Which statistical distribution best characterizes modern cellular traffic and what factors could predict its spatio-temporal variability?

Shruti Bothe, Haneya Naeem Qureshi, Student Member IEEE and Ali Imran, Senior Member IEEE

***Abstract*—Spatio-temporal characterization of user traffic is the first step in designing, optimizing and automating a mobile cellular network. While it is well known that voice telephony follows Poisson distribution, the distribution of SMS and internet data usage along with voice calls and the factors influencing the distribution, is still an open question. We characterize the distribution of multi-faceted cellular traffic while identifying the factors influencing the parameterization of the distribution. Eight latent features that play a statistically significant role to characterize the traffic distribution variations over time and space are determined by leveraging a large real dataset. The features used to characterize the dynamism of the traffic distribution are Points of Interest, day of the week, special events and region. Results show that Generalized Extreme Value distribution best describes SMS, call and internet activity and it does not change with spatio-temporal features. Also, traffic distribution is not stationary. Insights gained from this analysis can pave the way towards more precise and resource efficient planning, designing and optimization of future cellular networks.**

***Index Terms*—Big Data, Traffic Distribution, Milano Data, Distribution Analysis, Network Design**

I. Introduction

The first step and a key component towards planning, designing and optimizing networks is identification of user traffic requirements [1], [2] and to find a distribution model for network traffic. Characterizing the traffic enables us to determine the operational and capital expenditure, as well as predict the long-term performance of the network, which in turn can be used for optimal design of networks and maximizing the return on investment. This can help network operators to minimize total cost of ownership, boost network capacity, maximize coverage, minimize power consumption and even optimize handover zones [2]. Furthermore, scenarios such as fault prediction and detecting outages [3] can be better handled with prior knowledge of the traffic distribution, which enable better resource allocation and minimize the true negative impact of outages.

Characterization of user traffic has been done for voice telephony that has been the foundation of 2G and 3G system planning. However the characterization of data traffic i.e., SMS, internet, Voice over Internet Protocol (VoIP) still remains unexplored. In addition, variability of traffic distribution over space - Points of Interest (PoI), urban, sub-urban or rural region, or over time - weekdays, weekends, and special events remains under investigated. Current understanding of cellular traffic remains largely limited to two classic findings: first, the most widely used and oldest traffic models stating that voice telephony follows Poisson distribution and second, the traffic volume fluctuates over different times of the day [4].

A. Relevant Work

By harnessing the massive amount of data in mobile networks such as Call Detail Record (CDR) that remained largely unexploited in the past, user traffic distributions can be characterized. For example, some recent studies leverage CDR data to study mobility patterns [5], specifically in urban areas [6], anomaly detection [7] and to devise traffic prediction models [8]. Authors in [9] and [3] use CDR data for cell planning and resource allocation solutions. The authors in [10] investigate the use of temperature as a predictor for determining traffic activity levels. Authors in [11] proposed a shifted gamma distribution model to characterize internet traffic consisting of packet arrival time and size without considering spatial parameters. The authors in [12] utilize the fact that incoming packet cycles are quasi-periodic to determine anomalies in the network. However, the question of which distribution of call, internet and SMS traffic taking into account both spatial as well as temporal features remains unanswered.

The most relevant to this study are recent works in [3] and [13]. Using the same data set as used in this study, authors in [3] built support vector machine based traffic prediction model for devising an energy saving scheme. This study differs from [3] and [8] as we do not try to create a traffic prediction model, since such a model is only applicable to network for which it is trained and will have to be retrained for each network, location and time window. Instead we determine the underlying distribution of the traffic and quantify the parameterization of that distribution as a function of the factors that influence the shape of the distribution. Thus, this study offers more broadly applicable findings compared to specific traffic prediction model proposed in [3] and other similar studies. In [13] authors use k-means clustering to cluster base-stations with respect to activity levels. A spatial analysis was done to determine that time-correlation of call arrivals is influenced by time and the location of base stations. It is concluded that a Poisson process can be used to model call arrival rates. However, the authors in [13], highlight the inability of Poisson model to capture traffic burstiness which characterizes data traffic (as opposed to voice traffic in old telephone systems). Hence, the results in [13] further highlight the need for analysis to characterize non-voice traffic while taking into account spatio-temporal parameters that may influence the shape of distribution of the modern multi-faceted cellular traffic. The novelty of this paper lies in the finding that compared to 80 known distributions, the behavior of SMS, call and internet activity can be modelled best using Generalized

Extreme Value (GEV) distribution. Furthermore, we quantify the three parameters that define GEV as a function of spatial or temporal features such as PoI, non-PoI, urban, sub-urban, rural areas, weekdays, weekends as well as special events and holidays.

B. Contributions and Organization

To the best of authors' knowledge, this paper is the first attempt to characterize significant features in time and space that dictate the traffic distribution of SMS, call and internet values using a large real data set. More specifically, this paper provides answers to the following questions:

1) What is the distribution that SMS, call and internet activities follow?
2) Are the distributions of SMS, call and internet traffic stationary or non-stationary?
3) If the distributions are non-stationary, what are the parameters that affect the dynamism of these distributions?
4) How significant are these parameters and can they be quantified?

Section II provides the framework for traffic distribution analysis. Numerical results and traffic distribution parameters for various scenarios are presented in Section III. Section IV concludes the paper.

II. Framework for Traffic Distribution Analysis

A. Data Set

In order to develop a realistic model for traffic distribution, we have leveraged two real data sets that take into consideration spatial and temporal characteristics:

1) The Milano CDR data set: This extensive CDR dataset provides SMS, call and internet traffic activity values for two months. The data has been aggregated to 10-minute interval times.
2) Trip Advisor's Point of Interest data set: This data set provides 357 spots in the form of latitudinal and longitudinal values that have been classified as historical, administrative and institutional spots in the city of Milan. The urban region centered at latitude = 45.4642 and longitude = 9.1900, has a radius of 5.2 miles and includes 246 PoIs. The sub-urban region covers an additional 2.6 miles and has 85 PoIs. The rest of the region is classified as rural and has 26 PoIs.

To capture both spatial and temporal characteristics into the traffic distribution, the auxiliary data set obtained from Trip Advisor is super-imposed on the Milan grid using MapBox [14] as shown in Fig 1.

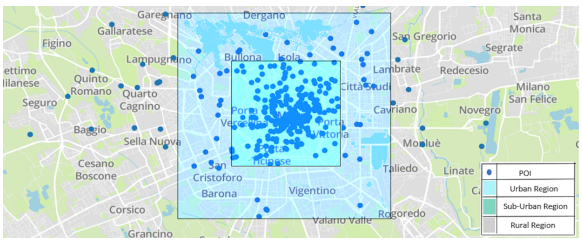

Fig. 1: Spatial classification of Milan data set.

B. Data Preprocessing

Two key problems are addressed in data preprocessing:
1. Recovery of missing values 2. Elimination of outliers.

1) Recovering the Missing Data: A major challenge in the used data, that is often the case with any real data, is that 45% of the entries are missing. These missing values were noted to be Missing at Random (MAR). First, a simple method of imputation for replacing the missing values with the average of previous and next time slot values was administered. However, this method did not yield satisfactory results because of large imbalances in the data. Therefore, a more sophisticated technique such as matrix factorization was adopted.

To better predict the missing values, we first perform singular value decomposition (SVD) of the multiply imputed matrix $\mathbf{R}$ to get a lower rank (i.e., smaller/simpler) approximation of the original matrix [15] as:

$$\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}} \tag{1}$$

where the diagonal entries of $\mathbf{\Sigma}$ are equal to the singular values of $\mathbf{R}$, and $\mathbf{V}^{\mathbf{T}}$ represents data matrix consisting of SMS, call and internet values. $\mathbf{U}$ denotes the features matrix. To get the lower rank approximation, we keep only the top 8 features, which include features such as day of the week, PoI and regions as well as rare features such as special events.

Missing values are then estimated by minimizing the sum of squared residuals, one feature at a time, using gradient descent with regularization [16]. The minimization problem in order to predict the missing values can be formulated as:

$$\min_{\mathbf{p}_u,\mathbf{q}_i} \sum_{r_{ui}\in\mathbf{R}} (r_{ui} - \mathbf{p}_u.\mathbf{q}_i)^2 \tag{2}$$

where $\mathbf{p}_u$ is a set of row vectors, where each vector consists of the u-th row of $\mathbf{U}$ and $\mathbf{q}_i$ is a set of column vectors, each consisting of i-th column of $\mathbf{V}^{\mathbf{T}}$. All the vectors $\mathbf{p}_u$ and $\mathbf{q}_i$ are mutually orthogonal.

After random initialization, vectors $\mathbf{p}_u$ and $\mathbf{q}_i$ are updated using the following rules: $\mathbf{p}_u = \mathbf{p}_u + \alpha.\mathbf{q}_i(r_{ui} - \mathbf{p}_u.\mathbf{q}_i)$ and $\mathbf{q}_i = \mathbf{q}_i + \alpha.\mathbf{p}_u(r_{ui} - \mathbf{p}_u.\mathbf{q}_i)$ where α is the learning rate and is set to 0.01. The goal is to bring r_{ui} as close as possible to the actual values of $\mathbf{p}_u$ and $\mathbf{q}_i$

Once the vectors $\mathbf{p}_u$ and $\mathbf{q}_i$ have been updated, we can estimate the missing values as:

$$\hat{r}_{ui} = \mathbf{p}_u.\mathbf{q}_i \tag{3}$$

This method can handle multi-dimensional data while being computationally efficient. It also outperforms multiple imputation techniques in particular when the qualitative variables have many categories and some of them are rare.

2) Detecting outliers and eliminating duplicate values: It is well known that real traffic in cellular network is multi-scale and even fractal (particularly at locations where special events occur). Therefore, while pre-processing it was ensured that any physiological fluctuation from the distribution was kept. This includes examples such as special events or certain hours of weekends when traffic pattern differs from regular days and PoI combined with special events. Data values lesser

than 1.5 times quartile 1 and greater than 1.5 times quartile 3 were dropped while preserving the traffic characteristics. Next, duplicate values were deleted. Elimination of outliers and duplicate values reduced the data set size by 22%.

C. Consolidating the data

For ease of processing, data points were aggregated from 10-minute interval data to 1-hour for each bin ID. Analysis was also done for 3-hour and 6-hour time intervals. However, the best data compression while conserving the traffic characteristics was observed for 1-hour aggregation.

III. Results and Analysis

Following the framework in Section II, the traffic distribution for SMS, call and internet is analyzed for a period of two months. As an example, Fig. 2 depicts the SMS traffic during the month of December. As seen in the figure, SMS traffic activity has a varied pattern. It can also be noted that for 25th and 31st of December, the SMS activity is exceptionally high. This motivated us to cluster the user activity for the entire data (over all cells) into weekdays and special events.

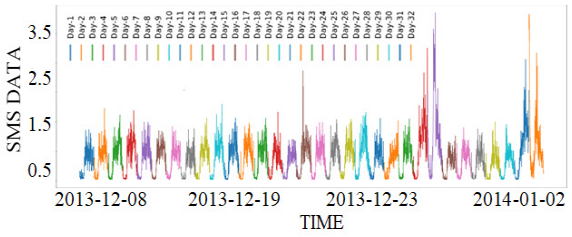

Fig. 2: SMS traffic pattern for 1 cell over the month of December.

For a robust analysis, the spatio-temporal features result in a total of 18 combinations as shown in Table I. Values 0 and 1 in special events, weekday and PoI columns indicate false and true values respectively. Values 1, 2 and 3 in the region column indicate rural, sub-urban and urban regions respectively. Eighty distributions were tested on each of the histograms generated. GEV distribution best describes SMS, call and internet traffic. The Probability Density Function (PDF) of GEV is given by:

$$f(x) = \frac{1}{\sigma} t(x)^{\zeta+1} e^{-t(x)} \tag{4}$$

where,

$$t(x) = \begin{cases} (1+\zeta(\frac{x-\mu}{\sigma}))^{-1/\zeta} & \text{if } \zeta \neq 0 \\ e^{-(x-\mu)/\sigma} & \text{if } \zeta = 0 \end{cases} \tag{5}$$

and μ, σ and ζ are the location, scale and shape parameters of the GEV distribution repectively.

Table I provides the shape, scale and location parameters of GEV distribution of SMS, call and internet traffic for each of the 18 cases. These parameters change over time and space, highlighting the non-stationarity of traffic distribution. This table can directly be used by network operators with similar demographics as Milan to configure various network parameters based on spatio-temporal characteristics in the area of interest by first classifying the given area spatially, i.e., into rural, sub-urban or an urban region and if it is a PoI or non-PoI. Next, classification needs to be done temporally i.e., whether the planning or optimization needs to be done on a weekend, weekday or on a special event. After the spatio-temporal classification, relevant rows can be matched to select the GEV parameters for SMS, call and internet. As an example, the SMS distribution for one particular case of non-PoI and rural region on a special event day is illustrated in Fig. 3.

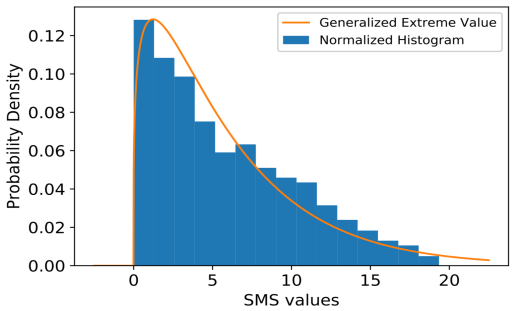

Fig. 3: Distribution fitting for SMS activity in cells for non-PoI and rural area on a special event.

A. Validation

In order to validate our findings, the Kolmogorov-Smirnov (K-S) test was done. K-S statistics for GEV distribution was compared to three closest distributions, namely, Generalized Pareto, Beta and Weibull and is represented as a box-plot in Fig. 4. As seen in Fig. 4, the variance in K-S values of

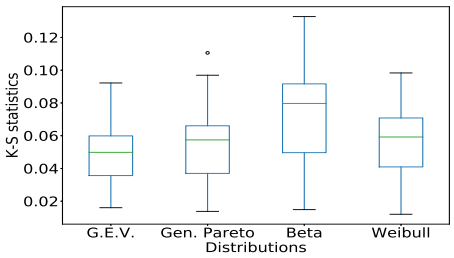

Fig. 4: Box plot of Kolmogorov-Smirnov statistic for different distributions over all scenarios.

GEV are the least as compared to the other distributions, thus validating that GEV is the best fitting distribution over all combinations of weekdays, weekends, special events, PoI and regions, making it ideal for traffic distribution modelling. The values of σ, ζ and μ were determined for all 18 combinations and are listed in Table 1. While GEV when characterized by appropriate parameter values as indicated in Table 1 seem to well represent the traffic distribution for a wide variety of spatial regions and temporal spans, a further fine tuning of the distribution parameters may be required for a region whose underlying demographics and user behavior do not match with that of Milan. While results show that GEV is the best fit, Generalized Pareto and Weibull distributions may also be used in practice with reasonable overall accuracy.

TABLE I: Parameters of GEV distribution of SMS, call and internet values in different scenarios

				SMS			CALL			INTERNET		
Special Events	Weekday	PoI	Region	σ_s	ζ_s	μ_s	σ_c	ζ_c	μ_c	σ_i	ζ_i	μ_i
1	-	0	1	-0.362	8.178	-0.137	-0.263	8.274	3.765	-0.467	102.680	6.338
1	-	0	2	-0.444	6.247	16.141	-0.490	6.847	12.992	-0.581	6.397	198.150
1	-	0	3	-0.583	4.869	20.328	-0.626	5.987	14.431	-0.728	10.837	167.843
1	-	1	1	0.248	10.376	21.593	0.265	18.825	15.293	0.162	109.857	200.369
1	-	1	2	0.190	20.364	36.032	0.197	20.055	28.932	0.156	212.340	360.42
1	-	1	3	0.188	71.532	177.720	-0.152	87.115	148.170	0.144	628.440	1804.5
0	1	0	1	-0.488	7.543	15.938	-0.498	9.436	11.650	-0.072	45.77	52.053
0	1	0	2	-0.605	10.457	21.698	-0.644	10.821	19.084	-0.642	74.066	205.540
0	1	0	3	-0.825	16.967	45.852	-0.746	13.875	30.567	-0.129	105.764	260.782
0	1	1	1	0.013	16.764	22.347	0.097	8.433	14.256	0.372	76.1857	201.389
0	1	1	2	0.166	20.181	28.240	-0.359	12.556	13.277	0.701	187.355	329.913
0	1	1	3	0.389	44.051	57.327	0.185	19.949	76.739	-0.379	336.620	410.840
0	0	0	1	0.037	3.549	6.692	0.148	4.815	7.173	-0.036	42.165	54.209
0	0	0	2	-0.046	6.158	16.214	-0.049	6.921	12.927	0.057	64.269	195.460
0	0	0	3	-0.189	8.274	20.991	0.498	10.937	15.484	0.195	101.389	232.295
0	0	1	1	0.357	11.543	28.163	0.396	14.844	21.773	0.193	114.054	248.185
0	0	1	2	0.198	19.893	35.546	0.192	19.732	28.193	0.164	204.410	352.290
0	0	1	3	0.181	71.056	178.560	0.043	86.236	148.770	0.142	624.640	1805.400

IV. Conclusion

In this paper, we have characterized the distribution of user traffic by using real CDR dataset from Telecom Italia, Trip Advisor's PoI data and public calendars for Milan, Italy. Extensive analysis on the fused dataset shows that network traffic distribution for SMS, call and internet for all combinations of spatio-temporal features follows GEV distribution. The results of this study can be used to predict, quantify and manage traffic in an area of interest. The insight gained can be used to optimize several aspects of networks like optimal base station placement, switching on and off cells for energy saving and various other self organizing network functions.

Acknowledgment

This material is based upon the work supported by National Science Foundation (NSF) under Grant Numbers 1559483 and 1619346. For more details, please visit: http://www.ai4networks.com

References

[1] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: How to empower SON with big data for enabling 5G," *IEEE Network*, vol. 28, no. 6, pp. 27–33, Nov 2014.

[2] A. Taufique, M. Jaber, A. Imran, Z. Dawy, and E. Yacoub, "Planning wireless cellular networks of future: Outlook, challenges and opportunities," *IEEE Access*, vol. 5, pp. 4821–4845, 2017.

[3] Y. Kumar, H. Farooq, and A. Imran, "Fault prediction and reliability analysis in a real cellular network," in *Wireless Communications and Mobile Computing Conference (IWCMC), 2017 13th International*. IEEE, 2017, pp. 1090–1095.

[4] A. Zoha, A. Saeed, H. Farooq, A. Rizwan, A. Imran, and M. A. Imran, "Leveraging intelligence from network CDR data for interference aware energy consumption minimization," *IEEE Transactions on Mobile Computing*, vol. 17, no. 7, pp. 1569–1582, July 2018.

[5] M. C. Gonzalez, C. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," vol. 453, pp. 779–82, 07 2008.

[6] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira Jr, and C. Ratti, "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example," *Transportation research part C: emerging technologies*, vol. 26, pp. 301–313, 2013.

[7] B. Hussain, Q. Du, and P. Ren, "Deep learning-based big data-assisted anomaly detection in cellular networks," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec 2018, pp. 1–6.

[8] C. Zhang, H. Zhang, D. Yuan, and M. Zhang, "Citywide cellular traffic prediction based on densely connected convolutional neural networks," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1656–1659, Aug 2018.

[9] P. D. Francesco, F. Malandrino, and L. A. DaSilva, "Assembling and using a cellular dataset for mobile network analysis and planning," *IEEE Transactions on Big Data*, pp. 1–1, 2017.

[10] M. N. Rafiq, H. Farooq, A. Zoha, and A. Imran, "Can temperature be used as a predictor of data traffic: A real network big data analysis," *in Proc. 5th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), Zurich*, pp. 1–1, 2018.

[11] S. Kim, J. Y. Lee, and D. K. Sung, "A shifted gamma distribution model for long-range dependent internet traffic," *IEEE Communications Letters*, vol. 7, no. 3, pp. 124–126, March 2003.

[12] A. D'Alconzo, A. Coluccia, F. Ricciato, and P. Romirer-Maierhofer, "A distribution-based approach to anomaly detection and application to 3G mobile traffic," in *GLOBECOM 2009 - 2009 IEEE Global Telecommunications Conference*, Nov 2009, pp. 1–8.

[13] S. Zhang, D. Yin, Y. Zhang, and W. Zhou, "Computing on base station behavior using erlang measurement and call detail record," *IEEE Transactions on Emerging Topics in Computing*, vol. 3, no. 3, pp. 444–453, Sept 2015.

[14] "Mapbox. Available at: https://www.mapbox.com/."

[15] G. Ding, J. Wang, Q. Wu, Y. Yao, F. Song, and T. A. Tsiftsis, "Cellular-Base-Station-Assisted Device-to-Device Communications in TV white space," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 107–121, Jan 2016.

[16] A. Paterek, "Improving regularized singular value decomposition for collaborative filtering," in *Proceedings of KDD cup and workshop*, vol. 2007, 2007, pp. 5–8.