The Predictability of Cellular Networks Traffic

Xuan Zhou*†, Zhifeng Zhao*†, Rongpeng Li*†, Yifan Zhou*†, and Honggang Zhang*†
*York-Zhejiang Lab for Cognitive Radio and Green Communications
†Dept. of Information Science and Electronic Engineering
Zhejiang University, Zheda Road 38, Hangzhou 310027, China
Email: {zhouxuan, zhaozf, lirongpeng, zhouyftt, honggangzhang}@zju.edu.cn

***Abstract*—In order to improve the energy efficiency and resource management of cellular networks, traffic modeling and prediction has been focused in recent years. In this paper, we take advantage of entropy theory to explore the limits of predictability of cellular network traffic based on large amount of traffic dataset gathered from real cellular network in China. By categorizing traffic according to voice, text and data group, we investigate random entropy of each type of traffic, as well as conditional entropy by temporal, spatial and service related information. Our key findings are that (1) traffic can be well predicted by preceding 15 hours traffic, (2) voice traffic has so close similarity to text traffic in the same cell that we can use one of them to predict the other, (3) knowledge of adjacent cells traffic can enhance the predictability of voice and text more than data. Considering the large amount traffic dataset which contains thousands of base stations and billions of records, the impact of dataset pre-processing, quantization and time resolution are also taken into account and are discussed. Moreover, macroscopic view of entropy distribution is presented by geo-location markers.**

I. Introduction

A. Backgrounds

With the popularization of mobile devices and development of smart phones, traffic of cellular networks is growing rapidly. The surge of traffic is incurring huge challenge to mobile operators to allocate resource, thus making the analysis of cellular networks an imperative. [1] analyzes application usage in a 3G cellular data network, [2] proposes a profile-based approach to green cellular infrastructure with real traffic traces and actual Base Station (BS) deployment map. The researchers in [3] present a study on the performance of HSPA networks in Hong Kong.

Besides that, some researchers have shown the possibility to improve the energy efficiency, depending on the prediction result of traffic [4], [5]. In order to characterize the traffic of the cellular network, [5] proposes algorithm of movement prediction to reduce energy consumption, [6] shows that practical traffic load profile is periodical and can be approximated by a sinusoidal function.

B. Motivation

Fig.1 shows average cell traffic of voice, text and data services in one Base Station Controller (BSC) during seven days. We can easily draw an intuitive impression from it that any type of service traffic in one day resembles those in other days. Besides, different types of service traffic has similar shape. Yet, the regular pattern of traffic would bring up one question: to what extent can the traffic be predicted with certain prior information such as historical traffic, other types of service traffic and adjacent cells' traffic. In order to answer that, we adopt anonymous traffic records of roughly 7000 BSs in one month from China Mobile. Afterwards, we utilize entropy theory to quantify the information measurement that temporal, service and spatial traffic will provide for prediction. In the field of network research, entropy theory is widely employed to characterize traffic pattern and subscriber mobility. [7] makes solid research in the predictability of human mobility, and [8] uses entropy theory to estimate QoS parameters in ATM traffic streams. To our best knowledge, our work is the first to adopt entropy theory to analyze the predictability of cellular networks traffic.

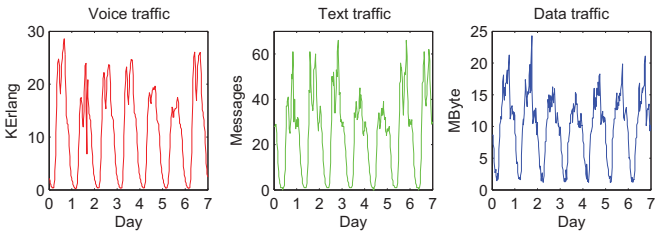

Fig. 1. Average cell traffic of a BSC (Base Station controller) in one week, each point represents measurement of traffic in half an hour.

C. Paper organization

The rest of the paper is organized as follows: In Section II, we cover the fundamentals over cellular networks and present the details of our dataset. Section III presents the results that we use entropy theory to describe the predictability of the traffic. Section IV discusses parameters we use in dataset processing and the geo-location distribution of high and low entropy cells. Section V concludes our findings.

II. Overview of the Dataset

A. Network architecture

In this section, we go through some basic concepts of cellular network involved in the paper. Fig.2 depicts the typical architecture of a GSM/UMTS hybrid cellular network, which consists of access network (AN) and core network (CN). AN is the edge part of a telecommunications network which connects subscribers to their service provider, and CN is the central part

978-1-4673-1157-1/12/$31.00 © 2012 IEEE

of the network that provides services to subscribers who are connected by AN.

In AN, subscriber uses mobile station to connect to Base Transceiver Station (BTS) or NodeB (we usually call them base stations), which can be identified by cell ID. In CN, MSC processes the voice and text demand in Circuit Switch (CS) domain, as well as Serving GPRS Support Node (SGSN) processes the data demand in Packet Switch (PS) domain. AN has complex forms of connections and is adjusted frequently, therefore it is difficult to monitor every BS. In this paper, we use Signaling Monitor System in CN to get the traffic information.

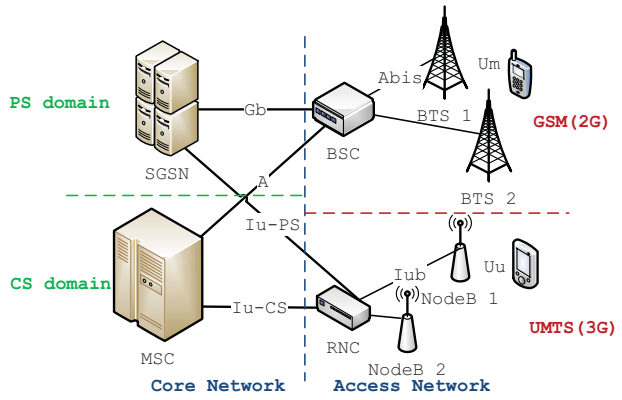

Fig. 2. The typical architecture of a GSM/UMTS hybrid cellular network.

B. Dataset description

In this study, we collected traffic records from 9 MSCs and SGSNs with about 7000 BSs with coverage of 780 km^2. Both rural and urban areas are covered in the dataset, containing residential areas, business buildings, hotels, schools and farms, etc. The dataset consists of both GSM and UMTS BSs traffic from March to April in 2012, serving about 3 million subscribers.

In AN resource view, subscriber's service demands fall into three categories: voice, text and data. In order to explore the characteristics of cellular traffic, we collect anonymous traffic records of the three services from MSCs and SGSNs. The information that records contain is listed in Table.I : $Time$ field in the dataset indicates the time that specific service completes; the serving BS of the record is identified by $Cell\ ID$; $Call\ Duration$ is measured in millisecond; $Download\ Volume$ and $Upload\ Volume$ describe packet data volume in terms of bytes.

TABLE I
THE RECORD FIELDS OF COLLECTED DATASET

Service	Record Fields
Voice	Time, Cell ID, Call Duration
Text	Time, Cell ID
Data	Time, Cell ID, Download Volume, Upload Volume

To conveniently process the massive dataset, we group the traffic records by time and cell ID, then sum up call duration and download/upload volume if they occur in the same time period and in the same BS. Suppose that time period is T_p minutes, each service with same cell ID has $24*60/T_p$ records in one day after processing. Then we get the processed traffic array $Tr(i)$ of each BS, where i denotes the ith time period, and $Tr(i)$ is measured in erlangs with voice service, messages with text service and bytes with data service.

III. THE UNCERTAINTY OF THE TRAFFIC

Entropy measures the disorder or more precisely the uncertainty of events. Shannon defined the entropy H of a discrete random variable X with possible values $\{x_1, \ldots, x_n\}$ as:

$$H(X) = \sum_{i=1}^{n} p(x_i) \log_b p(x_i), \tag{1}$$

where b is the base of the logarithm used whose common value of b is 2 when the unit of entropy is bit.

In this paper, traffic is recognized as a random variable, so we can use entropy to describe the predictability of traffic. We assign the following types of entropy to different prior information cases: (1) random entropy without any prior information, (2) temporal entropy with preceding T_{prior} minutes traffic, (3) service entropy with traffic of other services, (4) spatial entropy with adjacent cells traffic.

With the purpose of describing the traffic in the view of information theory with discrete values, we quantize the cell traffic to Q levels as $Tq(i)$:

$$Tq(i) = floor\left(Q \cdot \frac{Tr(i) - \min(Tr)}{\max(Tr) - \min(Tr)}\right), \tag{2}$$

where i denotes the ith time period, $\max(Tr)$ and $min(Tr)$ denotes the maximum and minimum traffic value of the cell during the timespan of interest. The $floor$ function rounds the elements to the nearest integer less than or equal to it, therefore $Tq(i) \in \{0, 1, \ldots, Q-1\}$. Then we calculate the probability of each level appears in traffic array $Tq()$:

$$p(X_k) = \frac{N_k}{\sum_{k=1}^{n} N_k} \tag{3}$$

where X_k is the event that $Tq(i)$ value equals k ($k \in \{0, 1, \ldots, Q-1\}$), and N_k is the count that elements of $Tq()$ equal k.

A. Random entropy

Now, random entropy of three services (voice, text and data) in all cells is calculated using (1), where $Q = 6$ and $T_p = 30$ minutes. We select a pair of cells with high and low random entropy of each service, whose entropy values and distributions are listed in Table.II.

Table.II shows that the low entropy cell has more zero-level traffic, and traffic distribution of the high one is more uniform. We observe the random entropy of all cells and draw the corresponding histograms and cumulative distributions with respect to voice, text and data in Fig.3. It can be inferred from

TABLE II
RANDOM ENTROPY VALUES AND DISTRIBUTIONS OF TRAFFIC.

Service	entropy	$p(X_0)$	$p(X_1)$	$\sum_{k=2}^{5} p(X_k)$
Voice	2.32	0.32	0.27	0.41
Voice	1.25	0.67	0.10	0.23
Text	2.24	0.43	0.29	0.28
Text	0.123	0.88	0.10	0.02
Data	2.05	0.64	0.23	0.13
Data	0.029	0.95	0.03	0.02

Fig.3 that more than 80% cells have random entropy larger than 1 bit in voice and text service, which means the difficulty to predict traffic in these cells is harder than guessing the right side of a coin. To make the traffic more predicable, prior information needs to be employed to reduce the uncertainty of the subsequent traffic.

Findings: voice traffic value is more uniform however data service has more low-valued traffic, and more than 70% cells have low random entropy of data traffic.

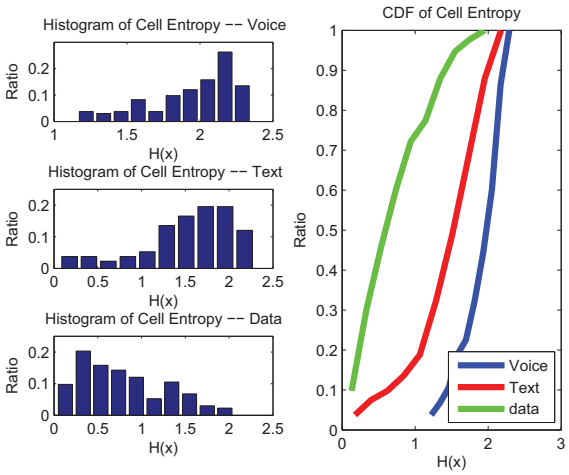

Fig. 3. Histograms and CDFs of cell random entropy.

B. Temporal entropy

From Fig.1, we may predict the traffic based on past traffic according to the intuition that traffic in each day has the same appearance. Therefore, we may concern the question: how much information can we get from preceding T_{prior} minutes traffic, or how much uncertainty is still there even though the prediction is based on preceding T_{prior} minutes traffic knowledge?

It is known that the conditional entropy of two events X and Y, which takes values x_i and y_j respectively, as

$$H(X|Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)} \quad (4)$$

where $p(x_i, y_j)$ is the joint probability of $X = x_i$ and $Y = y_j$. Here we define variable X as the current traffic and define Y as the historical traffic in preceding T_{prior} minutes, which contains T_{prior}/T_p data points with Q^{T_{prior}/T_p} possible values. In Fig.4, histogram and CDF of cell temporal entropy is shown when T_{prior} is set to 120 minutes and 600 minutes, $Q = 6$ and $T_p = 30$ minutes. We find that more than 90% cells have the temporal entropy less than 0.2 bit when preceding 10 hours traffic is employed. Fig.4 (d) demonstrates that introducing preceding traffic can increase traffic predictability much ,and temporal entropy of three services converges into the same distribution shape when T_{prior} augments.

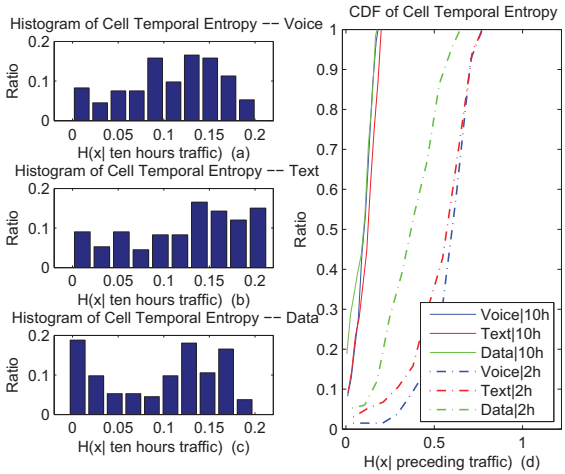

Fig. 4. Histograms and CDFs of cell temporal entropy with preceding 10 hours traffic.

To see the distinctions of traffic between high and low temporal entropy, we draw each type of service traffic in cell with high temporal entropy as well as that with the low one in Fig.5. The red line of voice traffic looks more regular, yet the blue line has relative low traffic in weekend, which leads to irregularity. When it comes to the blue line of text traffic, a peak value in Saturday evening, makes it hard to predict. Likewise, more irregular peak values appear in the blue line of data traffic. Compared to that, the red one is a little easier to recognize by daily division. Fig.6 (Left) sketches the variation of temporal entropy to historical time corresponding to traffic in Fig.5.

Given that larger T_{prior} would contribute more information to prediction, we show the average temporal entropy of all cells to the length of historical time in Fig.6 (Right). Although there is more entropy on Q=8 than Q=6 when $T_{prior} < 2$ hours, both curves decrease rapidly when T_{prior} increases, and converge to less than 0.1 bit when the historical time is larger than 15 hours. In other words, we can predict the traffic precisely once we obtain the more than 15 hours preceding traffic information. Furthermore, data traffic is a little easier to predict than voice/text traffic when $T_{prior} \leq 10$ hours, while if $T_{prior} > 10$ hours, data traffic becomes the hardest one to predict.

Findings: traffic can be well predicted by preceding 15 hours traffic ,and the predictability order is: voice>text>data when 15 hours' historical records are employed.

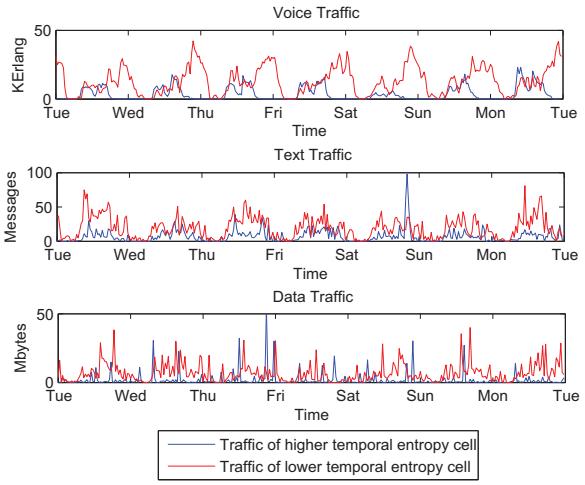

Fig. 5. Cell traffic of high and low temporal entropy. Temporal entropy is calculated with Q=6, T_{prior}=600 and T_p=30.

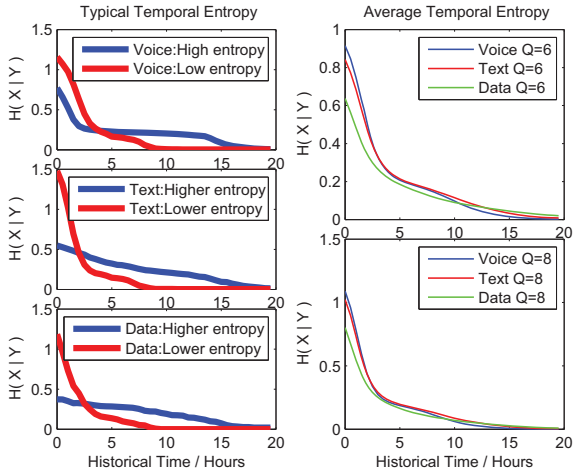

Fig. 6. Left: Temporal entropy to historical time in high and low temporal entropy cell. Right: Average temporal entropy to historical time in all cells with Q=6 and Q=8.

C. Service entropy

As mentioned above, three types of service traffic have something in common. Mutual information is used to describe the dependence of the two services which can be defined as:

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \qquad (5)$$

where X and Y is the traffic of two services in the same cell, $p(x,y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively. Fig.7 shows sample traffic of the high mutual entropy cell as well as the low one. To be specific, Fig.7 (a)(b) sketch the comparison of voice-text traffic between cells of high and low service entropy, Fig.7 (c)(d) sketch that of voice-data traffic. In Fig.7 (a)(c), two services

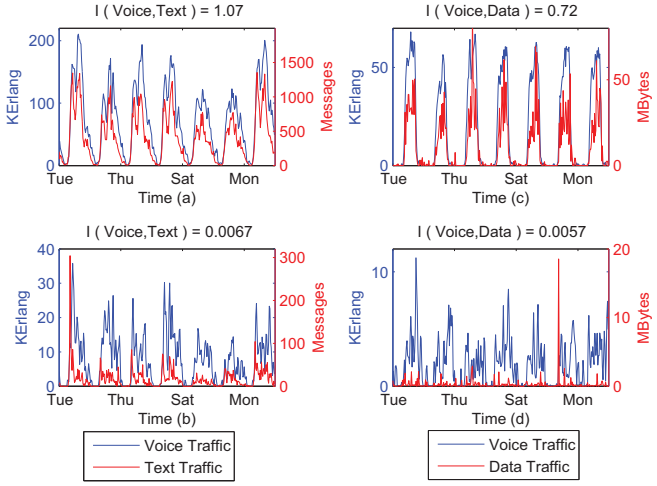

Fig. 7. Cell traffic of high and low mutual entropy. Mutual entropy is calculated with Q=6.

have the same appearance with different scales, which means they have high similarity. Now we consider mutual entropy of each pair of services in all cells. We summarize the mean values and variances of service mutual entropy in Table.III, and show the histograms and CDFs in Fig.8. Mutual entropy between voice and text traffic is distinct from others: not only the mean value of it is about 0.5 bit higher than the other two, but also more than half of cells in the other two mutual entropy pairs provide little valid information.

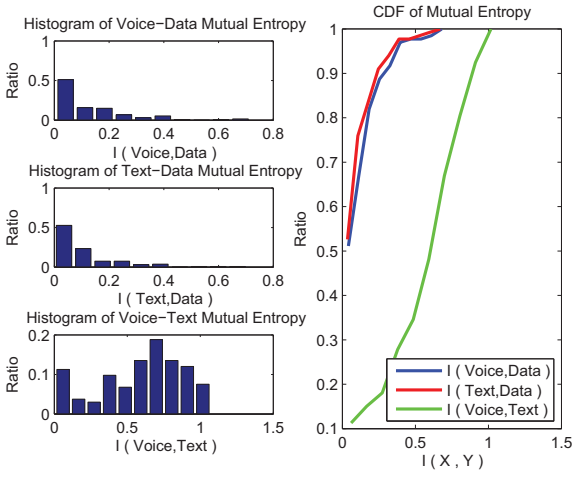

Fig. 8. Histograms and CDFs of mutual entropy in all cells.

TABLE III
MUTUAL ENTROPY STATISTICS IN FIG.8

Service	Mean	Variance	Min	Max
I(Voice,Data)–blue line	0.1305	0.0184	0.0057	0.7158
I(Text,Data)–red line	0.1091	0.0161	0.0002	0.7049
I(Voice,Text)–green line	0.5943	0.0833	0.0067	1.0709

Findings: data traffic in more than 50% cells has no contribution to voice and text prediction, whereas voice traffic has so close similarity to text traffic in the same cell that we can use one of them to predict the other.

D. Spatial entropy

In this part, we measure how much information we can get from the adjacent cells. We use the conditional entropy defined in (4) to describe the spatial entropy:

$$H(X|X_1 \dots X_n) = H(X, X_1 \dots X_n) - H(X_1 \dots X_n) \quad (6)$$

where X is the traffic of the concerned cell and $X_i (i \in \{1, 2 \dots, n\})$ denotes the traffic of ith adjacent cell around X. Firstly, we set n to 1, Fig.9 (d)(e) show the traffic in concerned cell and neighboring cells with high and low spatial entropy. In Fig.9 (e), there are so apparent lags in the blue line when traffic is decreasing at night, thus making it hard to predict the blue line given by the red one. Additionally, the red line holds the traffic in the weekend but the blue one doesn't. Next, given that 99% of cells have at least three neighbor cells in our dataset, n is enlarged to 3 to calculate $H(X|X_1, X_2, X_3)$ using spatial dependence. Fig.9(a)(b)(c) illustrate the histograms of spatial entropy in all cells, indicating that data traffic owns the highest relevance to neighboring cells.

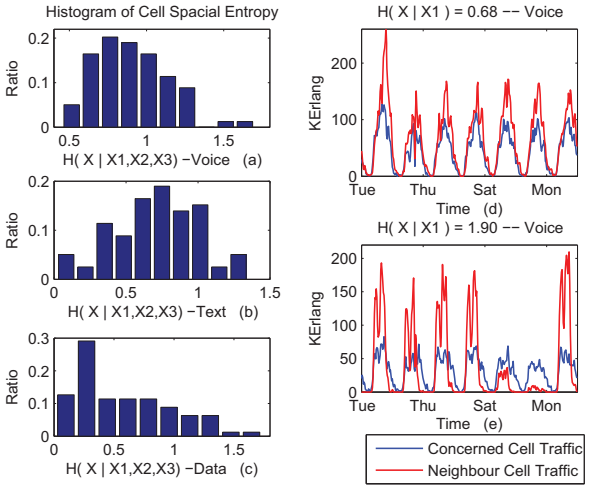

Fig. 9. (a)(b)(c): Histogram of cell spacial entropy conditioned by three adjacent cells. (d)(e): Sample traffic in high and low spatial entropy cell conditioned by one adjacency.

Besides, the CDFs of spatial entropy correspond to histograms in Fig.9 (a)(b)(c) is shown in Fig.10 (a). By the end, we analyze the impact of adjacent cell number to spatial entropy. Fig.10 (b) reveals that spatial entropy decreases evidently when the number of cells increases in voice and text services, but it has less impact on data traffic.

Findings: knowledge of adjacent cells traffic can enhance the predictability of voice and text more effectively than data.

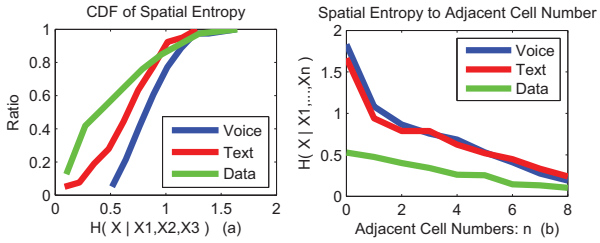

Fig. 10. (a): CDF of cell spacial entropy conditioned by three adjacent cells. (b): The variation of spatial entropy to adjacent cell number.

IV. Discussion

A. Quantization level

The less quantization levels are used, the more information is ignored from the dataset; however the more quantization levels the harder for us to guess the right level. Thus what is the suitable quantization level for traffic prediction? Fig.11 (a)(b) shows the CDF of random entropy and temporal entropy when quantization level varies from 4 to 16 on $T_p = 30$ minutes. In random entropy case, there is about 0.86 bit rise when level doubles, however in temporal case, entropy doesn't change as much as the random one, for the error of quantization has been dismissed in conditional entropy calculation. In Fig.11 (b), due to that traffic pattern is more clear in larger level case, temporal entropy decreases when level rises. In real BS, there are six frequency carriers at full load scenario [9], therefore six levels are enough to describe the working states of the BS, based on the precondition that frequency carriers can not be totally turned off. As a result, quantization level is set to six in most simulations to describe the traffic in this paper.

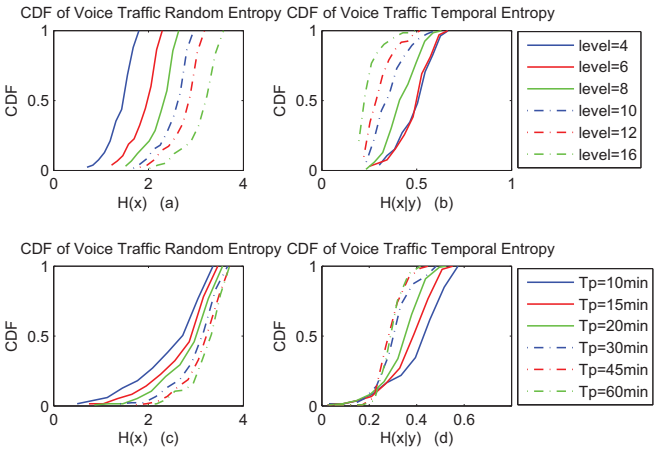

Fig. 11. (a)(b): Entropy variations to quantization levels. (c)(d): Entropy variations to the length of time period.

B. Time period

Small time period may magnify the volume of dataset, while large time period will ignore small scale fluctuations of the

traffic. Fig.11 (c)(d) illustrate the CDFs of random entropy and temporal entropy when time period varies from 10 to 60 minutes on $Q = 6$. We can see that random entropy increases with T_p, because traffic has a more uniform distribution when time period becomes large. Accordingly, temporal entropy decreases when time period rises, for there is less disturbance in traffic pattern recognition while T_p is large.

C. Geo-location analysis

In this part, we investigate the geo-locations of high and low entropy cells by means of BS deployment information in a $3km * 3km$ area which contains 133 cells, and then mark dedicated cells on it as in Fig.12. Cells are plotted to the figure if it has the top 3 high or low entropy of each category – random entropy, temporal entropy, service conditional entropy and spatial entropy. The traffic entropy here is not distinguished from service types, while we calculate average entropy of voice, text and data. After making field survey of the marked cells, we use Table.IV to conclude the geo-location rule of cell entropy. Moreover, we find that low random entropy cells usually have high service conditional entropy and vice versa; low temporal entropy cells sometimes have high spatial entropy and vice versa.

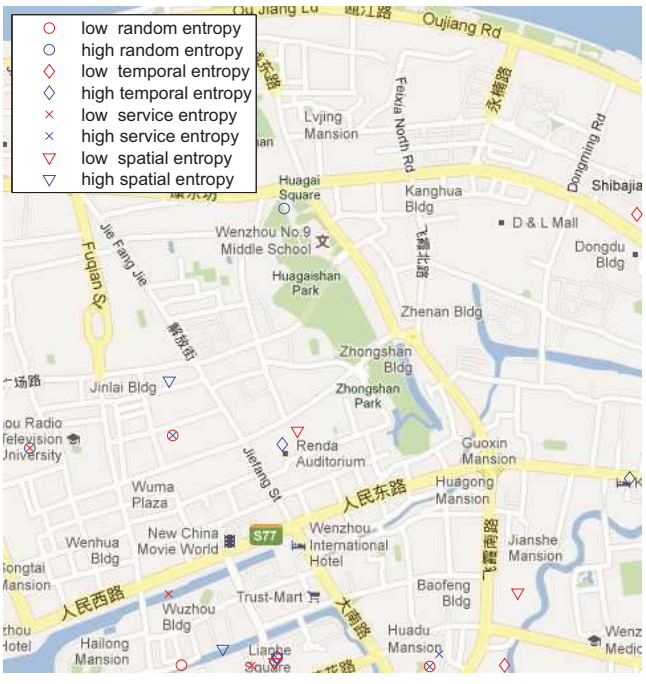

Fig. 12. Geo-location of high and low entropy cells.

V. Conclusion

In this paper, we categorize the traffic according to the type of service (i.e., voice, text or data) it belongs to. Afterwards, we investigate the traffic random entropy and the conditional entropy with temporal, service and spatial information a prior. Consequently, we summarize the predictability of traffic as follows: (1) voice traffic value is more uniform however data service has more low-valued traffic, and more than 70% cells have low random entropy of data traffic, (2) traffic at one moment can be well predicted when the preceding 15 hours traffic is known and voice service is the easiest to predict for its regularity, (3) voice traffic has so close similarity to text traffic in the same cell that we can use one to predict the other, however data traffic in more than 50 cells has no contribution to voice and text prediction, (4) knowledge of adjacent cells traffic can enhance the predictability of voice and text more effectively than data. In this paper, we also investigate parameters of traffic dataset processing and geo-location of cells with high and low entropy, and presents the geographic distribution view of traffic predictability. These conclusions demonstrate the upper bound of the traffic predictability, which provide guidance for us to design energy efficiency schemes and to manage network resources.

TABLE IV
The distribution of high and low entropy cells in Fig.12

Entropy category	Low entropy	High entropy
Random	Suburban buildings	Hotels and hospitals
Temporal	Schools and residential areas	Street market
Service	Business areas and hotels	Suburban buildings
Spatial	Business areas	Bars and hotels

Acknowledgment

This paper is partially supported by the National Basic Research Program of China (973 Program 2012CB316000) and the National Natural Science Foundation of China (NSFC) under grant number 61071130.

References

[1] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3G cellular data network," in *Proceedings of IEEE Infocom 2012*, Orlando, Florida, USA, Mar. 2012.

[2] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li, "Traffic-driven power saving in operational 3G cellular networks," in *Proceedings of MobiCom 2011*. New York, NY, USA: ACM, 2011, pp. 121–132.

[3] F. P. Tso, J. Teng, W. Jia, and D. Xuan, "Mobility: A double-edged sword for HSPA networks," in *Proceedings of IEEE MobiHoc 2010*, Chicago, Illinois, USA, Sep. 2010.

[4] R. Li, Z. Zhao, Y. Wei, X. Zhou, and H. Zhang, "GMPAB: A grid based energy saving scheme with predicted traffic load guidance for cellular networks," in *Proceedings of IEEE ICC 2012*, Ottawa, Canada, 2012.

[5] S. Chakraborty, Y. Dong, D. Yau, and J. Lui, "On the effectiveness of movement prediction to reduce energy consumption in wireless communication," *IEEE Transactions on Mobile Computing*, vol. 5, no. 2, pp. 157 – 169, feb. 2006.

[6] E. Oh and B.Krishnamachari, "Energy savings through dynamic base station switching in cellular wireless access networks," in *Proceedings of IEEE Globecom 2010*, Miami, Florida, USA, 2010.

[7] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[8] N. Duffield, J. Lewis, N. O'Connell, R. Russell, and F. Toomey, "Entropy of ATM traffic streams: a tool for estimating QoS parameters," *IEEE JSAC*, vol. 13, no. 6, pp. 981 –990, aug 1995.

[9] *Products and Application Scenarios of Radio Access Products*, http://www.huawei.com/.