计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(10):325−330 [doi: 10.15888/j.cnki.csa.008160]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

微信流量特性分析与建模①

龚　莲, 谭献海

(西南交通大学 信息科学与技术学院, 成都 611756)
通讯作者: 龚　莲, E-mail: 304260562@qq.com

摘　要: 微信是现代互联网的主要应用之一, 到目前为止有关微信流量特性分析与建模的研究较少. 本文以微信流量为研究对象, 分析验证微信流量同时具有自相似性和突发性. 针对这两种特性进行微信流量建模, 采用线性分形稳定噪声模型刻画微信流量特性, 完成了模型的参数估算和效果分析. 本文的研究成果是后续的网络性能分析、网络流量监管等的基础.

关键词: 微信; 网络流量; 自相似性; 突发性; 线性分形稳定噪声模型

引用格式: 龚莲,谭献海.微信流量特性分析与建模.计算机系统应用,2021,30(10):325–330. http://www.c-s-a.org.cn/1003-3254/8160.html

Characteristics Analysis and Modeling of WeChat Network Traffic

GONG Lian, TAN Xian-Hai

(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: WeChat is one of the main applications of the modern Internet, but there are few studies on the characteristics analysis and modeling of its network traffic. The study takes WeChat traffic as the research object and finds that it has self-similarity and burstiness. In view of these two characteristics, we use the linear fractional stable noise model to characterize WeChat traffic and carry out the parameter estimation and the effect analysis of the model. The research results provide a basis for subsequent network performance analysis and traffic monitoring.

Key words: WeChat; network traffic; self-similarity; burstiness; linear fractional stable noise model

早期的网络流量呈短相关性, 使用 Possion 或者 Markov 过程描述. 随着 Leland 发现局域网流量的自相似长相关性[1], 大量的研究结果表明传统互联网流量具有普遍的自相似长相关性, 因此许多学者提出了长相关流量模型, 包括 ON/OFF 模型、FARIMA 模型、FBM 和 FGN 模型等. 现代互联网无论是应用类型还是用户数量都与早期的互联网有较大的区别, 其网络流量特性也随之改变.

据研究机构 Trustdata 发布的《2020 年 Q1 中国移动互联网行业分析报告》显示[2], 微信在国内 APP 排行榜位列第一, 明显超过其他网络应用. 作为目前拥有最高用户活跃数的应用, 微信流量特性受到用户参与行为的深度影响. 目前关于微信流量的研究主要包括: 李玮提出一种基于 DPI 的识别方法对微信流量进行识别研究, 基于业务特征进行微信业务的识别与分类[3] . 燕飞鹏提出一种基于随机森林算法的微信流量分类模型, 基于流量分类提出微信用户阶段性行为识别技术[4]. 张江楠对微信流量进行特性分析, 发现微信流量呈自相似特性和幂律特性[5].

综上关于微信流量的研究多集中于流量识别、业务分类等方面, 缺乏微信流量特性分析与建模的研究. 分析微信流量特性并用时间序列建模是流量预测的基本原理, 基于模型预测可以研究微信流量在网络系统中的拥塞控制机制, 此外还可以依据微信流量模型计

① 基金项目: 国家科技支撑计划 (2015B14B01)
Foundation item: National Science and Technology Support Program of China (2015B14B01)
收稿时间: 2021-01-06; 修改时间: 2021-02-03, 2021-03-05; 采用时间: 2021-03-11

© 中国科学院软件研究所 http://www.c-s-a.org.cn

算流量在网络传输排队过程中的时延、丢包率和队列平均长度等网络性能指标. 微信流量作为互联网流量的核心入口, 对其进行研究可以为网络流量控制管理提供依据.

1 研究思路与研究数据

本文首先通过 Matlab 直观观察微信流量可能具有哪些特性, 然后定量分析微信流量确实具有这些特性. 在此基础上对微信流量进行建模, 模型中包含能同时刻画微信流量特性的参数, 最后分析模型效果.

本文的研究数据是在实验室局域网环境下使用 Wireshark 实时抓取 7.0.10 版本的微信自 2019 年 9 月 15 日 9:30–17:30 期间产生的流量, 这些流量由网络通信链路中多个更小的信源产生的流量组成, 并不能代表主干链路的流量, 然而主干链路的流量本质上是多个独立同分布信源流量的叠加, 所以实验室局域网环境下的微信流量与主干链路的微信流量为同一种分布. 通过对抓取的 timestamp、length 等数据项进行处理, 获得单位时间内到达的数据包个数. 为了使数据更具代表性, 除了抓取的微信流量之外, 本文还采用了文献 [5] 中的微信流量数据集. 由于采取的微信流量在该时段的变化趋势基本一致, 所以本文选取某个更小时间片段的流量进行实验分析.

2 微信流量特性分析

2.1 微信流量特性直观分析

首先从直观角度观察微信流量的变化, 不同时间尺度下到达的数据包数量如图 1 所示, 图中的时间间隔为 1 s 和 5 s, 不同时间尺度下数据包的到达数量的曲线变化非常相似, 并且在某些时间间隔出现非常高的数据包到达数量值, 可以直观看出微信流量同时具有自相似性和突发性.

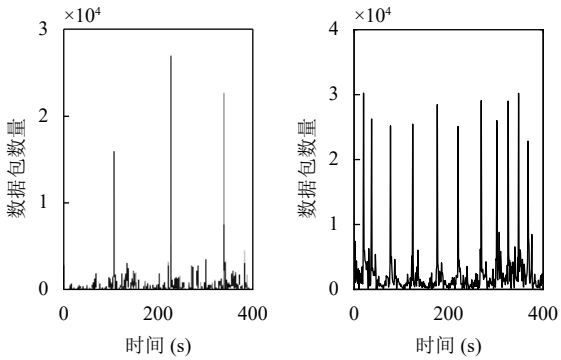
2.2 微信流量自相似性分析

流量自相似性是指流量的时间序列在局部与整体之间具有一定程度的相似, 其数学定义如下:

$$X(\lambda t) = \lambda^H X(t) \tag{1}$$

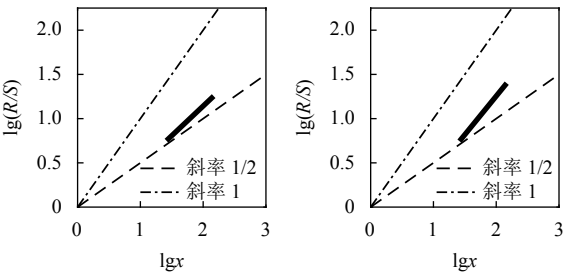
其中, $X(t)$表示第t个单位时间到达的数据包数量, H 为自相似参数[6].

本文对单位时间 1 s 内到达的微信文本类和音视频类流量进行自相似性分析, 采用 R/S 分析法[7] 计算两类流量的H参数如图 2 所示, 图中 x 表示 R/S 分析法中的每个子序列的长度大小, 实线的斜率即为H参数值, 可以看出两类流量的H参数值均满足$0.5 < H < 1$, 大小分别为 0.67、0.87, 从定量角度说明微信文字类和音视频类流量都具有自相似长相关性.



(a) 时间间隔 1 s (b) 时间间隔 5 s

图 1 不同尺度下微信流量数据包到达数量



(a) 微信文本类流量 (b) 微信音视频类流量

图 2 微信文本类和音视频类流量自相似参数估算

2.3 微信流量突发性分析

流量的突发性是指流量在幅度方面的突发, 这是网络流量的另一个特征, α 稳定分布可以很好地描述突发现象. 根据广义中心极限定理, 无穷多个独立同分布随机变量的叠加过程其归一化边缘分布收敛于 α 稳定分布函数簇, 而在网络链路中聚合流量本质上是无穷多个独立同分布信源的叠加, 所以本文采用 α 稳定分布来刻画微信流量的突发性, 其特征函数表示如下:

$$\phi(\theta) = \exp\{-\sigma^{\alpha}|\theta|^{\alpha}[1 - i\beta sgn(\theta)\omega(\alpha,\theta)] + i\mu\theta\} \tag{2}$$

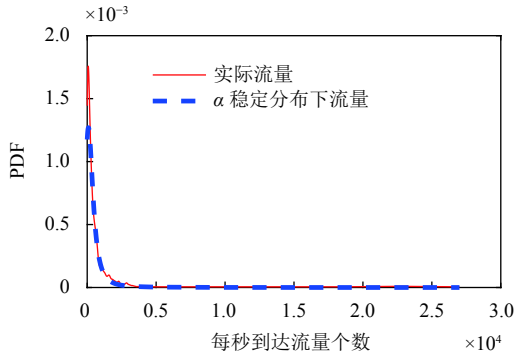
$$sgn(\omega) = \begin{cases} 1, \omega > 0 \\ 0, \omega = 0 \\ -1, \omega < 0 \end{cases}, \omega(\alpha,\theta) = \begin{cases} -\tan\dfrac{\pi\alpha}{2}, \alpha \neq 1 \\ \dfrac{2}{\pi}\ln|\theta|, \alpha = 1 \end{cases} \tag{3}$$

其中, α为特征指数, β为偏斜参数, σ为尺度参数, μ为

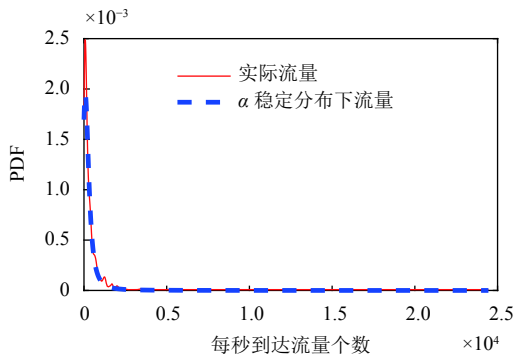
© 中国科学院软件研究所 http://www.c-s-a.org.cn

位置参数[8].

α 稳定分布中只有α参数表示突发程度, 其取值范围为 (0, 2], α越小则突发性越强, $\alpha = 2$时该分布不具有突发性, 所以本文重点关注α参数值. 验证微信流量是否具有突发性的步骤如下: 首先采用分位数法计算微信流量在 α 稳定分布下的 4 个参数值, 然后画出微信流量在该分布下的概率密度曲线 (PDF), 最后比较微信实际流量的 PDF 与 α 稳定分布下流量的 PDF. 通过计算得到微信文字类和音视频类流量的α参数值分别为 1.25、1.24, 说明两类流量都具有较大的突发性, 最后二者的概率密度曲线如图 3 所示.



(a) 微信文本类流量



(b) 微信音视频类流量

图 3 微信文本类和音视频类流量概率密度分布

3 微信流量建模

3.1 线性分形稳定噪声模型

上述流量特性分析表明微信流量同时具有自相似性和突发性, 需要能同时刻画这两种特性的模型对微信流量建模. 分形布朗运动是一种边缘分布为高斯分布的自相似随机过程, 而高斯分布是 α 稳定分布的一种特殊情况, 所以在 α 稳定分布条件下分形布朗运动可以扩展为线性分形稳定运动, 线性分形稳定运动的平稳增量过程是线性分形稳定噪声 (Linear Fractional Stable Noise, LFSN) 过程, LFSN 过程是目前唯一能描述随机变量的自相似性和突发性的随机过程, 其积分表达式的离散形式如下:

$$L_{\alpha,H}(i) = h_d * S^{(\alpha)}_{\beta,\sigma,\mu}(i) = \sum_{j=1}^{Km} h_d(j/m) S^{(\alpha)}_{\beta,\sigma,\mu}(i - j/m) \quad (4)$$

其中, α表示突发程度, H为自相似参数, $Ms(dx)$是具有 Lebesgue 控制测度的 α 稳定随机测度, K为积分截断点, m为积分离散化网络中的控制参数, $S^{\alpha}_{(\beta,\sigma,\mu)}$为 α 稳定分布, h_d为离散内核函数, 表达式如下:

$$h_d = \begin{cases} x^{H-1/\alpha} - (x-1)^{H-1/\alpha}, 1 < x \\ x^{H-1/\alpha}, 0 < x \le 1 \end{cases} \quad (5)$$

基于微信流量的自相似性、突发性和流量在任意时刻的非负性, 本文采用一种偏态 LFSN 过程的模型对微信流量建模[9], 表达式如下:

$$M(i) = c_1 \times \left(h_d * S^{(\alpha)}_{1,1,0}\right)(i) + c_2 \quad (6)$$

其中, $M(i)$是第i个单位时间到达的数据包个数, α表示网络流量的突发系数, 可以使用分位数法[10] 估算, H 是流量的自相似参数, 使用 R/S 分析法估算, c_1表示流量的偏差, c_2表示流量的均值.

3.2 线性分形稳定噪声模型参数估计

对式 (6) 取数学期望得到c_2的估算值:

$$E(M(i)) = c_1 h_d * E\left(S^{(\alpha)}_{1,1,0}(i)\right) + c_2 \quad (7)$$

因为$E\left(S^{(\alpha)}_{1,1,0}(i)\right) = \mu = 0$, 所以$c_2$是网络流量的平均值.

c_1是网络流量的偏差系数, 依据文献 [9] 使用下述公式计算c_1效果更佳:

$$c_1 = (x_{0.72} - x_{0.28})/1.654 \quad (8)$$

$$x_f = X(i) - (X(i+1) - X(i))\frac{f - q(i)}{q(i+1) - q(i)} \quad (9)$$

$$q(i) = (2i-1)/(2N) \quad (10)$$

其中, $X(i)$表示第i个样本数据, N表示样本数据个数, f应满足$(2i-1)/(2N) \le f < (2i+1)/(2N)$.

完成模型参数估算后, 使用文献 [11] 的方法生成 α 稳定分布随机数$S(i)$, 根据式 (5) 生成时间序列$H(i)$, 最后对$S(i)$和$H(i)$作离散傅立叶变换及其逆变换生成$M(i)$序列.

根据 LFSN 模型参数估算方法, 估算微信文字类

© 中国科学院软件研究所 http://www.c-s-a.org.cn

流量的参数值为: $\alpha = 1.25$、$H = 0.67$、$c_1 = 262.35$、$c_2 = 531.94$; 微信音视频类流量的参数值为: $\alpha = 1.24$、$H = 0.87$、$c_1 = 180.77$、$c_2 = 377.36$.

4 模型效果分析

4.1 微信流量建模仿真分析

为了分析 LFSN 模型对微信流量建模的效果, 本文对单位时间 1 s 内到达的微信流量进行建模. 目前关于微信流量特性分析与建模的研究非常少, 仅有文献 [5] 提出使用 Pareto 模型刻画微信流量, 此外 FBM 模型是常用的自相似网络流量模型[12], 所以本文将采用 Pareto 模型和 FBM 模型对微信流量建模, 并与 LFSN 模型效果进行对比, 证明 LFSN 模型的有效性.

基于 LFSN 模型对微信流量进行参数估算得到 $\alpha = 1.28$、$H = 0.55$、$c_1 = 211.58$、$c_2 = 410.17$, 根据参数值生成 LFSN 模型仿真序列如图 4 所示, 然后使用 Pareto 模型和 FBM 模型生成仿真序列如图 5 和图 6 所示. 从图中可以看出 LFSN 模型序列更加接近实际流量, Pareto 模型序列的大部分较小值都比实际流量的较小值大, FBM 模型序列在一个较小的范围内波动, 与实际流量差异较大; 在流量突发性方面, 相比于 Pareto 模型序列的突发值, LFSN 模型序列的突发值更接近于实际流量的突发值, 并且 LFSN 模型序列的突发值个数也明显多于 Pareto 模型序列的突发值个数, 而 FBM 模型序列完全不具有突发值.

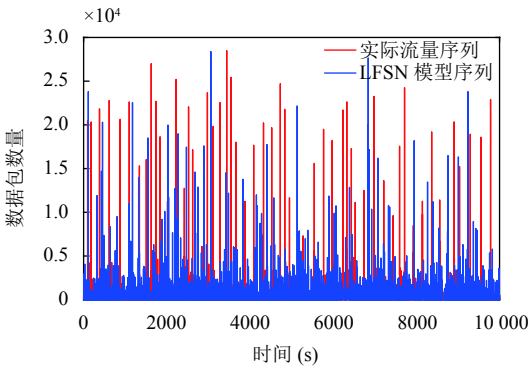


图 4 单位时间 1 s 内微信实际流量与 LFSN 模型仿真序列

接下来对 LFSN 模型序列、Pareto 模型序列和 FBM 模型序列的自相似参数和突发参数进行估算, LFSN 模型序列的自相似参数 $H = 0.53$、突发参数 $\alpha = 1.40$, Pareto 模型序列的自相似参数 $H = 0.47$、突发参数 $\alpha = 1.19$, FBM 模型序列的自相似参数 $H = 0.51$、突发参数 $\alpha = 2.00$. 在自相似性方面 LFSN 模型序列更接近于实际流量的自相似性, 尽管 FBM 模型是严格的自相似流量模型, 但是在保持微信流量的自相似性上仍然比 LFSN 模型差一些; 而在突发性方面, 由于 LFSN 模型序列的突发值个数比实际流量的突发值个数少, 所以突发性变小 (α越大突发性越小), 而 Pareto 模型序列由于突发值个数非常少, 并且个别突发值比大多数序列值大许多, 反而凸显了其突发性变强的特点, 但根据图 5 可以看出 Pareto 模型序列的突发值个数相比于实际流量突发值个数少许多, 而 FBM 模型则完全不能刻画微信流量的突发性. LFSN 模型序列的突发值个数比 Pareto 模型多的主要原因是 LFSN 模型中的自相似参数对突发参数的作用, 保持了一段时间内流量突发性的持续, 而 Pareto 模型中并没有自相似参数作用于突发参数. 此外本文还计算了 LFSN 模型序列、Pareto 模型序列和 FBM 模型序列与微信实际流量序列的拟合优度 R^2 值, 分别为 0.75、0.67、0.32. 综上 LFSN 模型能比 Pareto 模型和 FBM 模型更好的刻画微信流量的突发性和自相似性.

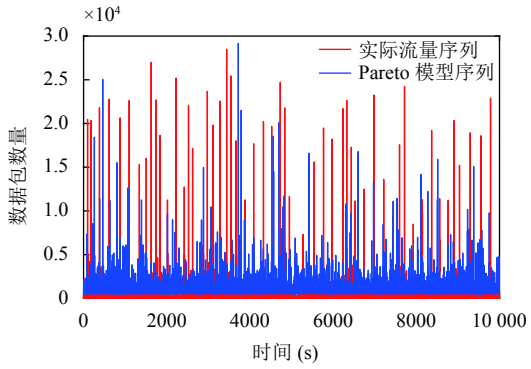


图 5 单位时间 1 s 内微信实际流量与 Pareto 模型仿真序列

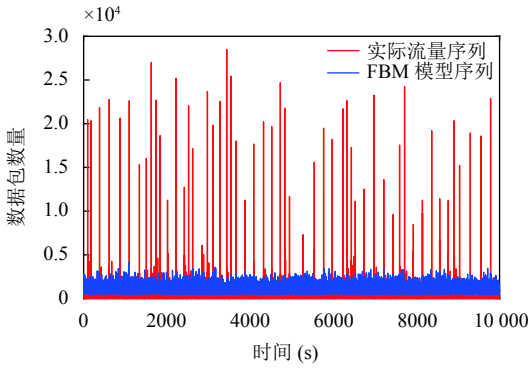


图 6 单位时间 1 s 内微信实际流量与 FBM 模型仿真序列

© 中国科学院软件研究所 http://www.c-s-a.org.cn

LFSN 模型和 FBM 模型都可以描述流量的自相似性, 所以本文为了分析微信实际流量与 LFSN 模型序列和 FBM 模型序列在自相似长相关性方面的变化趋势, 采用归一化样本自相关函数 (NACF) 进行比较[9], NACF 的表达式为:

$$\rho(k)=\frac{\sum_{i=1}^{n-k}X(i)X(i+k)}{\sum_{i=1}^{n}X^2(i)},1<k\le n-1 \tag{11}$$

微信实际流量和 LFSN 模型序列的 NACF 如图 7 所示, LFSN 模型序列的 NACF 衰减速率很慢, 并且近似于实际流量的 NACF 变化趋势, 说明二者在自相似长相关性的变化非常近似, LFSN 模型可以保持微信流量的自相似长相关变化趋势. FBM 模型序列的 NACF 如图 8 所示, FBM 模型序列的 NACF 衰减速率较快, 并且与微信实际流量的 NACF 变化趋势差异较大, 主要原因是微信实际流量具有较大的突发性, 根据流量自相似性成因可知, 具有突发性的流量叠加也会促进流量表现出自相似长相关性. 综上判断 LFSN 模型可以对微信流量建模, 在建模基础上可以对微信流量进行时延、丢包率等性能分析, 为网络流量监管提供依据.

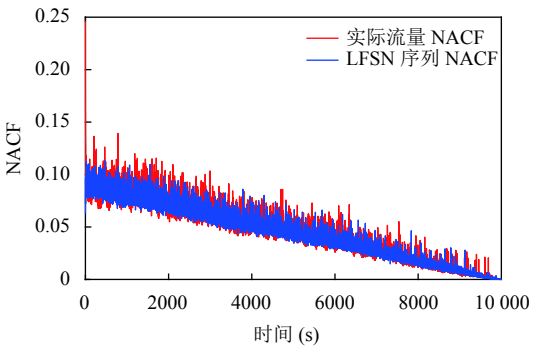


图 7 微信实际流量与 LFSN 序列 NACF

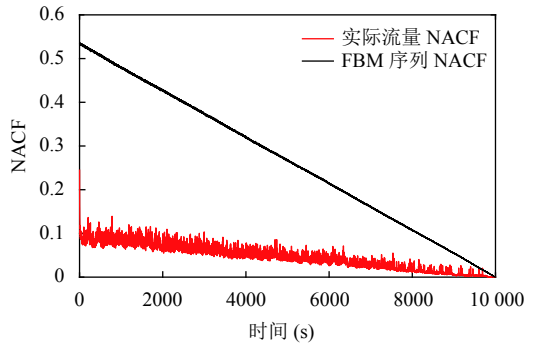


图 8 微信实际流量与 FBM 序列 NACF

4.2 微信流量网络性能指标估算

考虑在一般到达过程和确定服务速率的先来先服务的单个服务器队列情况下 (G/D/1), 基于 LFSN 模型给出大缓冲区条件下缓冲区溢出概率的渐进计算公式如下[9]:

$$\lim_{b\to\infty}P(Q(t)>b)\ge K(\alpha,H,c_1,c_2,c)b^{-\alpha(1-H)} \tag{12}$$

其中, $Q(t)$表示时间t时队列中的数据包数量, b是缓冲区长度, c是固定服务速率, $K(\alpha,H,c_1,c_2,c)$是由 5 个参数决定的常量, 其表达式如式 (13):

$$K=\frac{(1-\alpha)(\sigma_1c_1(1-H))^{\alpha}}{\Gamma(2-\alpha)\cos\left(\frac{\pi\alpha}{2}\right)}\left(\frac{(c-c_2)(1-H)}{H}\right)^{-\alpha H} \tag{13}$$

$$\sigma_1=\left[\int_0^{+\infty}\left((1+x)^{H-\frac{1}{\alpha}}-x^{H-\frac{1}{\alpha}}\right)^{\alpha}dx+\int_0^1(1-x)^{\alpha H-1}dx\right]^{\alpha} \tag{14}$$

令$K=-\alpha(1-H)$, 根据式 (13) 推导出平均队列长度、平均时延、丢包率等网络性能指标表达式如下[13]:

$$\text{丢包率: }\varepsilon=K(\alpha,H,c_1,c_2,c)b^{-\alpha(1-H)} \tag{15}$$

$$\text{平均队列长度: }E(b)=-\frac{C_\alpha K}{K+1}b^{K+2} \tag{16}$$

$$\text{平均时延: }E(T_d)=E(b)/c \tag{17}$$

根据 4.1 节估算的微信流量在 LFSN 模型下的 4 个参数值和固定服务速率c可以计算C_α的值, 由于式 (12) 要求固定服务速率必须大于流量的平均值, 那么假设 $c=600$, 从而$C_\alpha=0.5188$. 以丢包率和平均时延为例, 微信流量在不同缓冲区长度下的平均时延和丢包率 (P) 如图 9 和图 10 所示. 实际服务器设置的服务速率和缓冲区长度未知, 图 9 和图 10 的估算结果只是为了展示使用 LFSN 模型对微信流量建模可以估算其丢包率和平均时延, 从而可以调整服务器的服务速率和缓冲区长度以控制微信流量的丢包率和平均时延, 为微信流量监管提供依据, 而微信流量占据了互联网中的大部分流量, 所以也为网络流量监管提供参考.

5 结语

本文以微信流量为研究对象, 首先直观观察微信流量的自相似性和突发性, 然后计算 H 参数验证其具有自相似性, 通过 α 稳定分布验证描述微信流量的突发性. 在此基础上使用 LFSN 模型对微信流量建模, 并

© 中国科学院软件研究所 http://www.c-s-a.org.cn

且使用 Pareto 模型和 FBM 模型进行模型效果对比, 证明 LFSN 模型能更好地刻画微信流量的突发性和自相似性.

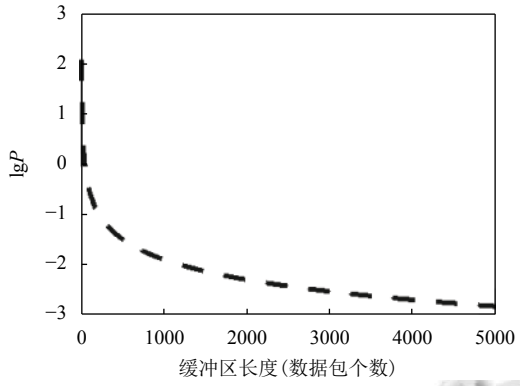


图 9 微信流量丢包率估算

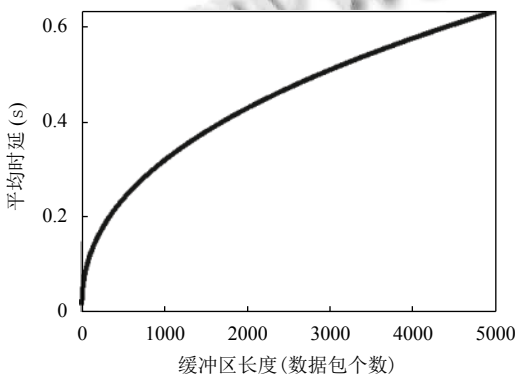


图 10 微信流量平均时延估算

参考文献

1 Leland WE, Taqqu MS, Willinger W, *et al*. On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM Transactions on Networking, 1994, 2(1): 1–15. [doi: 10.1109/90.282603]

2 Trustdata. 2020 年 Q1 中国移动互联网行业分析报告. http://www.itrustdata.com, [2020-05-22].

3 李玮. 微信流量模型与业务识别方法研究 [硕士学位论文]. 北京: 北京理工大学, 2015.

4 燕飞鹏. 基于网络流量的微信用户行为识别技术 [硕士学位论文]. 杭州: 杭州电子科技大学, 2019.

5 张江楠, 谭献海, 王帅. 微信数据流量特性的全面分析. 单片机与嵌入式系统应用, 2019, 19(7): 6–9, 14.

6 Cox DR. Long-range dependence: A review. Proceedings of 50th Anniversary Conference. Iowa: The Iowa State University Press, 1984. 55–74.

7 万贝利. 移动网络流量特性分析及预测研究 [硕士学位论文]. 重庆: 重庆大学, 2016.

8 单志明. α 稳定分布参数估计及自适应滤波算法研究 [博士学位论文]. 哈尔滨: 哈尔滨工程大学, 2012.

9 Karasaridis A, Hatzinakos D. Network heavy traffic modeling using α-stable self-similar processes. IEEE Transactions on Communications, 2001, 49(7): 1203–1214. [doi: 10.1109/26.935161]

10 McCulloch JH. Simple consistent estimators of stable distribution parameters. Communications in Statistics-Simulation and Computation, 1986, 15(4): 1109–1136. [doi: 10.1080/03610918608812563]

11 Chambers JM, Mallows CL, Stuck BW. A method for simulating stable random variables. Journal of the American Statistical Association, 1976, 71(354): 340–344. [doi: 10.1080/01621459.1976.10480344]

12 王晖, 季振洲, 朱素霞. 自相似网络流量模型研究. 智能计算机与应用, 2013, 3(2): 34–41. [doi: 10.3969/j.issn.2095-2163.2013.02.008]

13 Tan XH, Hu Y, Jin WD. Modeling and performance analysis of LFSN. 2007 IFIP International Conference on Network and Parallel Computing Workshops (NPC 2007). Dalian: IEEE, 2007. 549–554.

© 中国科学院软件研究所 http://www.c-s-a.org.cn