IET Communications

Review Article

IET Journals
The Institution of Engineering and Technology

Wireless big data in cellular networks: the cornerstone of smart cities

ISSN 1751-8628
Received on 20th November 2017
Revised 2nd February 2018
Accepted on 30th March 2018
E-First on 29th June 2018
doi: 10.1049/iet-com.2017.1278
www.ietdl.org

Rongpeng Li[1] ✉, Zhifeng Zhao[1], Chenyang Yang[2], Chunming Wu[1], Honggang Zhang[1]

[1]*Zhejiang University, Hangzhou 310027, People's Republic of China*
[2]*Beihang University, Beijing 100191, People's Republic of China*
✉ *E-mail: lirongpeng@zju.edu.cn*

Abstract: The rapid urbanisation has transformed cities to the preferential human settlement and allowed cities to quietly witness all range of human activities. As the key enabler in the information and communications technology industry, cellular networks play a decisive role in delivering communication messages and entertainment content. In particular, cellular network operators respond to human initiated service requests by gradually deploying necessary infrastructure and calibrating transmission protocols. Hence, cellular network records encompass the interesting interaction between human-initiated messages and network-triggered responses. In this study, the authors collect the 'big data' in urban cellular networks and try to dig out the human and urban planning properties. Specifically, they focus on the statistical modelling of three representative scenarios like spatial deployment density of base stations, packet length or traffic volume of mobile services, as well as inter-arrival time and dwell time of human mobility. Through extensive data mining, they validate the heavy-tailed feature universally existing in these scenarios. Afterwards, they discuss the implications of this heavy-tailed feature and talk about its fundamental contribution to intelligent resource adjustment, proactive content caching, and enhanced connection management in cellular networks. Finally, they highlight the applications of this feature towards smarter cellular networks and cities.

1 Introduction

With the rapid urbanisation of developing countries, cities have become the preferential habitat and more than 54% of the world's population is believed to live in cities [1]. Meanwhile, as a large and permanent human settlement, each city has quietly witnessed all range of human activities in daily life and recorded them in various types of 'big data'. For example, in Hangzhou, an eastern provincial capital of China with a population of over 9 million, more than 17 million subscribers connected the Internet by cellular networks, more than 2.3 million cars ran in the city and nearly 24 billion kilowatt-hour electrical power were consumed in 2016. All these societal and industrial processes produce or capture a significantly large number of logs. Despite many common features in the urban 'big data' like large volume, variety, value, and velocity, data in cellular networks is unique. First, cellular networks have millions of subscribers and hence gathered abundant patterns of human activities. Second, in order to transmit the human-initiated content, the existing protocols in cellular networks have to truncate some long messages. Hence, cellular network records implicitly encompass the interaction between network protocol designers and users. Third, the infrastructure in cellular networks indicates how networks evolve to respond to human demands. Therefore, it is meaningful to investigate the 'big data' of cellular networks and leverage the embedded properties to revolutionise the design of smarter cities.

On the other hand, in order to realise convenient network operation and maintenance, network operators collect anonymous cellular detail records (CDRs) to document the location of the base stations (BSs) carrying every voice call and data exchange, as well as the time stamp when the event occurred [2]. Implied by the well-known Shannon entropy theory, these CDRs in cellular networks explicitly or implicitly accumulate a large number of human activities and provide one effective means to further reduce the information uncertainty. However, given the explosive content volume and generating velocity, it is usually computational intensive to directly apply the 'big data'. Hence, it becomes more viable to extract some statistical models beforehand. Interestingly, based on the 'preferential attachment' [3], many large or complex networks should be heavy tailed. As its names implies, a heavy-tailed distribution has not-exponentially bounded tail. Mathematically, for a heavy-tailed variable X, the probability $\Pr(X > x)$ will satisfy $\lim_{x\to\infty} e^{\zeta x}\Pr(X > x) = \infty$, for all $\zeta > 0$. Many well-known statistical distributions including power-law distribution (also named as generalised Pareto distribution), Weibull distribution, log-normal distribution, and α-stable distribution [4] are proven to belong to the heavy-tailed family.

In this paper, we first investigate the correctness of the aforementioned bold argument and focus on the statistical modelling of three interesting cases like spatial deployment density of BSs, packet length or traffic volume of mobile services, as well as inter-arrival time and dwell time of human mobility. Later, inspired by the fact that human activities share some similarities between cellular networks and other typical networks (e.g. transport networks) in cities, we talk about the potential application of this heavy-tailed feature in cellular networks and bridge the gap between cellular networks and transport networks, so as to achieve smarter cities in a more harmonised manner.

2 How can we learn from the CDRs in cellular networks

In the past few years, we have collected a significant large number of CDRs in Hangzhou and Ningbo, Zhejiang Province from China Mobile, spanning from the longitude and latitude of the deployed BSs to traffic information with timestamps, subscriber IDs, and serving BSs. Contingent on these CDRs, we could conveniently learn some essential elements like *who* [user equipment (UE), *where* (locations of BSs), *when* (timestamp), *what* (content of messages). More importantly, we are able to extract statistical models in both spatial and temporal dimensions:

- *Spatial dimension*: The knowledge in spatial dimension includes the spatial correlation of human-deployed network resources or human-initiated resource requests. As shown in Figs. 1*a* and *b*, we can map the really deployed BSs with the geographical environment and intuitively examine the distribution of BSs. Besides, like Fig. 1*c*, we can plot the spatial traffic density

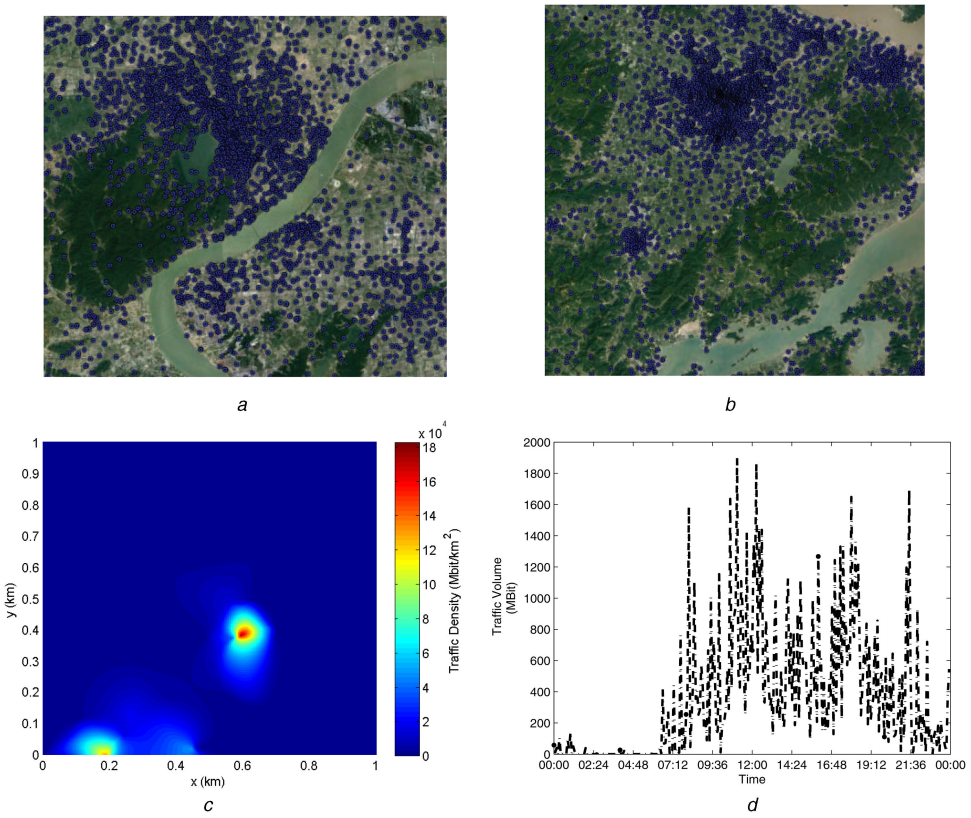

Fig. 1 *Snapshot for the realistic cellular network data*

(a, b) Geometric deployment of BSs in Hangzhou and Ningbo, Zhejiang Province, China, where each blue dot indicates a BS, ***(c)*** The distribution of spatial traffic density, where darker red regions imply heavier traffic loads, ***(d)*** The temporal traffic dynamics of one randomly selected BS

according to the location and volume of the related BSs and try to understand the related spatial relevancy in traffic.

- *Temporal dimension*: Similarly, the knowledge in temporal dimension means temporal characteristics embedded in human-initiated resource requests. Usually, for a particular user or BS, we can augment the traffic volume of specific services at different moments to be a traffic vector. Fig. 1*d* illustrates the variations of the traffic vector with respect to the time. Based on the traffic series, we can take advantage of common signal processing techniques to check out the embedded properties like periodicity, and leverage them to forecast future traffic loads.

Usually, the statistical modelling means to evaluate the fitting accuracy and find the most accurate statistical distribution from candidates. However, it is challenging to find out the exact distribution of a dataset, especially for some distribution lacking a closed-form expression for its probability density function (PDF). Therefore, an alternative means is to examine whether a dataset is consistent with some hypothetical distribution, so as to confirm the appropriate distribution of the dataset. Specifically, it is necessary to first estimate the unknown parameters from a given dataset, and then check the fitting error between the real distribution of the dataset and the estimated one. Notably, the unknown parameters in various candidate distributions (except α-stable distribution) could be estimated by the maximum likelihood estimation methodology. However, for α-stable distribution, due to the lack of its PDF in closed form, the relevant parameters should be obtained by quantile methods [4].

3 What have we learned about the statistical modelling results in cellular networks?

3.1 Spatial deployment density of BSs

The spatial deployment distribution of BSs lays the foundation for the performance analyses of emerging technologies in cellular networks. Recently, there have already existed substantial works towards discovering the spatial density distribution of BSs in cellular networks from various practical measurements. In the earliest stages, a two-dimensional hexagonal grid model was used and implied that BSs were spatially uniformly deployed, which is obviously deviated from the real scenarios. In next stages, Poisson distribution (i.e. PPP, Poisson point process) [5] was assumed to be able to roughly match the realistic BS deployment in cellular networks, meanwhile providing tractable analysis results. However, the modelling accuracy of Poisson distribution has always been questioned [6] and the clustering or repulsive feature among deployed BSs has been considered by introducing some variants of PPP to obtain precise analysis results [7]. On the other hand, the actual deployment of BSs during the past long evolution is highly correlated with human activities [8]. Humans tend to live together with societal relationship, and their social behaviours would lead to mobile traffic hotspots, thus causing BSs to be more densely deployed in certain areas as clusters. Therefore, heavy-tailed distributions appear to be more suitable to precisely characterise the clusteringly distributed BSs.

In this part, we collect the location information of the second/third/forth generation (i.e. 2G/3G/4G) BSs deployed by China Mobile in Hangzhou and Ningbo, which cover 8826 and 5746 BSs, respectively. Based on the large amount of BS location data, we spatially sample each city randomly with a fixed sample area size. Then, we compute the spatial density for 10,000 different sample areas and obtain the empirical density distribution, by counting and sorting the number of BSs in each sample area.

In the first place, we consider Hangzhou as a typical example and compute the PDF under the sample area size of $4 \times 4\,\text{km}^2$. Next, we model the corresponding PDF by various typical distributions and obtain the parameters in these disributions in Table 1. We provide the empirical BS density distribution with the candidate ones in Fig. 2*a* and summarise the numerial fitting error in Table 1, in terms of root mean square error (RMSE). As depicted in Fig. 2*a*, the statistical pattern of BSs obviously exhibits heavy-tailed characteristics. Besides, among all candidate distributions, α-stable distribution most precisely matches the empirical PDF. On the other hand, the RMSE results in Table 1 show α-stable distribution has the minimum RMSE (0.0177) while Poisson

Table 1 RMSE values after fitting candidate distributions to empirical data

Scenario	Case	α-Stable	Poisson	Exponential	Log-normal	Power-law	Weibull
BS deployment	Hangzhou 3×3 km^2	**0.0105**	0.1214	—	0.0207	0.0274	0.0361
	Hangzhou 4×4 km^2	**0.0177**	0.1465	—	0.0269	0.0339	0.0418
	Hangzhou 5×5 km^2	**0.0286**	0.1702	—	0.0293	0.0357	0.0432
	Ningbo 4×4 km^2	**0.0279**	0.2537	—	0.0905	0.1017	0.1151
service modelling	packet length	790×10^{-5}	—	56.0×10^{-5}	34.0×10^{-5}	9.76×10^{-5}	65.8×10^{-5}
	aggregated traffic	**0.0144**	—	0.0899	0.0491	0.0357	0.0470
human mobility	inter-arrival time	0.0266	—	0.0200	0.0062	**0.0019**	0.0169
	dwell time	0.0062	—	0.0019	5.76×10^{-4}	3.05×10^{-4}	9.29×10^{-4}

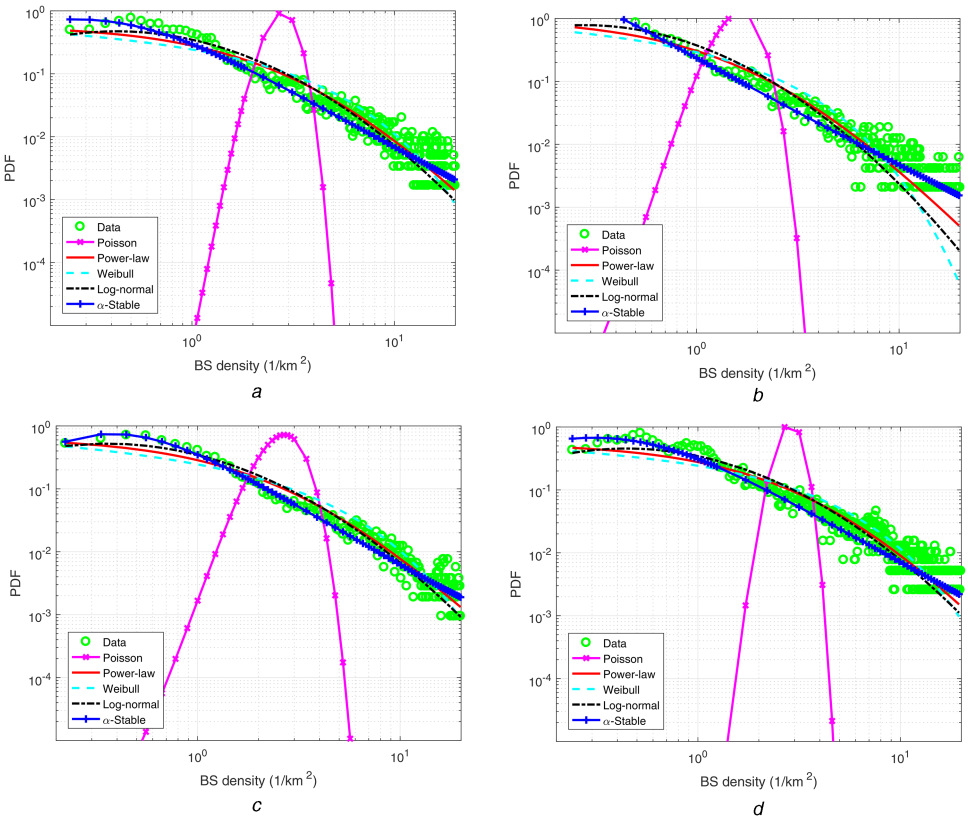

Fig. 2 *Results after fitting BS density distribution in Hangzhou and Ningbo to candidate distributions, when the size of sample area varies*

(a) Hangzhou, 4×4 km^2, ***(b)*** Ningbo, 4×4 km^2, ***(c)*** Hangzhou, 3×3 km^2, ***(d)*** Hangzhou, 5×5 km^2 Key insight: The spatial pattern of deployed BSs exhibits strong heavy-tailed characteristics, and α-stable distribution manifests itself as the most precise one. On the contrary, the popular Poisson distribution is an inappropriate model for the actual BS density distribution, in terms of the RMSE

distribution has the maximum one (0.1465), and once again strengthen this aforementioned conclusion. Meanwhile, we also change the sample area sizes to 3×3 and 5×5 km^2, respectively, so as to reach a more general result. The related results are provided in Figs. 2*c* and *d*. Obviously, compared with other candidate distributions, α-stable distribution still provides the most accurate fitting results for the BS density distribution in Hangzhou. Fig. 2*b* and Table 1 also indicate that similar observation could also be found for the BS deployment in Ningbo.

3.2 Packet length or traffic volume of services

Cellular networks have become the digital pipes for diverse mobile services like web browsing, video, instantaneous messaging (IM) services, and so on. Notably, our previous work [9, 10] has validated the accuracy to model the aggregated traffic for web browsing and video services by α-stable distribution. Therefore, we focus on the IM service here. IM service, which has built its reputation from the personal computer era by supporting real-time communications, and recently flourishes in mobile devices around the world and quickly generated significant amount of traffic loads within cellular networks. In this part, we use the mobile IM (MIM) service as an example to demonstrate the statistical modelling results for both individual packet length and aggregate traffic volume at a BS side. In traditional fixed broadband networks, researchers have shown that aggregate traffic traces exhibit strong burstiness and follow α-stable distributions [11]. On the other hand, the investigation over traffic characteristics of IM in wired Internet reveals heavy-tailed distribution phenomena in services like Skype [12]. Therefore, it is natural to raise a question, namely which one of the aforementioned statistical models is more suitable for the MIM traffic? As depicted in Fig. 3*a*, the MIM services distinguish themselves with the inborn packet-switching nature and its accompanied keep-alive mechanisms between UEs and cellular networks, and consume only a small amount of core network bandwidth but considerable radio resources of mobile access networks, thus making MIM services special [9, 10]. So, here comes the question again, that is, do we need a totally different

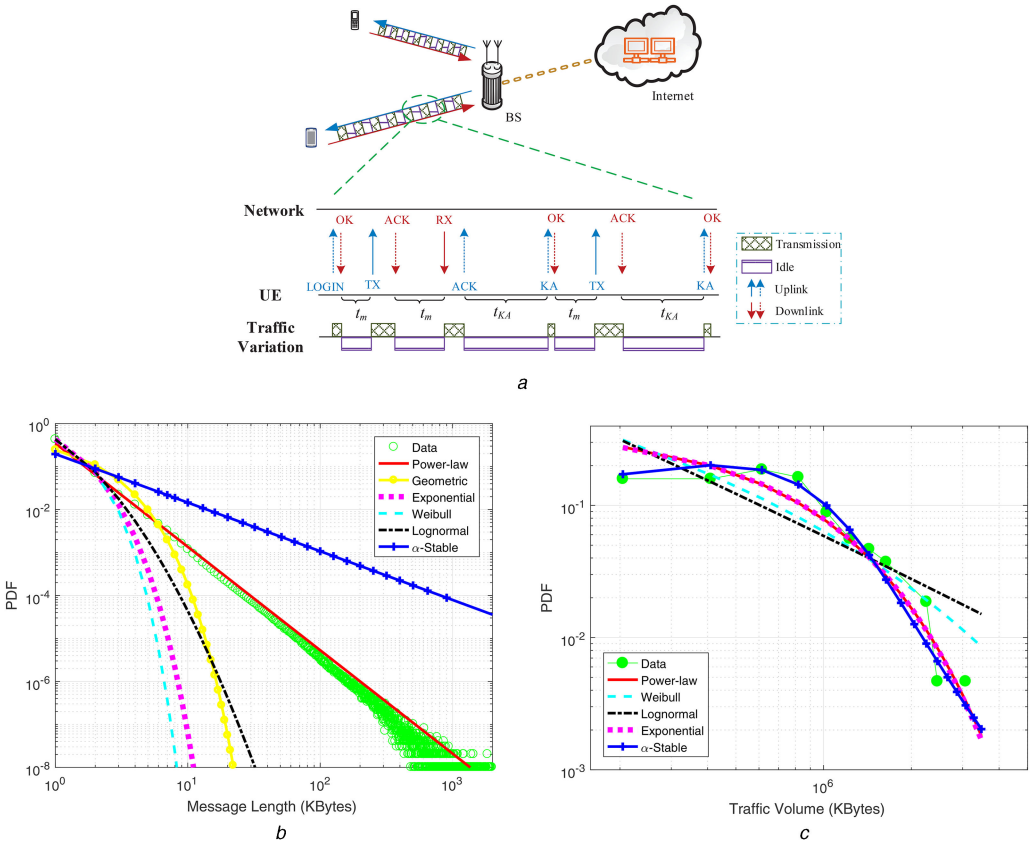

Fig. 3 *Working mechanism of MIM service like 'Wechat/Weixin'; statistical modelling results of the individual packet length and aggregated traffic at a BS side*

(a) Working mechanism of MIM service, ***(b)*** Individual packet length, ***(c)*** Aggregated traffic

Key insight: Different from the recommendation from 3GPP, packet length better follows power-law distribution. For aggregated traffic within one BS, the accuracy of applying α-stable distribution is proven to accurately characterise this statistical pattern, thus extending the suitability of α-stable distribution for traffic in both fixed core networks and cellular access networks

traffic model? We try to answer these questions by leveraging the CDRs of the widely booming MIM service 'WeChat/Weixin', which allows over 9 hundred million mobile users in China as well as around the world and transmits over over 38 billion text messages and 6.1 billion voice messages daily. Here, for the packet-level analyses, we collect the one-month 2G/3G/4G 'WeChat/Weixin' logs in April 2014, which contain timestamps, cell IDs, anonymous subscriber IDs, message lengths, and message types. For the aggregated traffic in a BS, we collect per 5 min 'WeChat/Weixin' traffic volume of roughly 6000 BSs in Hangzhou in September 2014.

We fit candidate distribution functions to the empirical PDF of message length and provide the corresponding results in Fig. 3*b*. Notably, this empirical PDF takes account of the 1-month traffic records collected from 7 million subscribers, originated from about 15,000 BSs of China Mobile within a region of 3000 km^2, by sorting the size of all incorporated packets and calculating the number of packets with the same specific packet size. From the figure, it can be observed that instead of the recommended geometric distribution by 3GPP, power-law distribution (i.e. $0.347x^{-2.407}$) could better approximate the empirical PDF of message length. The RMSE results in Table 1 also confirm that modelling the PDF of message length by power-law distribution yields superior accuracy.

Next, we evaluate of the appropriate PDF for the aggregated traffic from multiple UEs to one BS. We adopt the methodology same as that for individual packets. Fig. 3*c* presents the corresponding PDF comparison between the estimated results and the real aggregated traffic in one randomly selected BS. It can be observed that the traffic records in these selected areas could be better simulated by α-stable distribution. Similarly, Table 1 also indicates that α-stable distribution leads to better fitting accuracy in terms of RMSE. Hence, consistent with previous findings in fixed broadband networks [11], α-stable distribution is verified to accurately model the aggregated traffic from cellular access networks to core networks.

The reasons that aggregated traffic universally obeys α-stable distribution can be simply explained as follows: from Fig. 3*b*, the length of one individual message follows a power-law distribution. Meanwhile, the aggregated traffic within one BS can be regarded as the accumulation of messages from diverse UEs. The frequent MIM packet transmission implies that the aggregated traffic equals a large number of packets. Therefore, according to the generalised central limit theorem [4], the sum of a number of random variables with power-law distributions like $\propto x^{-\alpha-1}$ where $0 < \alpha < 2$ (and therefore having infinite variance) will tend to be an α-stable distributed one as the number of summations grows. On the other hand, the fitting values of α in different cells mostly fall between 1.136 and 1.515, while the slope of power-law distribution in Fig. 3*b* is 2.407. These findings are consistent with the theory from the generalised central limit theorem.

3.3 Inter-arrival time and dwell time of human mobility

Human mobility patterns contribute a lot to the protocol design and performance analysis of cellular networks. During the past 10 years, when users are on the move, the handoff performance of cellular networks have been designed and analysed based on the classical assumption that inter-arrival time and dwell time are either exponentially distributed or other similar distributions [13]. Nevertheless, as discussed earlier, several power-law distribution phenomena have been discovered in various areas of human behaviours lately, such as the inter-arrival time between consecutive IM packets sent by user obey power-law [3, 14]. Rhee *et al.* [15] analyses global positioning system traces and reports that travel length and pause time of human follow truncated power-law distributions. These findings above inspire us to check out that the inter-arrival time and dwell time in cellular networks might

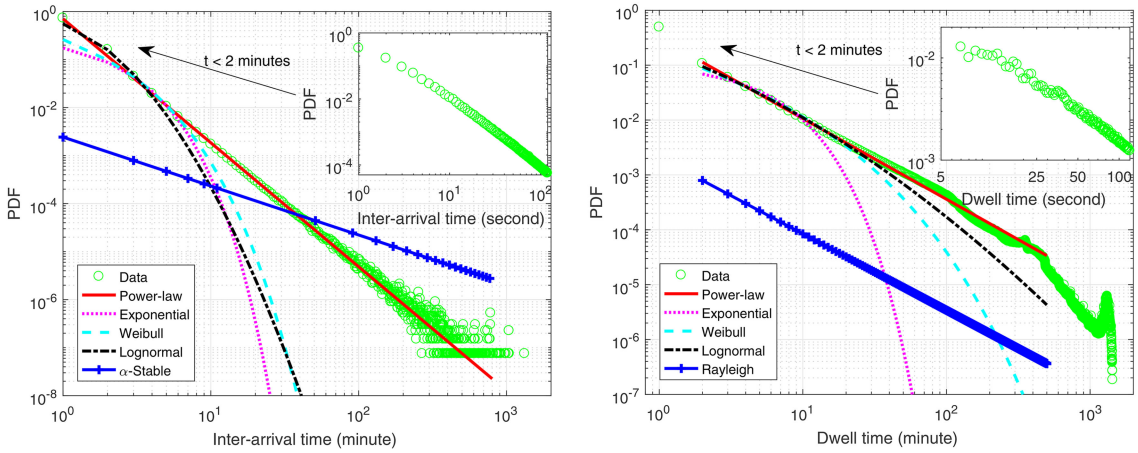

Fig. 4 *Fitting results of typical distributions to the measurement data of*
(a) Inter-arrival time, ***(b)*** Dwell time
Key insight: The inter-arrival time and dwell time could be more accurately modelled by the power-law distribution

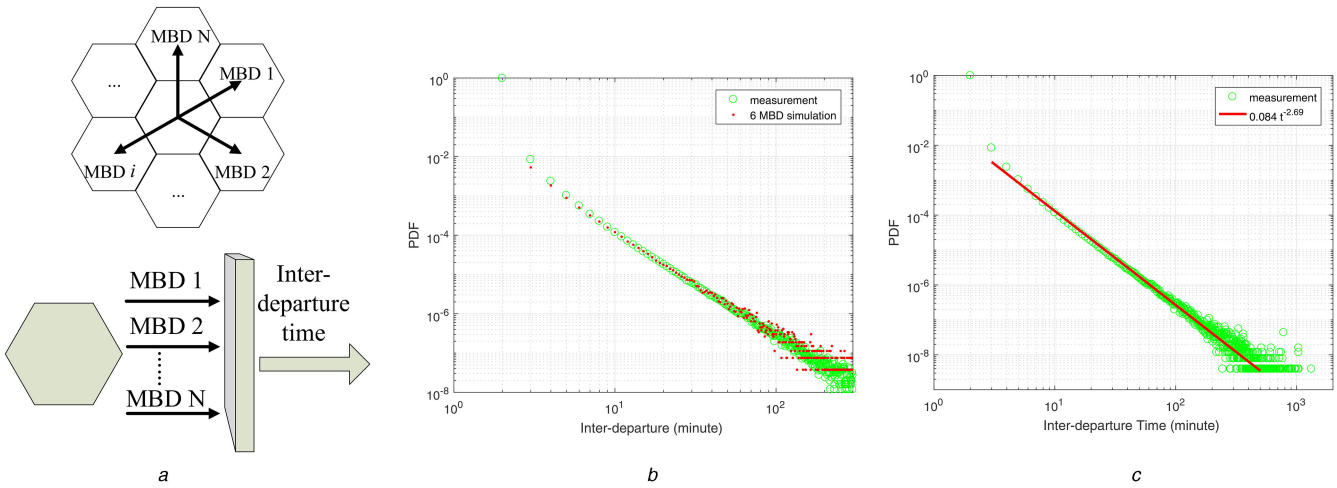

Fig. 5 *An illustration of human mobility modeling results*
(a) The illustration of memory birth-death process (MBD), ***(b, c)*** The fitting accuracy of MBD process and power-law distribution to the measurement data of inter-departure

follow heavy-tailed distribution rather than exponential distribution.

In this part, we collect one-month mobility-related information in 2012, which covers over 15,000 2G/3G BSs. Again, we extract the inter-arrival time and dwell time from the collected timestamps that every user enters and leaves one cell and further calculate the empirical PDF. Fig. 4*a* depicts the results of fitting typical distributions to the empirical PDF and Table 1 summarises the fitting error in terms of RMSE. Both of them indicate that the empirical inter-arrival time could be better modelled by a power distribution (i.e. $0.706x^{-2.58}$). Similarly, Fig. 4*b* implies that for the dwell-time, the power-law distribution (i.e. $0.3137x^{-1.47}$) again outperforms the counterpart ones.

As both the inter-arrival time and dwell time adhere to the power-law distribution and the previous study [8] unveils that the dwell time is independent of both inter-arrival time and the mobility trend of UEs, the arrival-departure relationship for a cell with N neighbouring BSs can be modelled as N MBD processes in Fig. 5*a*. According to the aforementioned fitting results, the PDFs of inter-birth time and life time should be modelled by the power-law distribution instead of the exponential distribution in classical memory-less birth-death processes. Moreover, the inter-death time of N MBD processes can be used to explicitly describe the inter-departure time of the cell. In Fig. 5*b*, we compare the simulation and measurement distributions where the simulation results are given by first generating power-law distributed inter-arrival time and dwell time of 6-MBD process and then computing the inter-departure time therein. From Fig. 5*b*, 6-MBD process could well capture the measurement results. Moreover, Fig. 5*c* depicts the fitting result and shows the inter-departure time is still power-law distributed.

4 What does the heavy-tailed cellular networks imply?

The statistical modelling of representative scenarios in cellular networks clearly exemplifies the widely existing heavy-tailed distributions, which brings tremendous opportunities to build smarter cellular networks:

- *Intelligent resource adjustment*: The BSs are intended to provision the high-peak traffic demands. The heavy-tailed deployments of BSs mean some BSs are clustered together. Therefore, when the predicted traffic loads is doomed to be low or a number of mobile subscribers are expected to leave, some BSs in the clusters could be turned into sleeping mode while neighbouring BSs are able to satisfy the traffic demands and avoid potential coverage holes by correspondingly adjusting the azimuth angle of antennas.
- *Proactive content caching:* Recently, it has been reported that proactive caching could enhance the network performance, such as relieving network congestion by offloading or multi-casting and reducing deployment and operational expenditure by replacing backhaul with cache. However, the biggest concern is that the additional cost of storage resource may overwhelm the potential gain. Fortunately, the heavy-tailed feature of cellular networks shows that the popular content is rather limited, thus indicating that proactive content caching could make sense.

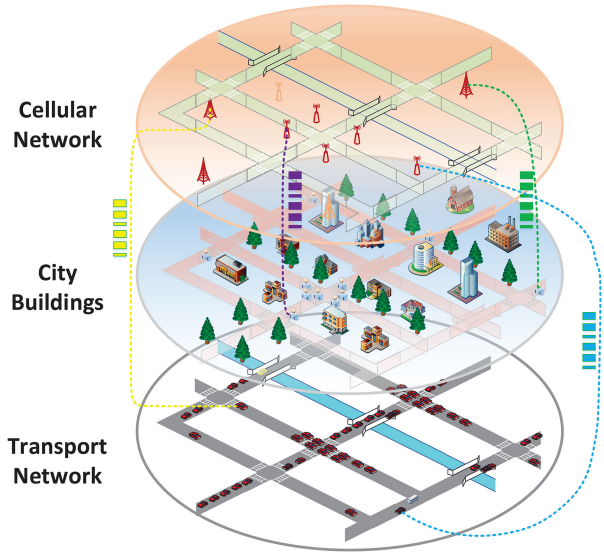

Fig. 6 *Illustration of the interaction between cellular networks and transport networks in smarter cities, where cellular networks act as the anchor delivering essential packets with self-similar sizes*

- *Enhanced connection management:* How to guarantee the service quality of UEs with highly mobility remains a critical challenge. However, the problem could be partially solved by dual connectivity. In particular, when the trajectory of UEs is well forecasted, a connection between the UE and the BS at the moving direction of this UE could be established in advance.
- *Smart network slicing:* The fifth-generation (5G) cellular networks aim to provision services with significantly distinct requirements (e.g. high throughput, low latency etc). One potential solution is to slice or reserve some resources for one type of service with the same or similar requirement. In order to improve the utilisation of the reserved resource, predicting the volume or bandwidth is of vital importance. In this regard, similar to the intelligent resource adjustment, the heavy-tailed feature will make a difference.

However, it is natural to raise a question on how to leverage the heavy-tailed feature and build smarter cellular networks? Beforehand, it is necessary to address several well-known features, which are coupled with the heavy-tailed distribution (especially α-stable distribution) [16]:

- *Self-similarity*: Given zero-mean, stationary time series $Y = (Y_t, t = 1, 2, 3, \ldots)$, Y is called as self-similar process if its m-aggregated series $Y^{(m)} = (Y_t^{(m)}, t = 1, 2, 3, \ldots)$ with each element $Y_t^{(m)} = \sum_{k=(t-1)m+1}^{tm} Y_k$ satisfy $Y_t^{(m)} = m^H Y_t$ for all $m > 0$, where $H \in (0, 1)$ is the Hurst parameter indicating the decay rate of statistical dependence of two points with increasing time interval or spatial distance between these two points. As $Y^{(m)}$ has the same distribution as Y rescaled by m^H, the self-similarity implies that the cellular network traffic is exactly or approximately similar to a part of itself. In other words, the traffic variations in different temporal resolution (time scaling) have similar shape.
- *Long-range dependence (LRD)*: The LRD property is closely related to self-similarity. It represents that the auto-correlation function $r(k)$ of the self-similar time series Y with the variance σ^2 follows the equation $r(k) = E(Y_t Y_{t+k})/\sigma^2 \sim k^{-\beta}$ where $\beta \in (0, 1)$. Moreover, $H = 1 - \frac{\beta}{2}$. Hence, if the series are long-range dependent, $H \in (\frac{1}{2}, 1)$, indicating the time series have persistent long memory effects. Therefore, the LRD in traffic variations allows sufficient time to predict human mobility and traffic loads and recommend necessary network changes.
- *Spatial clustering*: The spatial clustering represents the spatial inhomogeneity of human activities and shows the required network infrastructure (e.g. BSs) sometimes has to be deployed in some clusters.

Benefiting from these implications, predicting the heavy-tailed traffic or mobility trajectory becomes practical. One of the typical examples is to predict the self-similar traffic series [17] by long short-term memory (LSTM), one kind of deep learning algorithms [18]. The expression 'long short-term' refers to the fact that LSTM is a model for the short-term memory which can last for a long period of time. Therefore, LSTM could well match the LRD induced by the heavy-tailed feature. In [18], the authors take advantage of LSTM to forecast the cellular network traffic and show remarkable prediction accuracy over the classical auto-regressive integrated moving average and support vector regression. Specifically, [18] demonstrates that the mean square error, mean absolute error and log loss (also known as cross entropy) are as low as 0.042, 0.165, and 0.583, respectively, for downlink traffic load; while they become 0.031, 0.137, 0.556, respectively, for the uplink. In other words, based on the prediction results, the intelligent resource adjustment could be accurate and effective. Also, considering the flexibility to model traffic or mobility trajectory by time series, LSTM and the heavy-tailed feature behind provides interesting insight and huge imagination space on how to build smarter cellular networks.

In fact, cellular networks and other typical networks (e.g. transport networks) in cities are mutually dependent [19]. Changes in individual behaviours lead to altered travel, activity and calling patterns, which influence the loads on both transport network and cellular networks. For example, when an event (e.g. a concert) happens, people are prone to more frequently phone their friends, thus bringing the increase of loads in cellular networks. Meanwhile, more cars will jam into the neighbouring streets of the concert site. In spite of the challenges for precisely representing and reasoning about these networks together, the interaction between networks offers attractive opportunities. In particular, the location information embedded in CDRs contributes to the understanding and prediction of human mobility and provides auxiliary information to the routing path design in transport networks. On the other hand, as depicted in Fig. 6, vehicle-connected networks might be another opportunity to highly promote the integration of cellular networks and transport networks. In vehicle-connected networks, individual vehicles are granted more intelligence to timely transmit/receive information from neighbouring vehicles and form clusters for automatic driving in a more organised manner. Moreover, the travel direction of some lanes could be adjusted according to the population of cars, so as to bring superior driving experience and shorten the time towards the destination. In a word, smarter cellular networks could become an indispensable part of smarter cities.

5 Conclusion

Based on the 'big data' in cellular networks, we have investigated the statistical modelling of three interesting scenarios including spatial deployment density of BSs, packet length or traffic volume of mobile services, as well as inter-arrival time and dwell time of human mobility. Compared to the commonly assumed uniform and memory-less feature in 3GPP, we have validated the heavy-tailed feature in these three scenarios. Specifically, we have found that the spatial density of BSs and the aggregated traffic in a BS follow α-stable distribution. Moreover, we have observed that the inter-arrival time and dwell-time of human mobility satisfy the power-law distribution. Afterwards, we have talked about the potential contribution to cellular networks. We have also discussed the implications (e.g. self-similarity, LRD, and spatial clustering) of the universal existing heavy-tailed feature and addressed how to reach the goals of smarter cellular networks. Finally, we have shown that cellular networks and transport networks are mutually dependent, which implies that there exists sufficient room available for cooperatively optimise both networks for smart cities. In a word, the 'big data' in cellular networks offers the design and planning cornerstone for smarter cities.

6 Acknowledgments

This work was financially supported by the Program for Zhejiang Leading Team of Science and Technology Innovation (no. 2013TD20), National Natural Science Foundation of China (no. 61731002, 61701439), Zhejiang Key Research and Development Plan (no. 2018C03056), the National Postdoctoral Program for Innovative Talents of China (no. BX201600133), the Project funded by China Postdoctoral Science Foundation (no. 2017M610369), the Project funded by Ministry of Industry and Information Technology of China for Testing, Solution Verification and Application Promotion of Industrial Information Physics System, and the Fundamental Research Funds for the Central Universities 2016XZZX001-04.

7 References

[1] Thakuriah, P.V., Tilahun, N.Y., Zellner, M.: 'Big data and urban informatics: innovations and challenges to urban planning and knowledge discovery', in Thakuriah, P.V., Tilahun, N., Zellner, M. '*Seeing cities through big data, Springer geography*' (Springer International Publishing, 2016), pp. 11–45. Available at http://link.springer.com/chapter/10.1007/978-3-319-40902-3_2, dOI: 10.1007/978-3-319-40902-3_2

[2] Becker, R.A., Caceres, R., Hanson, K., *et al.*: 'A tale of one city: using cellular network data for urban planning', *IEEE Pervasive Comput.*, 2011, **10**, (4), pp. 18–26

[3] Barabási, A.-L., Albert, R.: 'Emergence of scaling in random networks', *Science*, 1999, **286**, (5439), pp. 509–512

[4] Samorodnitsky, G.: '*Stable non-Gaussian random processes: stochastic models with infinite variance*' (Chapman and Hall/CRC, New York, 1994). Available at http://www.amazon.com/Stable-Non-Gaussian-Random-Processes-Stochastic/dp/0412051710

[5] Andrews, J.G., Baccelli, F., Ganti, R.: 'A tractable approach to coverage and rate in cellular networks', *IEEE Trans. Wirel. Commun.*, 2011, **59**, (11), pp. 3122–3134

[6] Zhou, Y., Li, R., Zhao, Z., *et al.*: 'On the alpha–stable distribution of base stations in cellular networks', *IEEE Commun. Lett.*, 2015, **19**, (10), pp. 1750–1753

[7] Deng, N., Zhou, W., Haenggi, M.: 'The Ginibre point process as a model for wireless networks with repulsion', *IEEE Trans. Wirel. Commun.*, 2015, **14**, (1), pp. 107–121

[8] Zhou, X., Zhao, Z., Li, R., *et al.*: 'Human mobility patterns in cellular networks', *IEEE Commun. Lett.*, 2013, **17**, (10), pp. 1877–1880

[9] Li, R., Zhao, Z., 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**, pp. 1416–1422

[10] Li, R., Zhao, Z., Zheng, J., *et al.*: 'The learning and prediction of application-level traffic data in cellular networks', *IEEE Trans. Wirel. Commun.*, 2017, **16**, (6), pp. 3899–3912. Available at http://arxiv.org/abs/1606.04778

[11] Crovella, M., Bestavros, A.: 'Self-similarity in world wide web traffic: evidence and possible causes', *IEEE/ACM Trans. Netw.*, 1997, **5**, (6), pp. 835–846

[12] Xiao, Z., Guo, L., Tracey, J.: 'Understanding instant messaging traffic characteristics'. Proc. IEEE (ICDCS 2007), Toronto, ON, Canada, 2007

[13] Du, Y., Fan, J., Chen, J.: 'Experimental analysis of user mobility pattern in mobile social networks'. Proc. IEEE (WCNC 2011), Cancun, Quintana Roo, Mexico, 2011

[14] Zhou, X., Zhao, Z., Li, R., *et al.*: 'Understanding the nature of social mobile instant messaging in cellular networks', *IEEE Commun. Lett.*, 2014, **18**, (3), pp. 389–392

[15] Rhee, I., Shin, M., Hong, S., *et al.*: 'On the Levy-walk nature of human mobility'. Proc. IEEE (INFOCOM 2010), Phoenix, AZ, USA, 2008

[16] Weron, A., Burnecki, K., Mercik, S.: 'Complete description of all self-similar models driven by Levy stable noise', *Phys. Rev. E*, 2005, **71**, (1), pp. 016113

[17] Xu, F., Lin, Y., Huang, J., *et al.*: 'Big data driven mobile traffic understanding and forecasting: a time series approach', *IEEE Trans. Services Comput.*, 2016, **9**, (5), pp. 796–805

[18] Wang, J., Tang, J., Xu, Z., *et al.*: 'Spatiotemporal modeling and prediction in cellular networks: a big data enabled deep learning approach'. Proc. IEEE (INFOCOM 2017), Atlanta, GA, USA, 2017

[19] Barrett, C., Channakeshava, K., Huang, F., *et al.*: 'Human initiated cascading failures in societal infrastructures', *PLOS ONE*, 2012, **7**, (10), p. e45406