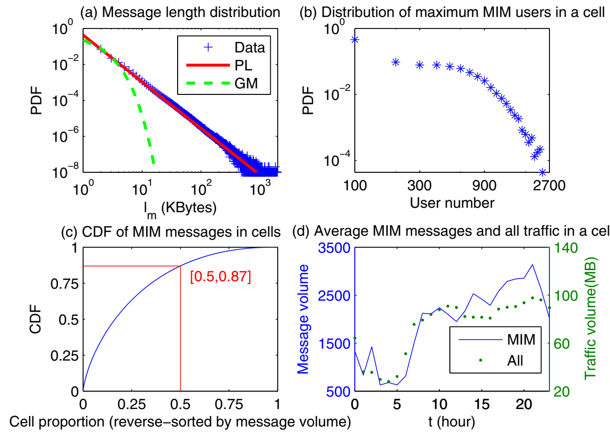

Fig. 3. Traffic characteristics of MIM service.

TABLE I
AIC TEST: THE INTER-ARRIVAL TIME AND MESSAGE LENGTH OF MIM.

Distribution	PDF	W_t	W_l
Power-law (PL)	ax^{-b}	1.8×10^{-87}	1.000
Exponential (EXP)	ae^{-bx}	0.000	0.000
Weibull (WB)	$abx^{b-1}e^{-ax^b}$	3.1×10^{-44}	9.4×10^{-74}
Lognormal (LN)	$\frac{1}{bx\sqrt{2\pi}}e^{\frac{(\ln(x)-a)^2}{-2b^2}}$	1.000	5.3×10^{-13}
Geometric (GM)	$(1-a)^x a$	0.000	0.000

connections with the servers, so the service providers usually set a timer (t_{KA}) to send KA messages periodically. Along with the login procedure, UE will receive an OK message if the KA message is successfully reported to the server.

Hereinafter, measurements of the MIM traffic from the operating cellular networks are presented, meanwhile several primary models of MIM are built up according to the actual measurements. Our dataset from the measurements consists of 1 month traffic records collected from 7 million subscribers, originated from about 15000 GSM and UMTS base stations of China Mobile within a region of 3000 km^2. The measurement records that contain timestamps, cell IDs, anonymous subscriber IDs, message lengths and message types, are collected from the Gb and Gn interfaces [9]. The *message* here refers to a series of packets transmitted between the UE and the servers of service provider on application layer. Generally, messages are usually transmitted with TCP connection, which cause occupations on both uplink and downlink data channel. From now on, we don't distinguish the directions of the messages as we are concerning with the data channels.

The interval t_m between two consecutive MIM messages of a subscriber (UE) can be obtained based on the timestamps of messages in the traffic records. Here, t_m are sampled in the order of second. Accordingly, the probability density function (PDF) of MIM's inter-arrival time can be calculated.

In order to model the activity pattern of user's behavior only, we firstly pay attention to the arrival intervals of messages except the KA messages. The distribution of inter-arrival time is fitted using the common heavy-tailed distributions listed in Table I, and the unknown parameters in the distributions are obtained by maximum likelihood estimation (MLE) method. Fig. 2 depicts the relevant fitting results compared to the empirical data (see legend: User behavior only) with the inter-arrival time from the 2nd to the 3000th second. Together with the MLE, an Akaike information criterion (AIC) [10] is also applied, so as to quantitatively find the fittest distribution. Interestingly, the result of AIC test in column W_t shows that the PDF of inter-arrival time of users' messages $f_u(t_m)$ is best fit by lognormal distribution instead of exponential distribution that is recommended by 3GPP [7]. Similar to the t_m, the PDF of the user message length denoted as $g_u(l_m)$ can also be calculated, where l_m is the user message length measured in terms of KByte. Column W_l in Table I shows that a power-law distribution fits best. In contrast with the geometric distribution recommended in [7], Fig. 3(a) illustrates the measurements and verifies the power-law distribution is the fittest one. Moreover, t_m and l_m are found independent from the measurements without regard to the KA messages.

Afterwards, we consider the arrive intervals of all kinds of messages including the KA messages. In Fig. 2, the intervals of KA messages among all the intervals can be easily recognized from the peaks with the legend: System measure. Due to various UE systems and MIM versions, the KA cycles observed in Fig. 2 are quite different but usually appear at multiples of 30 seconds. Moreover, the PDF of the inter-arrival time decreases sharply when t_m is larger than the maximal KA (300s) and the percentage of messages with inter-arrival time larger than the maximal KA period accounts only 2%. Therefore, it is reasonable to think that MIM has to send the KA messages if it has been out of touch with the servers for KA period unless the connection to the network is abnormal. On the other hand, the KA messages only contain basic status information of the MIM and thus have a constant length l_{KA}.

We also investigate the macroscopic view of MIM service in the cellular networks. Fig. 3(b) tells us the distribution of maximal number of MIM users in a cell during one day, while about 1% cells have more than 1400 users in the most crowded time, and the maximal MIM user number in a cell is about 2400. In Fig. 3(c), the imbalance feature of MIM messages in cells is described by CDF, within which all cells are reverse-sorted by their daily message volumes, and the X-axis and Y-axis represent the proportion of cells and messages respectively. We find that 87% MIM messages are transmitted on 50% cells. The comparison between the MIM message volumes and all data traffic volumes is expressed in Fig. 3(d). The left Y-axis represents the average MIM message volumes in a cell during one hour, while the right one is the average traffic volumes of all services in a cell. The MIM traffic has a distinct peak at 9PM, which is different with the volumes of all traffic. We guess people apt to use MIM applications more when they are free of both office and family affairs.

It is worthy to note that both the PDFs of the arrival interval and the message length are heavy-tailed. In other words, the inter-arrival time and the message length follow non-memoryless process. Therefore, the corresponding service characteristics could not be precisely modeled by the Markov process. In this letter, we take into account the intermittent activities to send the messages and adopt a revised ON/OFF model to formulate the service models. Generally, as illustrated in Fig. 1, the ON/OFF model is a two-state process alternating between the ON state and the OFF state, so t_m and $\frac{l_m}{r}$ could be the OFF period and ON period respectively, where r is the transmission speed. Moreover, in the classical ON/OFF model, ON-periods and OFF-periods are independent positive random variables. However, due to the specific KA mechanism in MIM services, we find that more than 99% OFF-periods to a length of t_{KA} are followed by an ON-period whose length is l_{KA}. In other words, the inter-arrival time is correlated with the message length. Specifically, once the OFF-time (inter-arrival time) equals t_{KA}, the consequent ON-time must be l_{KA}/r (KA messages transmission time). Therefore, a Joint ON/OFF Model (JOOM) is proposed to describe the MIM service pattern: one JOOM unit consists of an OFF-period and a sequential ON-period as shown in Fig. 1, with the joint distribution of the ON and OFF periods as follows:

$$f(t_m, \frac{l_m}{r}) = \begin{cases} \alpha \cdot \delta(t_m - t_{KA}); & t_m = t_{KA}, l_m = l_{KA} \\ (1-\alpha) \cdot f_i(t_m) \cdot g_u(l_m); & t_m < t_{KA} \\ 0; & t_m > t_{KA} \end{cases} \tag{1}$$

where g_u is a power-law distributed function according to Table I, $f_i(t_m)$ is the inter-arrival distribution influenced by the KA mechanism, α is the probability of occurrence of KA messages, and δ is the impulse function. For simplicity, we assume that all MIM clients have the same KA period. If the interval of two consecutive messages of the MIM exceeds the KA period, under the impact of KA mechanism, the new interval should be remainder of the old interval divided by t_{KA}. Therefore, the $f_i(t_m)$ is the sum of piecewise functions of the shifted f_u, while the shift values are multiples of t_{KA}:

$$f_i(t_m) = \sum_{i=0}^{[t_{max}/t_{KA}]} f_u(t_m + it_{KA}), 0 \leq t_m < t_{KA} \tag{2}$$

where $[\cdot]$ is the floor function, and t_{max} is the maximal arrive interval. After concatenating the consecutive OFF and ON periods and considering the OFF-ON pair independent with the others, the JOOM can better approximate the MIM traffic.

When it comes to the wireless interfaces in GSM EDGE Radio Access Network (GERAN) and Universal Terrestrial Radio Access Network (UTRAN), the inactivity timer has influences on the realistic resource allocation and further affects both the inter-arrival time and message length of MIM services. Fig. 4 illustrates the typical transmission patterns on the wireless interfaces. The UE and the network firstly need to negotiate data channel allocation by using control channel before the data transmission, and the allocation would be failed if all data channels are occupied by other UEs, just like the second data message of User B shown in Fig. 4. On the other hand, control channels may also be occupied by other UEs when the UE need to establish negotiation with networks, like the second control signaling of User A, which makes the network inaccessible for User A. When the UE is in the connected state, if the idle time is longer than the value of inactivity timer t_{ia} , the UE switches from the connected state to the idle state, so that the data channels can be assigned to other users. For example, for User A, two consecutive data messages are concatenated when the interval between them is smaller than t_{ia}, and the control channel isn't required for the followed message. Therefore, due to the impact of inactivity timer, the case where t_m is less than t_{ia} doesn't exist, and the length of some messages become larger because of message concatenation. The concatenation probability of two consecutive messages can be expressed as:

$$P_c = \int_0^{t_{ia}} f_i(t_m)dt_m \tag{3}$$

The PDF of JOOM in (1) would be adjusted by (3), and the corresponding deduction is omitted due to the page limitation.

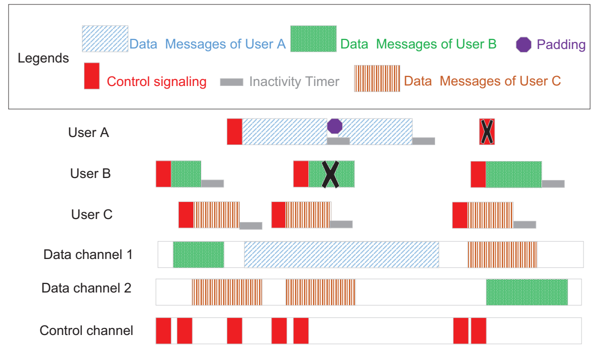

Fig. 4. An example of control and data channels allocation.

III. MIM Service Impact on Network Performance

In this section, we analyze the impact of MIM service on the network performance when the network parameters and the MIM traffic varies. Specifically, we use the proposed JOOM to simulate the MIM traffic and examine the influence of MIM on network blocking rate. The message length is divided by transmission rate to get the service time of message on the data channel, while the length of message during accessing the control channel is constant.

Since both the data and control channels are limited in the practical cellular networks, when service demands burst, transmission may be blocked. If the inactivity timer t_{ia} is set too small, channels would be assigned and released frequently, which leads to high occupation of control channel (i.e. signaling storm). If the t_{ia} is set too large, data channels would be occupied longer even if there is no data to transmit. Similarly,

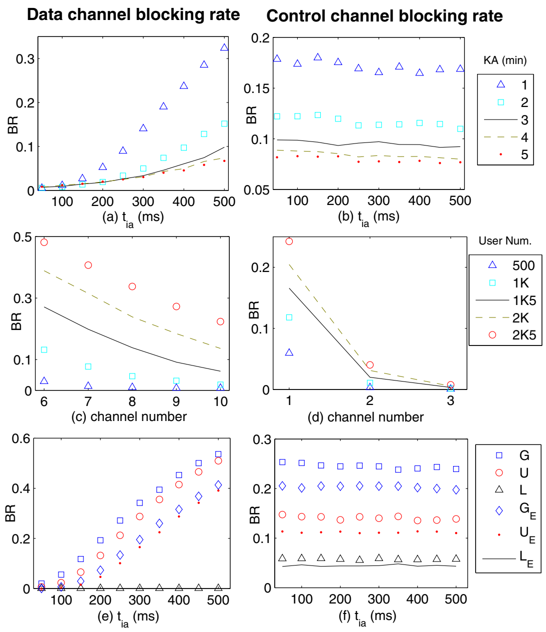

Fig. 5. Network performance versus network and MIM parameters.

small KA period may cause heavy burden on the channels while large one deteriorates the users' quality of experience (QoE). Accordingly, trade-off should be made between the control channel and the data channel usage, as well as between the network load and the QoE. Fig. 5(a)(b) illustrates the blocking rate (BR) of 1000 users within the GERAN network with 8 data channels and 1 control channel as the t_{ia} and the t_{KA} vary. It can be observed that the data channels would be significantly affected by the t_{ia} and the t_{KA}, while the control channel is mainly influenced by the t_{KA}.

Furthermore, in Fig. 5 (c)(d), we simulate the blocking rate with respect to the variations of the channel and user number when $t_{ia} = 0.3s$ and $t_{KA} = 120s$. The data channels are blocked heavier than the control channels when users aggregate, and recover slower if the same number of additional channels are allocated.

At last, we change the transmission rate and compare the corresponding performance of UTRAN (U) and LTE (L) networking scenarios with GERAN (G) in Fig. 5(e)(f). In order to test the performance under high networks workload, the maximal user number in Fig. 3(b) is applied, and the KA period is $120s$. Data channels in LTE system are shared by users, which is different from the other two networks [11]. Therefore the impact of inactivity timer on data channel can be neglected in LTE. Obviously, LTE has much better performance than the other two modes. However it still needs balance between the usage of control channel when mobile users aggregate, and about 5% BR still exists in the control channel. According to traditional network planning specification [12], the inter-arrival time of MIM has been widely modeled by exponential distribution while the message length obeys geometric distribution as specified in [7], whose performances are also plotted in Fig. 5(e)(f) as G_E, U_E and L_E. We observe that the traditional service model would definitely underestimate the strength and negative effect of traffic bursts, and the actual blocking rates wouldn't achieve the designed target performances. In particular, the KA period is recommended to be less than 3 minutes and the inactivity timer should be carefully set in GERAN and UTRAN systems, so that the blocking rate could be controlled lower than 10%.

IV. Conclusion

In this letter, we investigate the traffic characteristics of social MIM service and the corresponding impact on network performance based on real traffic measurements on a large-scale cellular network. Specifically, the heavy-tailed user behaviors have been characterized by lognormal distribution for the inter-arrival time and power-law distribution for the message length, respectively. Also, the keep-alive mechanism has been analyzed, and particularly the joint ON/OFF model for describing the traffic characteristics of MIM has been put forward. In regard to the influence of MIM service on network performance, the numerical results verify that the data channels are more sensitive to MIM traffic load than the control channels. Moreover, short keep-alive period has the non-negligible impact on the network performance, so traffic detection and QoE management of MIM are worth being our future work. Obviously, the previously deployed GERAN and UTRAN could not match well with the bursting service demands, while the cutting-edge LTE networks with channel sharing technology works much better than the former generation networks. However, the control channels of LTE would still be under the negative influence of MIM services when mobile users aggregate. Therefore, different service management strategies should be considered for different networking modes, while the shortage of LTE control channels under extreme network conditions needs to be paid attention.

References

[1] M. Donegan, "Android signaling storm rises in Japan," http://www.lightreading.com/blog/mobile-operating-systems/android-signaling-storm-rises-in-japan/240005553, Jan 2012.
[2] P. Sandle, "O2 says iPhone demand strained its London network," http://www.reuters.com/article/2009/12/29/o2-iphone-idUSLDE5BS0M520091229, Dec 2009.
[3] "WeChat/Weixin home page," http://www.wechat.com/en/.
[4] Z. Xiao, L. Guo, and J. Tracey, "Understanding instant messaging traffic characteristics," in *Proc. 2007 IEEE ICDCS*, pp. 51–58.
[5] J. Leskovec and E. Horvitz, "Planetary-scale views on a large instant-messaging network," in *Proc. 2008 ACM WWW*, pp. 915–924.
[6] A. B. Downey, "Lognormal and Pareto distributions in the Internet," *Computer Commun.*, vol. 28, no. 7, pp. 790–801, 2005.
[7] 3GPP TR 43.802 Technical Specification Group, "GERAN study on mobile data applications V0.4.0," 3GPP, Tech. Rep., 2013.
[8] M. Chuah, W. Luo, and X. Zhang, "Impacts of inactivity timer values on UMTS system capacity," in *Proc. 2002 IEEE WCNC*, pp. 897–903.
[9] 3GPP TS 23.060 Technical Specification Group, "General Packet Radio Service; service description; stage 2 V11.7.0," 3GPP, Tech. Rep., 2013.
[10] K. P. Burnham and D. R. Anderson, "Multimodel inference: understanding AIC and BIC in model selection," *Sociological Methods and Research*, vol. 33, no. 2, pp. 261–304, Nov. 2004.
[11] 3GPP TS 36.300 Technical Specification Group, "E-UTRA and E-UTRAN; overall description; stage 2 V12.1.0," 3GPP, Tech. Rep., 2013.
[12] J. Laiho, A. Wacker, and T. Novosad, *Radio Network Planning and Optimisation for UMTS*. Wiley, 2006.