# A New Method for Traffic Prediction in Emerging Mobile Networks

Yunjian Jia[1], Beili Wan[1], Liang Liang[1], Qian Zhao[1], Yu Zhang[1], and Liang Tang[2]
[1] College of Communication Engineering, Chongqing University, Chongqing, China
[2]China Mobile Group Chongqing Co., Ltd. Chongqing, China
Email: {yunjian, wanbeili, liangliang, zhaoqian, yzhang}@cqu.edu.cn; tangliang2@cq.chinamobile.com

***Abstract*—**With the increasing popularity of mobile devices and applications, emerging mobile network traffic exhibits special characteristics in temporal scale e.g., there is a scale variance between the network traffic on weekdays and on weekends. Although most existing methods have been applied to data traffic prediction, few of them take such characteristic into consideration. In this paper, by using real data in mobile networks, we adopt the entropy theory to reveal that the duration of time-series given for prediction doesn't always have a positive impact and that the uncorrelated preceding time-series also deteriorates the prediction accuracy. In view of this, partitioning the network traffic prediction into weekdays' and weekends' perspective, we propose a method to predict the data traffic. Finally, we evaluate the proposed method through predicting the data traffic for a future time according to the historical data traffic in a real mobile network. In comparison with the work based on ARMA (Auto Regressive Moving Average) method, our proposed method can reduce the Mean Absolute Percentage Error (MAPE) by 35.7% and 43.8% on weekdays' and weekends' prediction, respectively.

***Index Terms*—**Mobile networks, data traffic prediction, entropy theory, time-series, real data

## I. INTRODUCTION

With the rapid development of mobile networks, from the second generation (2G) to the third generation (3G) and the fourth generation (4G), the smart devices and mobile applications have been increasing rapidly. Toward the fifth generation (5G) of mobile networks, numerous devices will be joined in and the demand of network traffic will constantly rise. Therefore, precise traffic prediction is expected to ensure the normal operation of the network and achieve high resource utilization [1], [2]. When 3G starts, the popularity of services leads to dramatic changes in network traffic characteristics. Consequently, some methods of traffic prediction have been proposed in this era [3]-[8]. Authors in [3] took the notion of self-similarity that took place in 3G networks into account and proposed a method to estimate the main parameters of network traffic. In [4], the authors proposed an order-k Markov model to predict traffic patterns in 3G networks. Other methods, such as Auto Regressive Integrated Moving Average (ARIMA) [5], neural networks method [6], Kalman filtering method [7] and wavelet method [8] etc, have also been adopted to predict future traffic in 3G networks. However, the methods mentioned above usually predict the network traffic based on the preceding time-series and the prediction accuracy deteriorates quickly as the duration of time-series increases. In this respect, authors in [9]-[11] exploited the entropy theory to analyze the traffic predictability in mobile networks and concluded the appropriate duration of preceding time-series used for traffic prediction.

Along with the evolution of mobile networks, we find that the correlation between network traffic and social aspects in emerging mobile networks is more complex than that in traditional 2G or 3G networks, e.g., there is a scale variance between the network traffic on weekdays and network traffic on weekends with respect to user behaviors aspect. Therefore it is important to analyze this trend and exploit it to improve the performance of network traffic prediction.

Considering that the social behaviors of users have a great influence on the network traffic. Specifically, from social behavior aspect, users on weekends seem to spend more time on mobile devices and mobile applications than weekdays'. There is a scale variance between the data traffic on weekdays and data traffic on weekends. Therefore, in this paper, we take such characteristic into consideration, and adopt entropy theory to demonstrate that the prediction performance will be influenced by the combination of temporal duration. Partitioning the traffic prediction into weekdays' perspective and weekends' perspective, we propose a prediction method concentrating on combining the data from correlated days. Moreover, as the days on weekends are limited, we consider using a factor to compensate the variance of the data traffic on weekdays to improve the prediction on weekends. Numerical results show that our method can identify the optimal temporal combination as prior information and predict the objective data traffic accurately. Therefore, our work provides an essential understanding on traffic prediction in future networks.

The rest of this paper is organized as follows: Section II analyzes the network traffic predictability in real mobile networks. Section III reviews the model training.

Manuscript received June 5, 2015; revised December 7, 2015.
This work is supported by the National High-tech R&D Program of China (863 Program) under grant No. 2015AA01A706. This work is also sponsored in part by Hitachi, Ltd.
Corresponding author email: yunjian@cqu.edu.cn.
doi:10.12720/jcm.10.12.947-954

In Section IV, we introduce the data traffic prediction. Numerical results are presented in Section V, and Section VI gives the conclusions.

## II. Theoretical Analysis on Data Traffic Prediction

Information theory [12] has played a key part in illustrating the generality of information content. Entropy theory, as its basic idea, offers a precise definition of information content and gives an effective method to measure its uncertainty. In this section, we adopt the entropy approach to gauge the data traffic predictability according to certain prior information from historical data.

### A. Data Collection

We collect the data traffic from mobile networks in China Mobile Group Chongqing Co., Ltd., China. The data set covers a large area of about 82403 km2 and it covers more than one million mobile users. Fig. 1 shows the architecture of data collection. The following list shows the elements from the leaves to the root of the tree: nodes, core equipments, net stream, network management system, optical splitter and DPI (Deep Packet Inspection) analysis system. Each node, referring to a computer or other devices, connects to the core equipment through communication link. The data traffic is stored in net stream. Network management system is used for resource monitoring and analysis. During the transmission, a part of data traffic is copied for analysis by optical splitter. DPI analysis system works as a filtering to collect and analysis the statistical information and our data traffic is collected from this system.

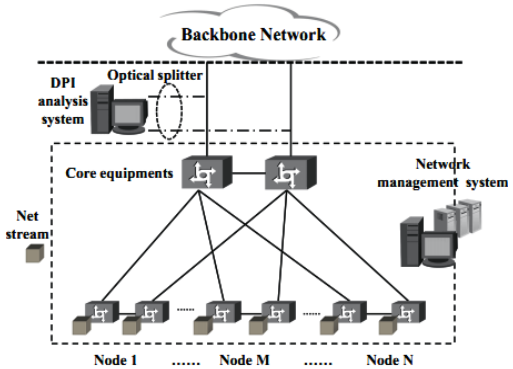


Fig. 1. The architecture of data collection in China Mobile Group Chongqing Co., Ltd.

### B. Theoretical Analysis on Prediction

The collected dataset includes all types of applications (HTTP, P2P, IM etc.) in both rural and urban areas in Chongqing, China. Each hour records the volume of total applications. After obtaining the dataset, a processing procedure is conducted to normalize the data traffic with the following equation:

$$Nor\_Data(t) = \frac{actualData(t) - \min Data}{\max Data - \min Data} \quad (1)$$

where *actualData*(*t*) denotes the collected dataset, and *Nor_Data(t)* denotes the normalized dataset. *maxData* and *minData* are the maximal and minimal data traffic values of the dataset, respectively.

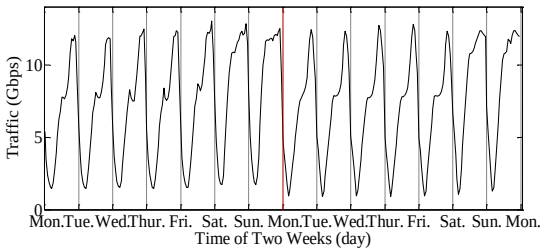


Fig. 2. Data traffic for two weeks.

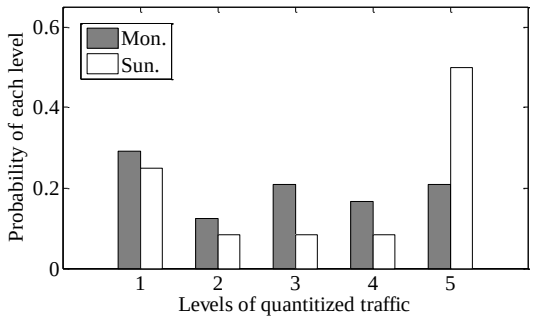


Fig. 3. The probability of each level on Monday and Sunday.

To ease the following analysis, the data during a certain period *i* is quantized into 5 levels, which represent five different volume of data traffic, from the lowest to the highest one. Thus, with the normalized dataset, the corresponding data traffic probabilities can be obtained. For example, Fig. 2 shows the data traffic two weeks. In this figure, a clear diurnal pattern of data traffic is illustrated. For all days, there is a scale variance between weekdays and weekends. However, there is little scale variance among days in weekdays and days in weekends respectively. This kind of pattern can be explained that user behaviors have relationship to data traffic. Fig. 3 depicts the probability of each level on Monday and Sunday in one week. The entropy is employed to measure the uncertainty of events [12]. Let $X$ be a discrete random variable with possible values $\{x_1, ..., x_n\}$ and the corresponding probability is defined as

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \quad (2)$$

where $b$ is the base of the logarithm. Unless otherwise specified, we commonly take the logarithm to base 2. According to the theory of entropy, the data traffic probability heavily depends on the data traffic characteristics. For example, the data traffic in each week which is shown in Fig. 2 would have a scale variance between weekdays and weekends. From social behavior perspective, users in weekends would spend more time on mobile devices and mobile applications than weekdays'. Given that the data traffic demand is highly linked to user behaviors. That's to say, the variety of entropy values is also more or less related to user behaviors. Therefore, we use entropy theory to describe the uncertainty of data traffic in mobile networks. In addition, the accuracy of

data traffic prediction relies on the adopted model, but also requires a certain quantity of prior information to reduce uncertainty [12]. In other words, prediction performance will be improved with the increase of the amount of prior information. Here, we take the conditional entropy of two random variables $X=\{x_1,x_2,\cdots x_n\}$ and $Y=\{y_1,y_2,\cdots y_n\}$ into consideration. The conditional entropy $H(X|Y)$ is defined as

$$H(X|Y)=\sum_{i,j}p(x_i,y_j)\log\frac{p(y_j)}{p(x_i,y_j)} \quad (3)$$

Taking the network traffic on Monday and on Sunday for example, we calculate the entropy values of different conditions. As indicated in Table I, among the dataset, both the conditional entropies of Monday and Sunday network traffic decrease rapidly, even though the random entropy of Monday or Sunday is comparatively larger. Moreover, for Monday, the conditional entropy with network traffic on weekdays (e.g., the present Monday, or the present Monday and Tuesday) decreases more rapidly than that on weekends (e.g., the present Saturday). For Sunday, the conditional entropy with network traffic on weekends (e.g., the present Sunday, or the present Saturday and Sunday,) decreases more rapidly than that on weekdays (e.g., the present Monday). This finding is also applied to other days. Therefore, we conclude that, the different combinations of time-series have different impacts on traffic predictability. In particular, the preceding time-series on weekends has little contribution to the performance of weekdays' prediction, and vice versa. In the next section, we present a method to determine the time-series combination from both weekdays' perspective and weekends' perspective.

TABLE I: CORRELATIONS BETWEEN SEVEN DAYS AFTER COMPENSATION

| Day | Calculation conditioned on | Entropy |
|---|---|---|
| **Next Mon.** | None | 2.2672 |
| | The present Mon. | 0.0254 |
| | The present Sat. | 0.0712 |
| | The present Mon. and Tue. | 0.0308 |
| | The present Mon. Tue. and Wed. | 0.0010 |
| | The present Mon. Tue. and Sat. | 0.0311 |
| **Next Sun.** | None | 1.8962 |
| | The present Mon. | 0.2230 |
| | The present Sun. | 0.0760 |
| | The present Mon. and Sun. | 0.0729 |
| | The present Sat. and Sun. | 0.0199 |
| | The present Mon. Sat. and Sun. | 0.0438 |

## III. MODEL TRAINING

As network traffic is generally continuous, we treat the volume of network traffic in time interval [$t$, $t+\Delta t'$] as an entry in the method. Therefore, $n$ length sequence is obtained to indicate the volume of network traffic in $n$ intervals for the given week.

We use the data from the first week to training the prediction model, than we adopt the training model to predict the traffic in next week. In this section, we mainly introduce the model training of our method.

### A. Weekdays' Perspective

As for the variation of data traffic on weekdays, we use the combination of previous days as the prior information to predict the data traffic of days in next weekdays, as shown in Fig. 4 (a). First, we use the temporal combination of correlated days in a week to describe the day (e.g., Monday) we want to predict in the next week. The temporal combination of correlation can be built up as equation (4).

$$y_{weekdays}(t)=\sum_{i=1}^{m}w_i f_i(t) \quad (4)$$

where $f_i(t)$ ($t$=1,2,...,$n$. $i$=1,2,...,$m$) reflects the combination sequence $i$ on interval $t$. $y_{weekdays}(t)$ ($t$=1,2,$\cdots$, $n$) denotes the combination of $f_i(t)$. The combination weight vector for m combination sequences is $w=(w_1,w_2,\cdots w_m)^T$, and it satisfies the following conditions:

$$e^T w=1 \quad (5)$$

$$w\geq 0 \quad (6)$$

where $e^T$=(1,1,...,1).

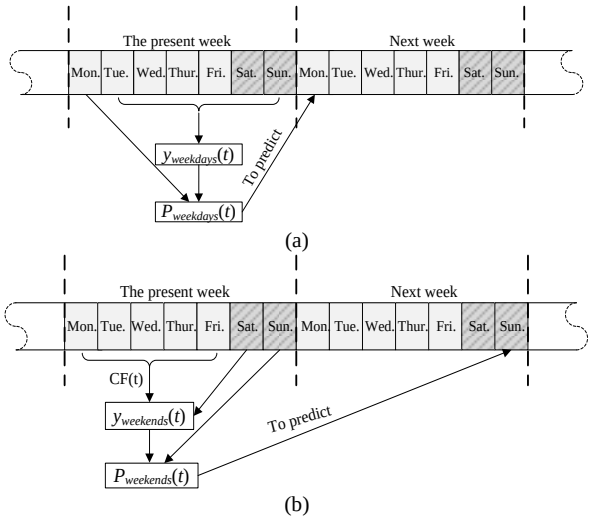


Fig. 4. (a) The flow diagram of the prediction method on weekdays' perspective (i.e., for the data traffic prediction on Monday). (b) The flow diagram of the prediction method on weekends' perspective (i.e., for the data traffic prediction on Sunday).

### B. Weekends' Perspective

Considering two days on weekends for a week, we propose a factor to compensate the previous data traffic on weekdays. After the reconstruction of the data, we use the combination of reconstructed data to forecast the day (e.g., Sunday) in next weekends, as Fig. 4 (b) shown. The model can be built up by the following steps.

*Step one*: we propose a factor to compensate the small scale variance between weekdays and weekends. The factor in each interval can be calculated by equations (7) and (8).

$$p(t) = \frac{f_j(t)}{f_s(t)} \tag{7}$$

where $f_s(t)$ ($t$=1,2,...,$n$) is the sequence of the predicted day on weekends, and $f_j(t)$ ($j$=1,2,...,5. $t$=1,2,...,$n$ ) is the combination sequence $j$ on weekdays (5 days). $p(t)$ is the ratio of $f_j(t)$ and $f_s(t)$. $CF(t)$ is the compensate factor on interval $t$.

$$CF(t) = \frac{1}{5}(p(1) + p(2) + ... + p(n)) \tag{8}$$

*Step two*: we use the compensate factor to modify the data traffic on weekdays. The new sequence on weekdays is obtained by equation (9).

$$f_{new\_j}(t) = CF(t) \cdot f_j(t) \tag{9}$$

where $f_{new\_j}(t)$ indicates the new combination sequence compensated by factor $CF(t)$. Then, we use the new sequences to update the prior data on weekdays.

*Step three*: we use compensated sequences to predict the data traffic on weekends as described by equation (10).

$$y_{weekends}(t) = \sum_{j=1}^{m} w_j f_{new\_j}(t) \tag{10}$$

where $y_{weekends}(t)$ ($t$=1, 2,..., $n$) is the combination of data traffic from correlated days that has been compensated.

In order to get the best effect of combination, we propose the optimization model that is described below. By solving the model, we could get the optimal combination weight vector.

For weekdays:

$$\max R_{weekdays} = \frac{\sum_{t=1}^{n}(f_k(t) - f_A)\cdot(y_{weekdays}(t) - y_A)}{\sqrt{\sum_{t=1}^{n}(f_k(t) - f_A)} \cdot \sqrt{\sum_{t=1}^{n}(y_{weekdays}(t) - y_A)}} \tag{11}$$

$$s.t. \begin{cases} e^T w = 1 \\ w \geq 0 \end{cases} \tag{12}$$

For weekends:

$$\max R_{weekends} = \frac{\sum_{t=1}^{n}(f_s(t) - f_B)\cdot(y_{weekends}(t) - y_B)}{\sqrt{\sum_{t=1}^{n}(f_s(t) - f_B)} \cdot \sqrt{\sum_{t=1}^{n}(y_{weekends}(t) - y_B)}} \tag{13}$$

$$s.t. \begin{cases} e^T w = 1 \\ w \geq 0 \end{cases} \tag{14}$$

where $f_k(t)$ and $f_s(t)$ ($t$=1,2,...$n$) are the corresponding data traffic of the predicted day in last weekdays and weekends, respectively. $f_A$ is the average value of $f_k(t)$ and $f_B$ denotes the average value of $f_s(t)$ ($t$=1,2,...$n$). $y_A$ and $y_B$ are the average values of $y_{weekdays}(t)$ and $y_{weekends}(t)$, respectively. $R$ is adopted to weigh the correlation between two variables. In this work, we use $R_{weekdays}$ to weigh the correlation between variable $f_k(t)$ and variable $y_{weekdays}(t)$ and $R_{weekends}$ to weigh the correlation between variable $f_s(t)$ and variable $y_{weekends}(t)$. When $R > 0$, it indicates that the two variables are positively related. When $R < 0$, it indicates that the two variables are negatively related. When $R = 0$, it indicates that there is no linear correlation between the two variables. When $R = 1$, it indicates that there is total linear correlation between the two variables, that is, the function relationship. When $R > 0$ and $0 < R < 1$, there is a linear correlation between the two variables. And if $R$ is closer to 1, the linear relationship between the two variables is strong close; if $R$ is closer to 0, the linear correlation between the two variables is weak close.

After determining the optimal combination weight vector, we use the equations (15) and (16) to get the predicted value.

$$P_{weekdays}(t) = \frac{1}{2}(y_{weekdays}(t) + f_k(t)) \tag{15}$$

$$P_{weekends}(t) = \frac{1}{2}(y_{weekends}(t) + f_s(t)) \tag{16}$$

$P_{weekdays}(t)$ and $P_{weekends}(t)$ denote the predicted data traffic value on weekdays and weekends, respectively.

## IV. Mobile Data Traffic Prediction

In this paper, we consider the data traffic in one week, and denote the time points of a day by [0, 1, 2,..., 23]. So there are $24 \times 27$ time points. As is explained in Section II, the value of n is 24 and the value of m is 6. Here we analyze the optimal combination of data traffic sequences on weekdays and weekends, respectively.

### A. Prediction on Weekdays

We assume that the data traffic on next Monday is to be predicted. The combination of data traffic in this week except for Monday is used to predict it. According to the prediction model mentioned above, we should find the optimal number of combination and proper weight vector.

In order to get the optimal number of combination, we consider the possible combination from 1 to 6, according to the degree of correlation between other 6 days. First, we calculate the degree of correlation between 7 days. As is depicted in Table II, the data traffic on Monday has a high degree of correlation with that on weekdays. Then, we take 6 trials to get the probable prediction of Monday in next week. For each trial we take the data traffic from

the days that are most correlated to Monday into consideration. The composition of combination for each trial is listed as follows:

Trial #1: {Tue.};
Trial #2: {Tue., Thur.};
Trial #3: {Tue., Wed., Thur.};
Trial #4: {Tue., Wed., Thur., Fri.};
Trial #5: {Tue., Wed., Thur., Fri., Sat.};
Trial #6: {Tue., Wed., Thur., Fri., Sat., Sun.}.

After the determination of the composition, we take the data traffic on Monday as the actual data $f_k(t)$, and the combination of data traffic from correlated days denotes as $y_{weekdays}(t)$. Then we calculate the best combination weight factor for each trial through equation (11) and (12). Table III depicts the best weight factor for each trial.

TABLE II: CORRELATIONS BETWEEN SEVEN DAYS

| Correlations | Mon. | Tue. | Wed. | Thur. | Fri. | Sat. | Sun. |
|---|---|---|---|---|---|---|---|
| **Mon.** | 1 | 0.997 | 0.996 | 0.997 | 0.992 | 0.956 | 0.931 |
| **Tue.** | 0.997 | 1 | 0.998 | 0.999 | 0.992 | 0.962 | 0.939 |
| **Wed.** | 0.996 | 0.998 | 1 | 0.999 | 0.993 | 0.953 | 0.924 |
| **Thur.** | 0.997 | 0.999 | 0.999 | 1 | 0.992 | 0.955 | 0.929 |
| **Fri.** | 0.992 | 0.992 | 0.993 | 0.992 | 1 | 0.969 | 0.934 |
| **Sat.** | 0.956 | 0.962 | 0.953 | 0.955 | 0.969 | 1 | 0.988 |
| **Sun.** | 0.931 | 0.939 | 0.924 | 0.929 | 0.934 | 0.988 | 1 |

TABLE III: THE BEST WEIGHT FACTOR FOR EACH TRIAL ON MONDAY PREDICTION

| Factor value | $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ |
|---|---|---|---|---|---|---|
| **Trial #1** | 1 | Null | Null | Null | Null | Null |
| **Trial #2** | 0.4998 | 0.5002 | Null | Null | Null | Null |
| **Trial #3** | 0.3333 | 0.3332 | 0.3335 | Null | Null | Null |
| **Trial #4** | 0.2679 | 0.0001 | 0.5972 | 0.1348 | Null | Null |
| **Trial #5** | 0.2859 | 0.2335 | 0.3872 | 0.0933 | 0.0001 | Null |
| **Trial #6** | 0.2182 | 0.0001 | 0.6457 | 0.1351 | 0.0001 | 0.0008 |

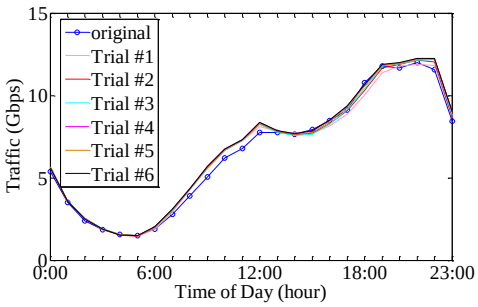


Fig. 5. Prediction of 6 trials on Monday.

With the obtained weight factor, we can easily calculate the predicted data on Monday in the next week by equations (15). Fig. 5 shows the prediction of 6 trials. As it demonstrates, the curves for 6 trials fit well with the original one.

### B. Prediction on Weekends

We first analyze the new degree of correlation with the data traffic on weekdays which is updated by equation (9). As is depicted in Table IV, the correlation between Sunday and weekdays in a week has been much improved. According to the correlation, we listed the following 6 trials:

Trial #1: {Tue.};
Trial #2: {Tue., Wed.};
Trial #3: {Tue., Wed., Thur.};
Trial #4: {Mon., Tue., Wed., Thur.};
Trial #5: {Mon., Tue., Wed., Thur., Fri.};
Trial #6: {Mon., Tue., Wed., Thur., Fri., Sat.}.

After that, we take the data traffic on Sunday as the actual data $f_s(t)$, and let the combination of data traffic from correlated days be $y_{weekends}(t)$. The best combination weight factor for each trial can be calculated by equation (13) and (14). Then, we use equation (16) to calculate the predicted data on Sunday in the next week with the obtained weight factor listed in Table V. Fig. 6 shows the prediction of 6 trials and it illustrates that the predicted values of 6 trials close to the original data on Sunday.

TABLE IV: CORRELATIONS BETWEEN SEVEN DAYS AFTER COMPENSATION

| Correlations | Mon. | Tue. | Wed. | Thur. | Fri. | Sat. | Sun. |
|---|---|---|---|---|---|---|---|
| **Mon.** | 1 | 0.995 | 0.995 | 0.995 | 0.992 | 0.985 | 0.998 |
| **Tue.** | 0.995 | 1 | 0.998 | 0.998 | 0.991 | 0.982 | 0.999 |
| **Wed.** | 0.995 | 0.998 | 1 | 0.999 | 0.993 | 0.989 | 0.999 |
| **Thur.** | 0.995 | 0.998 | 0.999 | 1 | 0.993 | 0.989 | 0.999 |
| **Fri.** | 0.992 | 0.991 | 0.993 | 0.993 | 1 | 0.992 | 0.995 |
| **Sat.** | 0.985 | 0.982 | 0.989 | 0.989 | 0.992 | 1 | 0.988 |
| **Sun.** | 0.998 | 0.999 | 0.999 | 0.999 | 0.995 | 0.988 | 1 |

TABLE V: THE BEST WEIGHT FACTOR FOR EACH TRIAL ON SUNDAY PREDICTION

| Factor value | $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ |
|---|---|---|---|---|---|---|
| **Trial #1** | 1 | Null | Null | Null | Null | Null |
| **Trial #2** | 0.4999 | 0.5001 | Null | Null | Null | Null |
| **Trial #3** | 0.3333 | 0.3333 | 0.3334 | Null | Null | Null |
| **Trial #4** | 0.2904 | 0.1603 | 0.2257 | 0.3236 | Null | Null |
| **Trial #5** | 0.1569 | 0.2575 | 0.1467 | 0.2596 | 0.1793 | Null |
| **Trial #6** | 0.214 | 0.2355 | 0.1884 | 0.2128 | 0.1257 | 0.0236 |

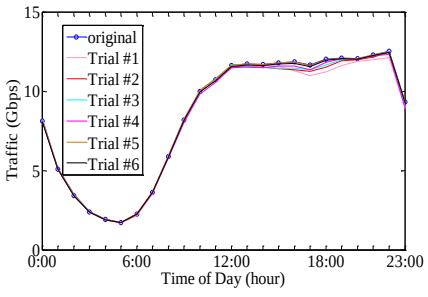


Fig. 6. Prediction of 6 trials on Sunday

## V. NUMERICAL RESULTS

In this section, we analyze the performance of our prediction. To begin with, we use the predicted data minus the actual data, as equation (17) shows:

$$e(t) = predictedData(t) - actualData(t) \quad (17)$$

where $e(t)$ ($t$=0, 2,..., 23) is the error of prediction. Then, the curve of prediction error is obtained. Fig. 7 depicts the curves of prediction error of 6 trials on Monday, and Fig. 8 depicts the curves of prediction error of 6 trials on Sunday. Fig. 7 and Fig. 8 show the prediction errors of 6 trials are obviously different.

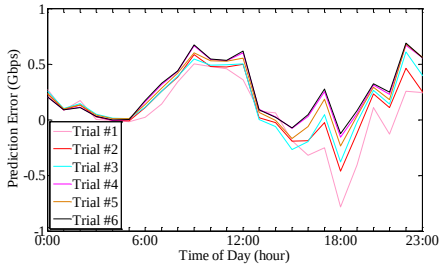


Fig. 7. Prediction error of 6 trials on Monday.

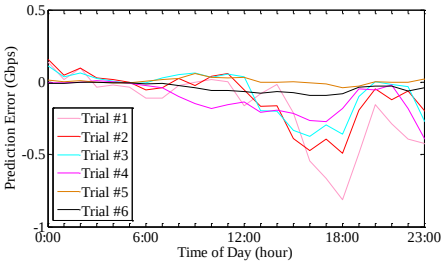


Fig. 8. Prediction error of 6 trials on Sunday

To evaluate the accuracy of each trial, we consider the mean absolute percentage error (*MAPE*) as the performance index. The indexes of weekdays and weekends are given by equation (18) and (19), respectively.

$$MAPE_{weekdays} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{f_k(t)-P_{weekdays}(t)}{f_k(t)}\right|, t=1,2\cdots n \quad (18)$$

$$MAPE_{weekends} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{f_s(t)-P_{weekends}(t)}{f_s(t)}\right|, t=1,2\cdots n \quad (19)$$

As is explained in Section IV, the value of n is 24 and there are 24 time points in our prediction. We can get the MAPE for 6 trials by equation (18) or equation (19). Fig. 9 and Fig. 10 depict the MAPE of each trial on Monday and Sunday prediction, respectively. In view of this, we conclude that the optimal combination as prior information on weekdays' prediction is the two days that most correlated to the predicted day in the present weekdays. On weekends' prediction, the optimal combination is the updated 5 days that most correlated to the predicted day in the present weekends. Based on that finding, we can conduct the prediction in a week. Considering ARMA method have good performance in prediction up to date [13], we compare our method with ARMA method on the prediction in a week, as shown in Fig. 11. Fig. 12 shows the comparison of our method and ARMA method on Monday and Sunday prediction. For the large number of vertical coordinates, the differences between predicted data and actual data have been confused. To better illustrate the performance of prediction, we amplify the local part of the figure. From Fig. 12, we can clearly see that our method has better performance than conventional ARMA method. Finally, we calculate the MAPE of the methods mentioned above on the prediction in a week. Table VI depicts, our method can reduce the MAPE by 35.7% and 43.8% on weekdays' and weekends' prediction compared with ARMA method.

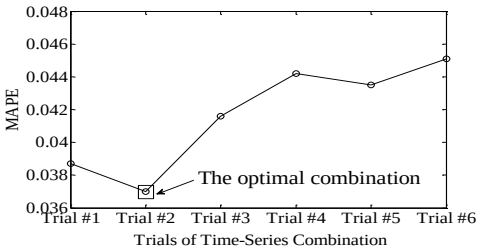


Fig. 9. MAPE of 6 trials on Monday.

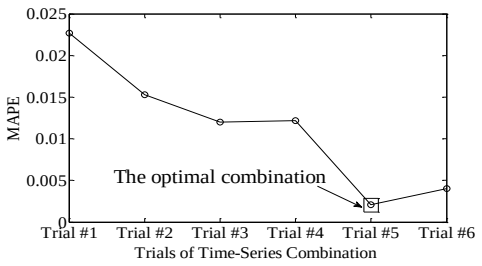


Fig. 10. MAPE of 6 trials on Sunday.

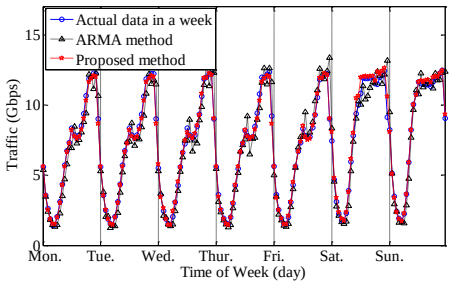


Fig. 11. The comparison of our method and ARMA method on the prediction of the traffic in a week.

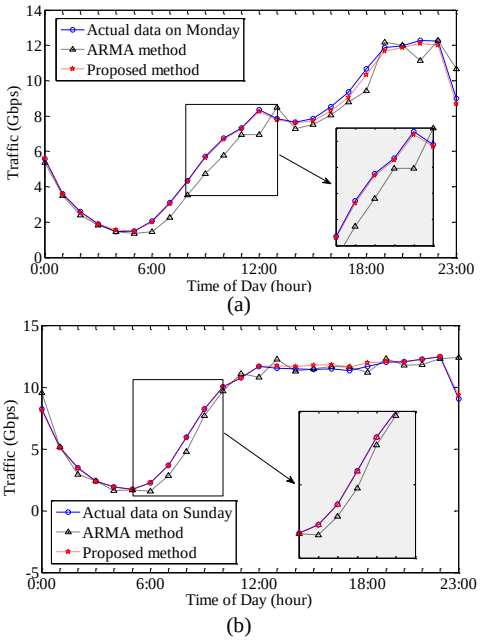


Fig. 12. The comparison of our method and ARMA method: (a) on Monday prediction, and (b) on Sunday prediction.

TABLE VI: THE MAPE OF OUR METHOD AND ARMA METHOD

| *MAPE* | Proposed method | ARMA method |
|---|---|---|
| Weekdays | 0.0538 | 0.0842 |
| Weekends | 0.0473 | 0.0861 |

## VI. CONCLUSIONS

This paper has presented a novel data traffic prediction method from social behavior aspect. The proposed method considers the characteristics of temporal variation from data traffic in a week. To detect the distinct characteristics, we have adopted entropy theory to analyze the effect of different temporal preceding duration on data traffic prediction. The study shows that the preceding duration in weekends has little contribution to the performance of weekdays' prediction, and the preceding duration in weekdays leads to the same result on weekends' prediction. Therefore, our method has partitioned the data traffic prediction into weekdays' and weekends' perspective and focused on the combination of data from correlated days. In addition, we have conducted extensive simulation experiments based on the actual data traffic in previous week to predict the data traffic in future week. Numerical results depict that our method can identify the optimal temporal combination as prior information for prediction, and provide accurate data traffic prediction afterwards.

## ACKNOWLEDGMENT

This work is supported by the National High-tech R&D Program of China (863 Program) under grant No. 2015AA01A706. This work is also sponsored in part by Hitachi, Ltd.

## REFERENCES

[1] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, *et al*., "5G on the Horizon: Key Challenges for the Radio-Access Network," *IEEE Vehicular Technology Magazine*, vol. 8, no. 3, pp. 47-53, 2013.

[2] R. Sattiraju and H. D. Schotten, "Reliability Modeling, Analysis and Prediction of Wireless Mobile Communications," in *Proc. IEEE VTC 2014 Conference*, Seoul, Korea, May 18-21, 2014.

[3] A. Krendzel, Y. Koucheryavy, J. Harju, and S. Lopatin, "Method for Estimating Parameterers of 3G Data Traffic," in *Proc. IEEE ICC*, Paris, France, June 20-24, 2004.

[4] K. Zhang and L. Cuthbert, "Traffic Pattern Prediction in Cellular Networks," in *Proc. IEEE ICCS*, Guangzhou, China, November 19-21, 2008.

[5] Y. Shu, M. Yu, J. Liu, and O. Yang, "Wireless Traffic Modeling and Prediction Using Seasonal ARIMA Models," in *Proc. IEEE ICC Conference*, Anchorage, Alaska, May 11-15,2003.

[6] L. Zhu, L. Qin, K. Xue, and X. Zhang, "A novel BP neural network model for traffic prediction of next generation network," in *Proc. IEEE ICNC Conference*, Tianjin, China, Augest 14-16, 2009.

[7] M. Achir, Y. M. Ghamri-Doudane, and G. Pujolle, "Predictive resource allocation in cellular networks using kalman filters," in *Proc. IEEE ICC Conference*, Anchorage, Alaska, May 11-15, 2003.

[8] X. Wang, Y. Ren, and X. Shan, "WDRLS: A wavelet-based on-line predictor for network traffic," in *Proc. IEEE GLOBECOM Conference*, San Francisco, USA, December 1-5, 2003.

[9] R. Li, Z. Zhao, X. Zhou, and H. Zhang, "Energy saving scheme in radio access networks via compressive sensing-based traffic load prediction," *Transactions on Emerging Telecommunications Technologies*, vol. 25, no. 4, pp. 25-29, Apr. 2014.

[10] X. Zhou, Z. Zhao, R. Li, Y. Zhou, and H. Zhang, "The predictability of celluler networks traffic," in *Proc. IEEE ISCIT*, Gold Coast, Australia, October 4, 2012.

[11] R. Li, Z. Zhao, X. Zhou, J. Palicot, and H. Zhang, "The prediction analysis of cellular radio access network traffic: From entropy theory to networking practice," *IEEE Comuunications Magazine*, vol. 52, no. 6, Jun. 2014.

[12] T. M. Cover and J. A. Thomas. (2006). Elements of Information Theory, Wiley-Interscience. [Online]. Available: http://www.amazon.com/Elements-Information-Theory-Telecommunications-Processing/dp/0471062596

[13] M. Wei and K. Kim, "Intrusion detection scheme using traffic prediction for wireless industrial networks," *Journal of Communications and Networks*, vol. 14, no. 3, pp. 310-318, 2012.

**Yunjian Jia** received his B.S. degree from Nankai University, China, and his M.E. and Ph.D. degrees in Engineering from Osaka University, Japan, in 1999, 2003 and 2006, respectively. From 2006 to 2012, he was with Central Research Laboratory, Hitachi, Ltd., where he engaged in research and development on wireless networks, and also contributed to LTE/LTE-Advanced standardization in 3GPP. He is now a professor at the College of Communication Engineering, Chongqing University, Chongqing, China. He is the author of more than 60 published papers, and 20 granted patents. His research interests include radio access technologies, mobile networks, and IoT. Dr. Jia has won several prizes from industry and academia including the IEEE Vehicular Technology Society Young Researcher Encouragement Award, the IEICE Paper Award, the Yokosuka Research Park R&D Committee YRP Award, and the Top 50 Young Inventors of Hitachi. Moreover, he was a research fellowship award recipient of International Communication Foundation in 2004, and Telecommunications Advancement Foundation Japan in 2005.

**Beili Wan** received her B.E. degree in School of Information Engineering from Southwest University of Science and Technology (SWUST), China, in 2012. She is currently working toward her M.E. degree in Information and Communication Engineering in Chongqing University. Her research interests include traffic model and traffic prediction for mobile networks, with emphasis on emerging mobile networks.

**Liang Liang** received her B.E. and M.E. degrees from the Southwest University of Science and Technology (SWUST), China, in 2003 and 2006, respectively, and the Ph.D. degree in communication and information system from the University of Electronic Science and Technology of China (UESTC) in 2012. She is currently a lecturer in College of Communication Engineering, Chongqing University, Chongqing, China. Her research interests include wireless communication and optimization, green radio, and wireless sensor networks.

**Qian Zhao** received her B.E degree in Communication Engineering from Qufu Normal University, China, in 2013. She is currently working toward her M.E degree in Electronic and Communication Engineering from Chongqing University. She mainly engages in the Internet of Things.

**Yu Zhang** received her B.E. degree in Communication Engineering from Chongqing University, China, in 2014. She is currently working toward her M.E. degree in Information and Communication Engineering in the same university. Her research interests include radio resource management and optimization for mobile networks, with emphasis on quality-of-experience provision.

**Liang Tang** received his M.E degree from Chongqing University, China, in 2009. He is currently working toward his Ph.D. degree in Information and Communication Engineering in the same university. Moreover, his is a director in data bearing network center, China Mobile Group Chongqing Co., Ltd., Chongqing, China. He mainly engages in wireless communication and radio resource management for mobile network.