



# 移动业务特征认知及其分布模型: 以即时消息为例

李荣鹏<sup>1\*</sup>, 赵志峰<sup>1</sup>, 张宏纲<sup>1</sup>, 钟晓峰<sup>2</sup>

1. 浙江大学信息与电子工程学院, 杭州 310027

2. 清华大学电子工程系, 北京 100084

\* 通信作者. E-mail: lirongpeng@zju.edu.cn

收稿日期: 2017-02-16; 接受日期: 2017-03-20; 网络出版日期: 2017-05-04

浙江省重点科技创新团队 (批准号: 2013TD20)、浙江省科技计划 (批准号: 2015C01075) 和中国博士后创新人才支持计划 (批准号: BX201600133) 资助项目

**摘要** 众所周知, 业务流量特点建模有助于设计能效低、稳定性强的网络协议. 近年来新兴业务如雨后春笋一般涌现, 重新考察这些新业务的特点也就至关重要了. 本文以即时消息业务为例, 了解当前移动互联网下的流量新特点. 为了得到可信的结果, 采集了大量的蜂窝网即时消息实测数据, 并从用户层面和基站层面两个维度去研究其统计规律. 首先在用户层面, 通过用户级的数据包包长和到达间隔建模发现, 相关结果同传统上包长符合几何分布、到达间隔满足指数分布的 3GPP 报告明显不同; 与之相反, 微信包长和到达间隔分别遵循幂律和对数正态分布. 其次在基站层面, 通过随机选取基站发现,  $\alpha$ -稳定模型能更好地刻画基站层面基站流量——这一结果同传统固定宽带网络业务研究相吻合. 最后, 建立了用户层面数据包同基站层面流量的理论联系.

**关键词** 业务特征, 即时消息, 统计规律建模, 重尾分布,  $\alpha$ -稳定模型

## 1 引言

近年来, 随着互联网的日益普遍, 腾讯 QQ<sup>1)</sup>和 Skype<sup>2)</sup>等即时消息应用走进了千家万户, 依托互联网发送和接收文字、图片等已经成为网民的风尚; 而随着智能终端的普及和移动互联网的发展, 微信 (Wechat)<sup>3)</sup>、Whatsapp Messenger<sup>4)</sup>和 LINE<sup>5)</sup>等移动即时消息应用也如雨后春笋般涌现出来. 这些基于互联网数据连接应用的机制相较于使用标准通信协议的传统短信业务有很大不同: 对于传统短

1) Tencent Inc. QQ international chat, video calls, groups - get a better communication experience. <http://www.imqq.com/>, 2001.

2) Microsoft Inc. Skype. <http://www.skype.com/>, 2013.

3) Tencent Inc. Wechat — the new way to connect. <http://www.wechat.com/en/>, 2011.

4) WhatsApp Inc. Whatsapp messenger. <https://www.whatsapp.com/>, 2009.

5) LINE Corporation. LINE: free calls & messages. <http://line.me/en/>, 2011.

**引用格式:** 李荣鹏, 赵志峰, 张宏纲, 等. 移动业务特征认知及其分布模型: 以即时消息为例. 中国科学: 信息科学, 2017, 47: 637-647, doi: 10.1360/N112017-00044

Li R P, Zhao Z F, Zhang H G, et al. The characteristics study of mobile instantaneous messaging traffic in cellular network (in Chinese). Sci Sin Inform, 2017, 47: 637-647, doi: 10.1360/N112017-00044

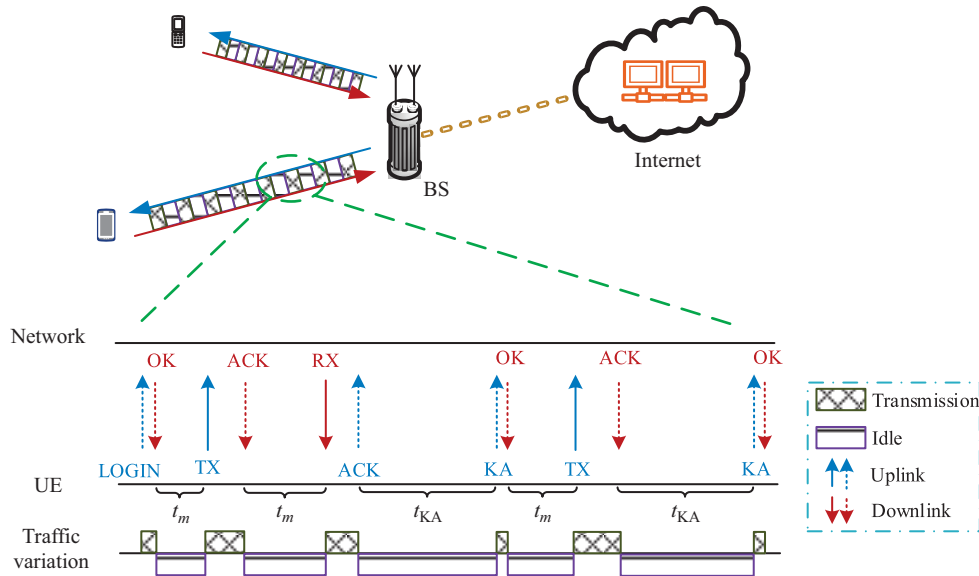


图 1 (网络版彩图) 即时消息业务中终端和网络交互流程示意图

Figure 1 (Color online) The working mechanism of mobile instantaneous messaging service in cellular networks

信业务, 得益于蜂窝网中一套较为完善的移动性管理 (mobility management) 策略可以粗略获取终端的大致位置, 当有传统的短信到达时, 只需要蜂窝网在特定区域内下发寻呼 (paging) 消息就可以找到用户, 再利用电路交换中标准的通信协议承载不超过 140 字的内容 [1]. 而即时消息业务利用分组交换的数据连接发送包括文字在内的各种多媒体内容, 这就需要会话保持状态下才可以收到消息. 为了防止没有数据内容传送导致会话超时、连接被释放, 即时消息软件需要定时发送一个几乎没有数据内容的消息来“保活” (keep-alive, KA) 会话连接 —— 这种保活连接的消息往往又被称作“心跳”消息. 基于此, 图 1 描述了移动即时消息的终端软件和网络交互过程: 用户登录 (login) 后, 开始陆续地发送和接收消息 (图 1 中的 TX 和 RX 数据包); 当完成数据发送后, 接收端会回复应答消息 (图 1 中的 ACK 数据包); 为了保证用户及时接收到网络发来的消息并“保活”网络连接, 软件会向服务端以周期  $t_{KA}$  发送“心跳”消息; 无论“心跳”消息还是登录消息, 服务端收到后都会回应成功消息 (图 1 中的 OK 数据包) 给用户.

由于移动即时消息的普及, “心跳”消息设置频繁甚至不合理, 造成非常大的网络资源开销, 并对移动网络的稳定性产生影响 [2,3]. 近年来的研究表明, 在传统宽带网络中即时消息应用业务呈现出幂律分布、Weibull 分布和对数正态分布等重尾现象 [4~6]; 而移动网领域重要的标准化组织 3GPP 针对即时消息业务进行建模问题, 建议消息到达时间和消息分布分别使用指数分布和几何分布 [7]. 因此, 基于充实的数据, 重新审视移动消息的业务特征是非常必要的. 鉴于此, 本文通过分析采集到的蜂窝网即时消息流量数据, 分别研究微信业务流量在用户层面 [8] 和基站层面 [9] 的统计规律. 具体而言, 用户层面是指流经用户级的数据包的统计规律, 而基站层面是指一段时间内通过某个基站的数据 —— 显然基站层面流量应该是基站内用户数据包的叠加, 探求其中的理论关联性是一件很有意思的工作.

在第 2 节首先介绍文章使用的重尾分布和  $\alpha$ - 稳定分布的数学背景; 在第 3 节重点介绍文章使用的分析数据集和分析结果; 并在第 4 节总结本文的主要工作.

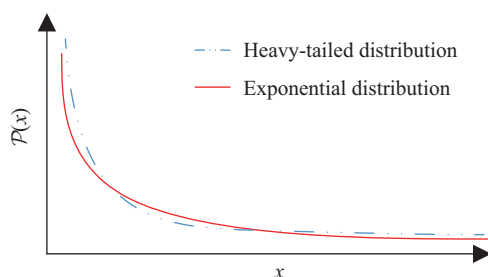


图 2 (网络版彩图) 重尾分布同指数分布的比较示意图

Figure 2 (Color online) The comparison between heavy-tailed distribution and exponential distribution

## 2 数学模型介绍

### 2.1 重尾分布背景介绍

自然界中广泛存在着一种现象, 比如空气中占比高达 99% 的只有氮气和氧气两种气体, 又如 80% 社会财富集中在 20% 的人手中, 海量的英文单词中仅有少数几个词汇经常被使用等; 换句话说, 少量个体往往做出大量贡献, 占据大量资源. 即使在信息产业界, 针对网络拓扑等研究也表明, 少量网络节点通常会处理大量数据<sup>[10,11]</sup>. 正因如此, 重尾分布 (heavy-tail phenomena) 被广泛地用来描述这种现象. 准确地讲, 如图 2 所示, 重尾分布意味着概率密度函数 (probability density function, PDF)  $P(X)$  的“尾部”比指数分布更大 (“重”)<sup>[12]</sup> 的分布, 即满足如下定义.

**定义 1** 当服从某分布的随机变量  $X$  的概率密度函数  $P(X)$  满足式 (1) 时,  $X$  被称作服从重尾分布:

$$\lim_{x \rightarrow \infty} e^{\mu x} P(X \geq x) = \infty, \quad \forall \mu > 0, \quad (1)$$

其中  $\mu$  是衡量重尾程度的一个参数.

除了解释前述自然现象外, 重尾分布还有一些奇怪但非常“漂亮”的性质<sup>[12]</sup>: 它拥有无穷大的方差, 甚至可能有无穷大的均值; 服从重尾分布的事件中远大于均值的也会比其他分布更为“频繁地”发生.

重尾分布的例子有很多, 包括幂律分布 (power-law distribution, 又被称作 Pareto 分布)、Weibull 分布 (Weibull distribution)、对数正态分布 (lognormal distribution) 以及第 2.2 小节介绍的  $\alpha$ -稳定模型 (分布) 等. 其中, 最经典的当属幂律分布, 其累计密度函数 (cumulative density function, CDF) 表达式为

$$P(X \geq x) = \begin{cases} \left(\frac{x_m}{x}\right)^\alpha, & x \geq x_m; \\ 1, & x < x_m. \end{cases} \quad (2)$$

通过对 CDF 进行求导, 可以得到相应的 PDF 表达形式:

$$P(X \geq x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, & x \geq x_m; \\ 1, & x < x_m. \end{cases} \quad (3)$$

特别地, 非零部分可以表述成  $ax^{-b}$  的形式, 这里  $a$  和  $b$  是分布中的未知参数. 幂律分布的一个重要性质是具有尺度不变 (scale invariance) 的特点, 即  $P(X \geq \lambda x) = \left(\frac{x_m}{\lambda x}\right)^\alpha = P(X \geq x) \left(\frac{x_m}{\lambda x}\right)^\alpha$ .

另外, 相较于无记忆的指数分布, 幂律分布则具有记忆性, 即  $\mathcal{P}(X \geq \lambda x + t | X \geq \lambda x)$  与  $x$  大小<sup>6)</sup>.

另一类重要的分布是对数正态分布, 其名称则是因为对于服从对数正态分布的  $X$ , 其对数  $Y = \ln(X)$  将服从正态分布. 对数正态分布的概率密度函数如下:

$$\mathcal{P}(X \geq x) = \frac{1}{\sqrt{2\pi}bx} e^{-\frac{(\ln x - a)^2}{2b^2}}, \quad (4)$$

$a$  和  $b$  是分布中的未知参数. 根据式 (4) 可以发现

$$\ln \mathcal{P}(X) |_{X=x} = -\frac{1}{2b^2} (\ln x)^2 + \left(\frac{a}{b^2} - 1\right) \ln x - \ln \sqrt{2\pi} - \frac{a^2}{2b^2}, \quad (5)$$

这样, 当  $b^2 \gg 1$  并且  $b^2 \gg a$  时,  $\ln \mathcal{P}(X) |_{X=x} \sim -\ln x + C$  ( $C$  表示一个常量), 换句话说, 近似为幂律分布. 因此对数正态分布和幂律分布性质极其相似, 但一般认为随机相乘的增长方式将产生对数正态分布; 但在此基础上对最小值加以限制的情况将产生幂律分布<sup>[13]</sup>.

Weibull 分布也属于重尾分布, 其概率密度分布则为

$$\mathcal{P}(X \geq x) = \begin{cases} abx^{b-1}e^{-ax^b}, & x \geq 0; \\ 0, & x < 0, \end{cases} \quad (6)$$

$a$  和  $b$  表示分布中的未知参数.

## 2.2 $\alpha$ - 稳定模型介绍

根据归一化中心极限定理 (generalized central limit theorem),  $\alpha$ - 稳定模型<sup>7)</sup> 描述了相当大的独立同分布 (independent identically distributed) 随机变量归一化求和后的分布<sup>[14]</sup>. 此外  $\alpha$ - 稳定模型能够产生“爆发性”和长时相关性 (long range dependence). 正因如此,  $\alpha$ - 稳定模型很自然地用于刻画传统宽带网络的流量分布<sup>[15, 16]</sup>, 并且被广泛应用于资源优化的建模中<sup>[17, 18]</sup>.

一般而言, 除了少数特例外,  $\alpha$ - 稳定模型对于概率密度函数缺少闭式 (closed-form) 的表达形式. 退而求其次,  $\alpha$ - 稳定模型通常由其特征函数决定. 数学意义上讲,  $\alpha$ - 稳定模型应该满足下述定义.

**定义2** 一个随机变量  $X$  被称作遵守  $\alpha$ - 稳定模型, 需要遵守: 存在变量  $0 < \alpha \leq 2$ ,  $\sigma \geq 0$ ,  $-1 \leq \beta \leq 1$  和  $\mu \in \mathbb{R}$  使变量  $X$  的特征函数 (characteristic function) 符合下式:

$$\begin{aligned} \Phi(\omega) &= E(\exp j\omega X) \\ &= \begin{cases} \exp \left\{ -\sigma^\alpha |\omega|^\alpha \left( 1 - j\beta(\operatorname{sgn}(\omega)) \tan \frac{\pi\alpha}{2} \right) + j\mu\omega \right\}, & \alpha \neq 1; \\ \exp \left\{ -\sigma |\omega| \left( 1 + j\frac{2\beta}{\pi}(\operatorname{sgn}(\omega)) \ln |\omega| \right) + j\mu\omega \right\}, & \alpha = 1, \end{cases} \end{aligned} \quad (7)$$

这里,  $\alpha$  是属性指数 (characteristic exponent), 表征了分布稳定程度;  $\beta$  表示  $\alpha$ - 稳定模型的偏度 (skewness);  $\alpha$  和  $\beta$  决定了模型的基本形状. 此外,  $\sigma$  和  $\mu$  分别表示尺度 (scale) 和偏移 (shift). 特别地, 当  $\alpha = 2$  时  $\alpha$ - 稳定模型可以简化成正态分布.

6) 与记忆性密切相关的一类分布被称作长尾分布 (long-tail distribution), 相关讨论可以参阅文献 [10].

7)  $\alpha$ - 稳定模型, 又被简称作稳定模型, 同时也有文献将其称为 Levy Skew  $\alpha$ -Stable 分布. 在本文中, 这几个名称除非特殊说明, 则表述相同含义.

$\alpha$ -稳定模型具有很多重要的性质,最重要的当属独立同分布服从  $\alpha$ -稳定模型的随机变量之和及它们本身具有相同的分布.此外,中心极限定理阐明了随着有限方差的随机变量数量增长,它们的和的分布趋向正态分布;而中心极限定理的一个重要推论则表明,随着服从以  $|x|^{-\alpha-1}$  ( $0 < \alpha \leq 2$ ) 递减的幂律尾分布(因此具有无限方差)的随机变量数量增长,它们的和的分布趋向  $\alpha$ -稳定模型<sup>[19,20]</sup>.另外,对于服从  $\alpha$ -稳定模型的随机变量  $X$ ,其采样特征函数  $\Psi(\omega) = \ln\{-\text{Re}[\ln(\Phi(\omega))]\}$  在  $\ln(\omega)$  之间存在线性关系

$$\Psi(\omega) = \ln\{-\text{Re}[\ln(\Phi(\omega))]\} = \alpha \ln(\omega) + \alpha \ln(\sigma). \quad (8)$$

由于  $\alpha$ -稳定模型的概率密度函数缺乏闭式表达式,因而证明一个分布服从  $\alpha$ -稳定模型将非常富有挑战性.这样,当描述一个数据集满足  $\alpha$ -稳定模型时,这通常意味着数据集反映出的一些特点同假设检验得到的  $\alpha$ -稳定模型和它的性质相吻合.换句话说,  $\alpha$ -稳定模型的验证过程需要首先利用数据集估计  $\alpha$ -稳定模型中的未知参数;接下来将数据集的实际分布和  $\alpha$ -稳定模型模拟出来的分布进行比照<sup>[16]</sup>.特别地,  $\alpha$ -稳定模型中的参数可以使用最大似然估计法、分位数法(quantile methods)和采样特征函数法(sample characteristic function methods)等方法求得<sup>[15,16]</sup>.

### 3 微信业务的时域统计建模

#### 3.1 分析数据集介绍

本文将以下流行的移动消息应用微信为基础,利用现网数据分析移动即时消息的时域业务特点及统计规律.腾讯公司开发的微信提供文字、语音和图片等一系列多媒体应用,目前已经拥有 3 亿活跃用户,所呈现出来的业务特征因此将具有很大的代表性.为了夯实分析结论的有效性,本节使用在 Gb 和 Gn 接口(可以参见图 3 所示现行 GSM/UMTS/LTE 混合组网的网络架构图)采集的现网 2013 年 4 月的微信业务数据集,其中涉及 700 万用户及服务其的 2G 和 3G 基站,包括时间戳、小区号、消息长度和类型等基本消息.基于这些采集到的数据,本节将计算得到微信业务的消息长度和消息到达间隔的概率密度函数  $\mathcal{P}(L)$  和  $\mathcal{P}(T)$ ,然后将常见的概率分布(包括指数分布、幂律分布、几何分布、对数正态分布和 Weibull 分布)和  $\mathcal{P}(L)$  或  $\mathcal{P}(T)$  分别用最大似然估计(maximum likelihood estimation, MLE)进行拟合,求得模型中的参数;接着,使用均方根误差(root mean square error, RMSE)评估各个候选分布模型的准确性,从而得到更为匹配的概率分布模型.图 4 给出了在不同时刻(上午 4 时、10 时,下午 4 时、10 时),一个随机选择区域(活跃基站数 23 个)的即时消息业务在基站层面的流量密度,从中可以清楚地发现用户微信业务在基站层面的稀疏特性.

#### 3.2 用户层面数据包

图 5 画出了微信消息长度使用常见分布进行拟合的概率密度函数  $\mathcal{P}(L)$  曲线,并提供了常见分布的拟合结果.从图 5 可以看出,幂律分布和消息长度的分布最为匹配;而表 1 也表明,幂律分布的均方根误差最小.因此,在所列的几种分布中,幂律分布( $0.347x^{-2.407}$ )最为准确地拟合消息长度.相对而言,3GPP 报告<sup>[7]</sup>中认为的几何分布有很大不准确性.

另一方面,图 6 画出了仅仅考虑用户消息时微信消息到达间隔分布的概率密度函数  $\mathcal{P}(T)$  曲线.在常见的概率分布中,从图 6 可以看出对数正态分布( $\frac{1}{\sqrt{2\pi \times 2.975x}} e^{-\frac{(\ln x - 2.36)^2}{2 \times 2.975^2}}$ )进行建模最为准确;表 1 则表明,对数正态分布的 RMSE 最小,并远远超过 3GPP 认为的服从指数分布的消息到达间隔模型.

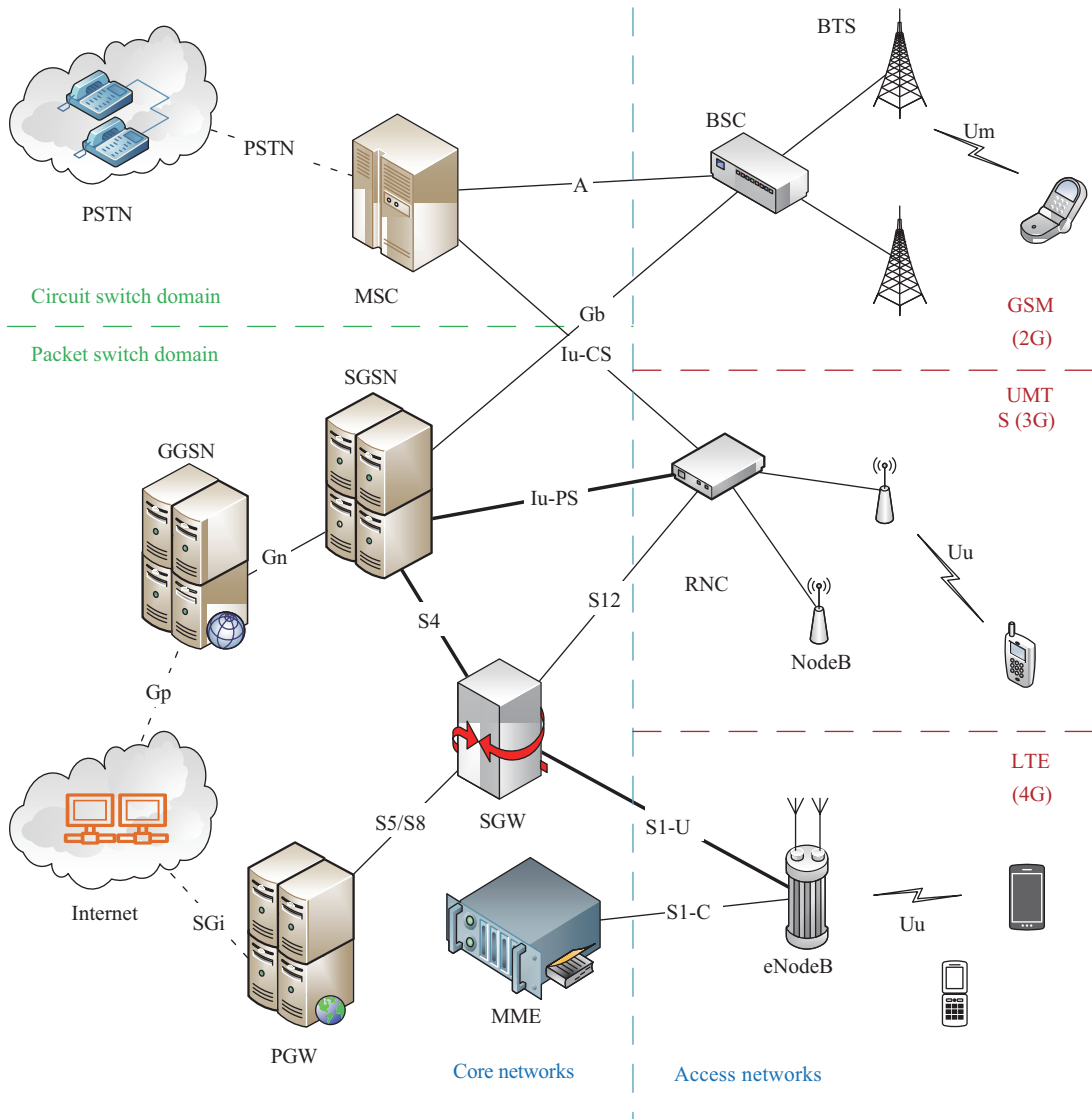


图 3 (网络版彩图) 现行 GSM/UMTS/LTE 混合组网的网络架构图

Figure 3 (Color online) The network architecture of GSM/UMTS/LTE co-operating cellular networks

当考虑心跳消息时, 微信消息到达间隔分布会发生变化, 在图 6 中会表现为一些尖峰. 由于微信版本和用户终端的原因, 微信“心跳”消息间隔会略有不同, 但通常为 30 s 的倍数. 此外, 可以发现当统计到达间隔超过 300 s (即“心跳”消息的最大心跳间隔) 时, 消息到达间隔的比例下降很快, 超过 300 s 的“心跳”消息仅占总数的 2%. 此外, 数据分析结果还证明了, 消息长度和到达间隔统计分布是相互独立的.

综上所述, 通过微信相关的现网数据分析可以发现, 移动消息业务的消息长度可以使用幂律分布进行拟合建模, 消息到达间隔可以使用对数正态分布进行建模分析. 换句话说, 移动消息业务同样呈现出明显的重尾特征.

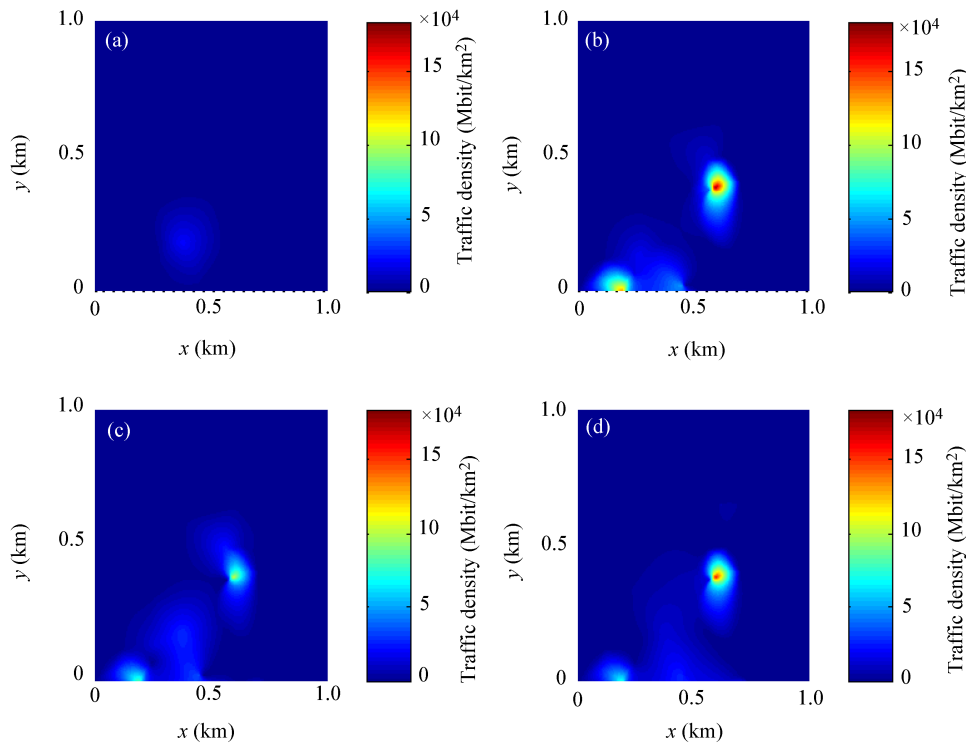


图 4 (网络版彩图) 上午 4 时、10 时, 下午 4 时、10 时, 一个随机选择区域 (活跃基站数 23 个) 的即时消息业务在基站层面的流量密度

Figure 4 (Color online) The spatial traffic density of mobile instantaneous messaging service in a random selected region with 23 active base stations. (a) 4:00 a.m.; (b) 10:00 a.m.; (c) 4:00 p.m.; (d) 10:00 p.m.

表 1 基于 RMSE 的微信数据建模准确性权重

Table 1 The modeling accuracy in terms of RMSE for Wechat

Name	PDF	Message length	Inter-arrival time	Aggregated traffic
Power-law	$ax^{-b}$	9.76E-5	9.25E-5	0.0357
Geometric	$(1-a)^x a$	607E-5	48.0E-5	0.0258
Exponential	$ae^{-bx}$	56.0E-5	22.9E-5	0.0899
Weibull	$abx^{b-1}e^{-ax^b}$	65.8E-5	8.08E-5	0.0470
Lognormal	$\frac{1}{\sqrt{2\pi}bx} e^{-\frac{(\ln x - a)^2}{2b^2}}$	34.0E-5	7.44E-5	0.0491
$\alpha$ -stable	-	790E-5	170E-5	0.0144

### 3.3 基站层面流量建模

图 7 描绘了某一随机选择基站内基站层面流量分布情况, 并比较了不同分布模型的建模结果. 同用户层面业务建模类似, 图 7 表明, 基站层面流量用  $\alpha$ - 稳定模型建模最为准确. 与此同时, 表 1 表明,  $\alpha$ - 稳定模型的均方根误差最小. 更进一步, 图 8(a)~(d) 刻画了  $\alpha$ - 稳定模型建模另外 4 个随机基站的结果; 第 2.2 小节表明使用  $\alpha$ - 稳定模型建模时,  $\Psi(\omega)$  和  $\ln(\omega)$  具有线性关系, 据此, 图 8(e) 刻画了所有基站内线性函数建模误差的累计概率密度函数 (cumulative distribution function, CDF). 从图 8(e)

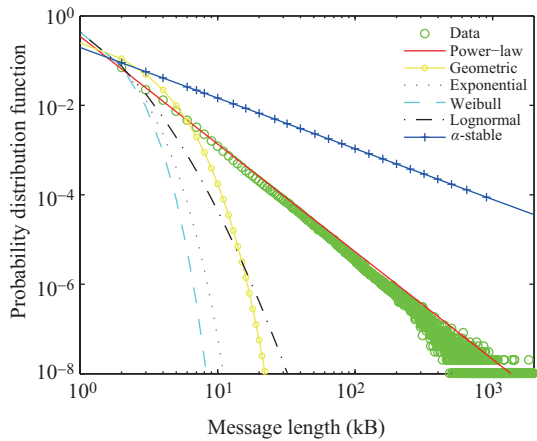


图 5 (网络版彩图) 微信消息长度的概率密度函数及拟合结果  
 Figure 5 (Color online) The PDF and fitting results of message length for Wechat

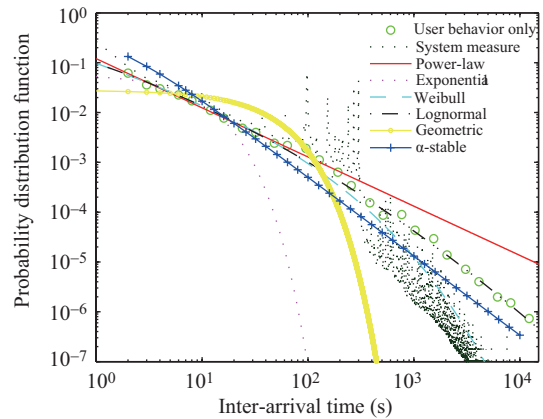


图 6 (网络版彩图) 微信消息到达间隔分布的概率密度函数及拟合结果  
 Figure 6 (Color online) The PDF and fitting results of inter-arrival time for Wechat

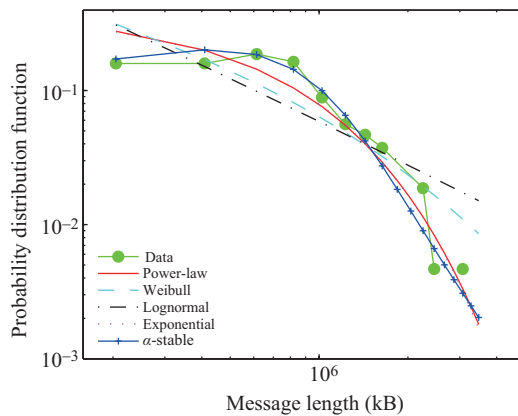


图 7 (网络版彩图) 在一个随机选择基站内, 基站层面微信流量建模结果

Figure 7 (Color online) The modeling results of the aggregated traffic in a randomly selected base station for Wechat

不难看出, 误差很小, 换句话说, 图 8(e) 表明了基站层面流量符合  $\alpha$ - 稳定模型.

另一方面,  $\alpha$ - 稳定模型预示着以微信为代表的即时通讯业务的自相似性<sup>[19]</sup>. 而即时消息业务之所以满足  $\alpha$ - 稳定模型, 是因为用户层面数据包长符合幂律分布, 一段时间内基站层面流量可以看做大量用户层面数据包的叠加, 那么根据广义中心极限定理, 大量满足幂律分布  $|x|^{-\alpha-1}$ , 其中  $0 < \alpha < 2$  随机变量的和, 将随着变量数目的增多而服从方差无限大的  $\alpha$ - 稳定模型. 有意思的是, 图 8(f) 给出了基站流量用  $\alpha$ - 稳定模型建模后, 其  $\alpha$  值所遵循的概率密度函数, 从中可以看出,  $\alpha$  值大部分位于 1.136 至 1.515 的区间, 而用户层面幂律分布建模的斜率为 2.407, 这些数据有力地印证了广义中心极限定理<sup>[20]</sup>.

值得注意的是,  $\alpha$ - 稳定模型相较于对数正态分布等模型虽然具有更好的建模准确性, 但其模型参数也更多, 因此参数拟合复杂度也会更高, 对模型的深入研究、理论分析也就更为困难. 幸运的是, 得



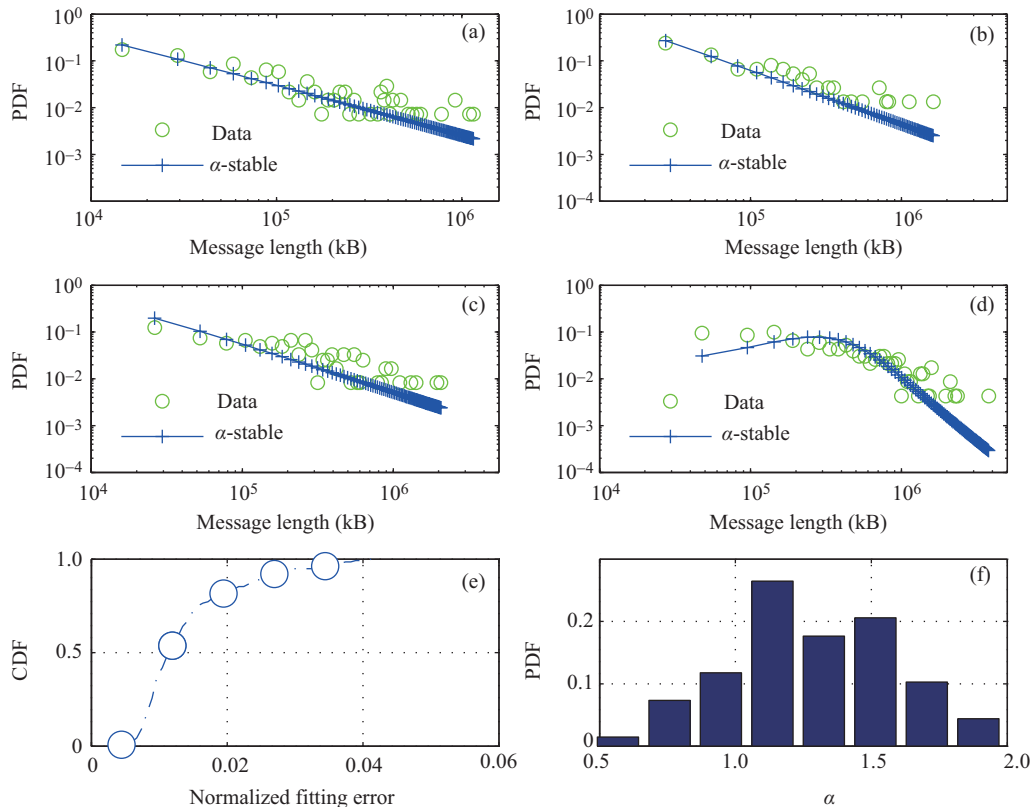


图 8 (网络版彩图) (a)~(d) 4 个随机选择基站  $\alpha$ - 稳定模型的建模准确度; (e) 在所有基站  $\Psi(\omega)$  同  $\ln(\omega)$  线性拟合误差的概率密度函数; (f) 所有基站内  $\alpha$  值的分布情况

**Figure 8** (Color online) (a)~(d) The accuracy after fitting the aggregated traffic in four randomly selected base stations by  $\alpha$ -stable models; (e) the PDF of fitting  $\Psi(\omega)$  with respect to  $\ln(\omega)$  by using a linear function; (f) the PDF of fitted  $\alpha$  values

益于广义中心极限定理带来的普遍意义,  $\alpha$ - 稳定模型在统计学的理论研究已经积累了一系列的研究成果<sup>[20]</sup>, 并为应用于基站层面的流量预测<sup>[16, 21]</sup> 奠定了基础, 从这个角度说,  $\alpha$ - 稳定模型依然具有广阔的应用场景.

## 4 总结

本文分析了微信为代表的即时消息业务并得出了如下结论: (1) 与 3GPP 报告认为的几何分布和指数分布不同, 幂律分布和对数正态分布可以更为准确地拟合其消息长度和消息到达间隔分布, 换句话说, 微信为代表的即时消息业务在统计规律上呈现出明显的重尾现象; (2)  $\alpha$ - 稳定模型可以用于刻画移动互联网中即时消息、业务在基站层面流量分布; (3) 根据广义中心极限定理, 本文还解释了即时消息业务的统计规律同基站层面  $\alpha$ - 稳定模型的关系.

## 参考文献

- 1 European Telecommunications Standards Institute. Services and facilities to be provided in the GSM system. GSM Doc 28/85. [http://www.etsi.org/deliver/etsi\\_gts/01/0102/05.00.00.60/gsmsts\\_0102v050000p.pdf](http://www.etsi.org/deliver/etsi_gts/01/0102/05.00.00.60/gsmsts_0102v050000p.pdf)

- 2 Sandle P. O2 says iPhone demand strained its London network. Reuters, 2009. <http://www.reuters.com/article/2009/12/29/o2-iphone-idUSLDE5BS0M520091229>
- 3 Donegan M. Android signaling storm rises in Japan. Light Reading, 2012. <http://www.lightreading.com/mobile/device-operating-systems/android-signaling-storm-rises-in-japan/a/d-id/693138>
- 4 Leskovec J, Horvitz E. Planetary-scale views on a large instant-messaging network. In: Proceedings of the 17th International Conference on World Wide Web, Beijing, 2008. 915–924
- 5 Xiao Z, Guo L, Tracey J. Understanding instant messaging traffic characteristics. In: Proceedings of the 27th International Conference on Distributed Computing Systems, Toronto, 2007. 51
- 6 Downey A B. Lognormal and pareto distributions in the Internet. *Comput Commun*, 2005, 7: 790–801
- 7 3GPP. GERAN study on mobile data applications. 3GPP TR 43.802. [http://www.3gpp.org/ftp//Specs/archive/43\\_series/43.802/](http://www.3gpp.org/ftp//Specs/archive/43_series/43.802/)
- 8 Zhou X, Zhao Z F, Li R P, et al. Understanding the nature of social mobile instant messaging in cellular networks. *IEEE Commun Lett*, 2014, 3: 389–392
- 9 Li R P, Zhao Z F, Qi C, et al. Understanding the traffic nature of mobile instantaneous messaging in cellular networks: a revisiting to  $\alpha$ -stable models. *IEEE Access*, 2015, 3: 1416–1422
- 10 Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the Internet topology. In: Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, Cambridge, 1999. 251–262
- 11 Mahanti A, Carlsson N, Arlitt M, et al. A tale of the tails: power-laws in Internet measurements. *IEEE Netw*, 2013, 1: 59–64
- 12 Nair J, Wierman A, Zwart B. The fundamentals of heavy tails: properties, emergence, and identification. In: Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems, Pittsburgh, 2013. 387–388
- 13 Dodds P. Lognormals and friends. Technical Report. 2013. <http://www.uvm.edu/~pdodds/teaching/courses/2013-01UVM-300/docs/2013-01UVM-300lognormals-flat.pdf>
- 14 Samorodnitsky G. Stable Non-Gaussian Random Processes: Stochastic Models With Infinite Variance. New York: Chapman and Hall/CRC, 1994
- 15 Gallardo J R, Makrakis D, Orozco-Barbosa L. Use of alpha-stable self-similar stochastic processes for modeling traffic in broadband networks. In: Proceedings of the SPIE International Society for Optical Engineering, Boston, 1998. 281–296
- 16 Ge X H, Zhu G X, Zhu Y T. On the testing for alpha-stable distributions of network traffic. *Comput Commun*, 2014, 5: 447–457
- 17 Song W, Zhuang W. Resource reservation for self-similar data traffic in Cellular/WLAN integrated mobile hotspots. In: Proceedings of the IEEE International Conference on Communications, Cape Town, 2010. 1–5
- 18 Chuang J C-I, Sollenberger N. Spectrum resource allocation for wireless packet access with application to advanced cellular Internet service. *IEEE J Sel Area Commum*, 1998, 6: 820–829
- 19 Crovella M, Bestavros A. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Trans Netw*, 1997, 6: 835–846
- 20 Kolmogorov A N, Chung K L, Gnedenko B V. Limit Distributions for Sums of Independent Random Variables. 2nd ed. Upper Saddle River: Addison-Wesley, 1968
- 21 Li R P, Zhao Z F, Zheng J C, et al. The learning and prediction of application-level traffic data in cellular networks. arXiv:1606.04778, 2016

# The characteristics study of mobile instantaneous messaging traffic in cellular network

Rongpeng LI<sup>1\*</sup>, Zhifeng ZHAO<sup>1</sup>, Honggang ZHANG<sup>1</sup> & Xiaofeng ZHONG<sup>2</sup>

1. *College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China;*

2. *Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*

\* Corresponding author. E-mail: lirongpeng@zju.edu.cn

**Abstract** Understanding traffic characteristics plays a vital role in network protocol design, which aims to minimize network resource consumption and maximize network stability. In this paper, we use the example of the mobile instantaneous messaging (MIM) service in cellular networks and try to understand its traffic characteristics. Specifically, in order to reach credible conclusions, our research uses practical measurement records from the MIM services of China Mobile at two different levels. First, a dataset of individual message level (IML) traffic is examined and reveals power-law distributed message length and lognormal distributed inter-arrival time. The heavy-tailness of which completely diverts from the geometric and exponential models recommended by 3GPP. Second, another dataset is used to examine the statistical patterns of aggregated traffic within a base station, and demonstrates the accuracy of  $\alpha$ -stable models for aggregated traffic. Furthermore, we verify that  $\alpha$ -stable models are suitable for characterizing traffic in both conventional fixed core networks and cellular access networks. Finally, by using the generalized central limit theorem, we build a theoretical relationship between the distributions of IML traffic and aggregated traffic.

**Keywords** traffic characteristics, mobile instantaneous messaging (MIM) services, statistical modeling, heavy-tail distribution,  $\alpha$ -stable models



Rongpeng LI received his B.E. and Ph.D. degrees from Xidian University, Xi'an, China and Zhejiang University, Hangzhou, China in June 2010 and June 2015, respectively, both as "Excellent Graduates". He is currently a postdoctoral researcher at the College of Computer Science and Technologies, Zhejiang University, Hangzhou, China. His research interests include applications of artificial intelligence, data-driven network design, and resource allocation for cellular networks (especially full-duplex networks).



Zhifeng ZHAO is an associate professor at the College of Information Science and Electronic Engineering, Zhejiang University, China. He received a Ph.D. degree in communication and information system from the PLA University of Science and Technology, Nanjing, China, in 2002. From September 2002 to December 2004, he was a postdoctoral researcher at Zhejiang University. His research interests include cognitive radio, wireless multi-hop networks (Ad Hoc, Mesh, WSN, etc.), wireless multimedia networks, and green communications.



Honggang ZHANG is a professor at Zhejiang University, China and was the international chair professor of Excellence for Université Européenne de Bretagne and Supélec, France from 2012 to 2014. He is also an honorary visiting professor of the University of York, UK. He served as the chair of the Technical Committee on Cognitive Networks of the IEEE Communications Society from 2011 to 2012. He is currently involved in research on green communications and has the role of series editor for the IEEE Communications Magazine series on Green Communications and Computing Networks.



Xiaofeng ZHONG received his Ph.D. degree from Tsinghua University in 2005 and joined Tsinghua immediately thereafter. In 2008, he became the Master's Advisor for the Department of Electronic Engineering at Tsinghua. His research interests include traffic modes and services in mobile networks, and network architecture and protocol design for wireless communication systems, especially on the medium access control (MAC) and network layers. His recent research has focused on energy efficiency mobile service in 5G systems based on user behaviors.