Future Generation Computer Systems 37 (2014) 429–437

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Analysis of user behaviors by mining large network data sets

Zhenhua Wang [a], Lai Tu [a,*], Zhe Guo [b], Laurence T. Yang [c], Benxiong Huang [a]

[a] Department of EIE., Huazhong University of Science and Technology, Luoyu Rd. 1037, Wuhan, China

[b] Controller Technology Development Department, Shanghai Institute of Huawei Technologies Co.Ltd, XinJinQiao Rd. 2222, Pudong New District, Shanghai, China

[c] School of Computer Science and Technology, Huazhong University of Science and Technology, Luoyu Rd. 1037, Wuhan, China

HIGHLIGHTS

- We process 16-week-long CDR data of one million users with fuzzy clustering.
- We find some relations between ARPU level and users behavior patterns.
- We prove that the mobility of users is related to ARPU and communication behavior.
- Results indicate that the top ARPU level users are the most "lonely" ones.

ARTICLE INFO

Article history:
Received 29 March 2013
Received in revised form
16 January 2014
Accepted 17 February 2014
Available online 26 February 2014

Keywords:
User behavior
Data mining
Customer segmentation
Fuzzy *c*-means clustering

ABSTRACT

Understanding the intelligence of human behaviors by mining petabytes of network data represents the tendency in social behaviors research and shows great significance on Internet application designing and service expansion. Meanwhile, the running mobile networks that generate huge data can be the best social sensor for these studies. This paper investigates a practical case of mobile network aided social sensing which uncovers some features of users' behaviors in mobile networks by intelligently processing the big data. The paper studies the users' behaviors with regard to communication, movement, and consumption based on large user data sets. The main contribution of the study is some findings on the relations among these behavior features. We find that the users' calling behaviors are different despite their monthly expenditures being similar, though different consumption level users may have similar communication behaviors. We also find that statistically users with the higher mobility contribute more ARPU than those with lower mobility. Additionally, we also find that the top consumption level users are the most "lonely" ones by exploring the movement clustering patterns of users. These findings are significant to instruct marketing strategies for telecommunication industry.

1. Introduction

Although technologies have been pushing our world and society to a smarter one, human behaviors in the society still keep some inherent features and complexities that are hard to explain. Understanding the intelligence of human behaviors in the real world has great significance in practical application, such as mobile network deployment, traffic engineering, urban planning and service recommendation. While studies on human behaviors were not new in social science, quantitative analyses were not common due to the lack of source of data. Thanks to the computers and networks as they can now give plenty of computational ways of collecting and analyzing data for social studies, which used to depend on surveys in traditional methodology. Thus in the new era of "Big Data", never before have researchers had the opportunity to mine such a wealth of information that promises to provide insights about the complex behaviors of human societies [1,2]. One goal of these researches is to quantitatively uncover the inherent feature of human behaviors and track how our behaviors evolve by mining petabytes of network data.

The studies on human behaviors can be traced since 19th century [3]. The discipline covers psychology, sociology, anthropology, etc., which study different aspects of the nature of human intelligence. However, due to lack of ways of measurement and analysis of large scale of data, the studies mostly focused on individuals or a certain small group or rough qualitative estimation of social behaviors. This was kept almost unchanged until the emergence of modern computing science and network technology. Recent

* Corresponding author. Tel.: +86 18086484600.
E-mail addresses: tulai.net@gmail.com, tulai@hust.edu.cn (L. Tu).

http://dx.doi.org/10.1016/j.future.2014.02.015
0167-739X/

advances in computing and network science have driven the studies on social behaviors to a newly high stage. The universal usage of computational devices leaves huge amount of data that are tightly related with human behaviors. Scientists were able to use computer data, and network data to sense and analyze human behaviors in the society. During this period, a lot of theoretical and practical achievements in social behaviors studies arose. The prevalence of social network in the Internet and the progress in complex network research are best examples for this [4,5]. Meanwhile, with the development and widespread of smartphones and mobile network, they began to show their dominance in sensing human behaviors over the traditional way of mining Internet social media. Therefore, in this paper, we investigate a practical case of mobile network aided social sensing. We study the users' behaviors with regard to communication, movement, and consumption based on mining a large set of mobile user data. The findings may be significant to instruct marketing strategies for the service providers to increase their revenues and lead them to success through Customer Relationship Management (CRM) [6].

When discussing the communication behaviors, movement behaviors and consumption behaviors of mobile users, we aim at three questions:

1. Does each ARPU (Average Revenue Per User) level imply similar communication behavior?
2. Does each ARPU level imply similar mobility level? And what is the relationship between a user's consumption capacity and mobility?
3. Do people with similar mobility patterns have similar communication behaviors?

We explore the above three issues in this paper. We notify that the detailed information about human mobility across a large population can be collected by mobile operators that record the closest base station when a call is generated. Herein, we use 16-week-long (from September to December, 2009) calling records of about one million users in a metropolis in China as our data set to conduct research. Each record includes the serving base station's ID, the start and end times of each phone call (outgoing and incoming calls are distinguished), as well as the monthly billing information of each user. For preserving the users' privacy, the record of each phone call is anonymous.

We apply both the value-based and behavior-based segmentation methods to investigate the difference between the users who have different or similar ARPU. That is, we firstly divide the users based on their ARPU into different groups as preliminary division. Then, we study their behaviors by using fuzzy c-means (FCM) clustering. The contributions of our paper are:

- Verify that the same ARPU level users have different behavior patterns, while different ARPU level users may have similar communication behaviors;
- Prove that the users' mobility levels have relation with their ARPUs and communication behaviors. People who make less calling or like making calls at night have less mobility than others;
- Find that the top ARPU level users are the most "lonely" ones, which imply that they are willing to move alone and do not like being with others.

The rest of paper is organized as follows. Section 2 briefly reviews related work. Section 3 introduces a communication behavior indicator and a user segmentation algorithm based on the FCM that are used to study the questions proposed by us. In Section 4, we investigate user behavior patterns and extract salient characteristics that indicate the relationships among users' consumption capacity, communication time, mobility, and locations. Finally, we further discuss our results and conclude the paper in Section 5.

2. Related work

Telecommunication industry has been developing rapidly since early 1990s. With increased market competition and years of development, the number of mobile users is becoming saturated. Notably, the proportions of mobile users are even more than 100% in some metropolises. This implies that some users subscribe more than one mobile number. Such a situation forces mobile operators to turn their efforts from increasing the number of subscribers to retaining the existing ones. As a result ensuring the quality of mobile services becomes the most important factor to enhance competitiveness. Typically, the ISP's are in search of data that possess key information about their users and try to distinguish users and in-depth understand the needs of different user groups.

To know the relationship between customer value and behavior, an appropriate method for customer segmentation is critical in CRM. The customer value, in most instances, only reflects the consuming capacity of a user. Segmentation based on it cannot distinguish the difference of user behaviors, which, however, contains useful and valuable information about the characteristics, preferences and desires of users.

Customer segmentation methods are classified in terms of segmentation dimensions and the purposes of segmentation. Generally, there are four kinds of customer segmentation methods [7]: demographics segmentation [8], lifestyle segmentation [9–12], behavior segmentation [13] and benefit segmentation [14]. Demographic segmentation treats geography as an important dimension. However, the globalization of markets and the rapid development of information technology weaken the relevance between customer and geographic characteristics. Lazer firstly proposed the method to identify and segment customers based on their lifestyle [11], which includes: Activity, Interests and Opinion (AIO) [9,10,12]. However, this kind of segmentation is hard to do since it is generally impossible to get customers' comprehensive lifestyle data in practice. Behavior segmentation classifies customers by analyzing their behavior patterns [13]. Supported by information technology, we can handle a large amount of data to get useful results with this method. But simply using behavior segmentation cannot disclose customer value or benefit, which is mostly concerned by the operators. There are many ways to calculate the value of customers. The most popular one is segmenting users based on ARPU. The famous pyramidal model typically divides the users based on the ARPU into 3 clusters: high, medium, and low ARPU, respectively. In this paper, we choose user behavior as an indicator to give an in-depth insight of the relationship between customer value and behaviors by considering the weakness and advance of the existing segmentation methods.

CRM is a broadly recognized, widely implemented strategy to manage and nurture a company's interactions with its users and sales. Its overall goal is to find, attract, and win new users, retain existing ones and entice former users back, and reduce the costs of marketing and customer services [6]. In order to achieve successful customer relationships, data mining (DM) is generally applied to understand the characteristics and desires of users [15]. The DM is a process of finding hidden patterns, associations, rules and statistically significant structures in large databases [16]. Clustering is a DM technique with applications in the areas of data exploration, segmentation, targeted marketing, and cross-selling [17,16]. With clustering, the CRM aims to segment users into discrete groups that share similar characteristics, such as age, gender, interests, and consumption habits. However, we notify that DM methods in the previous work seldom cluster users based on both their behaviors and ARPU.

Recently, analysis of user behaviors is becoming a popular approach to understand users. A number of applications ranging from city planning to resource management in mobile communications rely on the understanding of human behaviors. For example, Lopez-Paris et al. found that pedestrians usually walk along straight pavements in cities. Based on this, they proposed an approach to automatically agent navigation in realtime applications [18]. Eagle and Pentland found that different MIT staff's mobility patterns are different: higher entropy of junior staff indicates

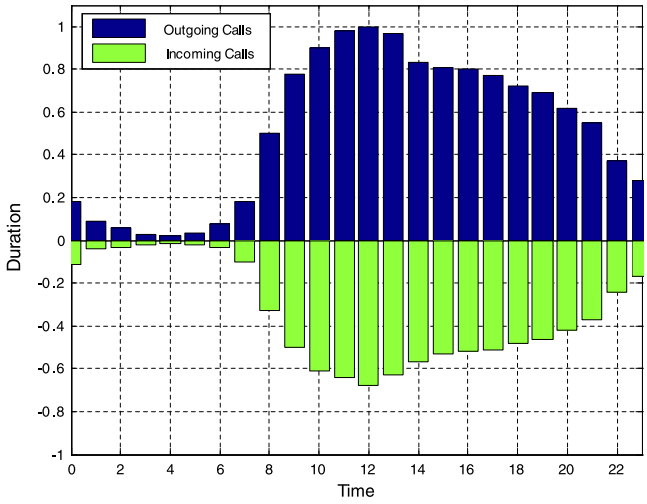

Fig. 1. The outgoing and incoming call durations of a user in a day.

more random movement of juniors than seniors [1]. Sohn et al. [19] explored how coarse-grained GSM data collected by mobile phones can be used to recognize properties of user mobility through the daily number of walking steps. Pan and Jon analyzed community structures by tracking human mobility [20]. They used delivery ratio and delivery cost to design efficient forwarding algorithms for mobile networks. E. Lu et al. proposed a method, named Cluster-based Temporal Mobile Sequential Pattern Mine (CTMSP-Mine) to discover Cluster-based Temporal Mobile Sequential Patterns (CTMSPs) in the context of Location-Based Services (LBS). This work mines and predicts mobile behaviors by simultaneously considering both user relations and temporal property [21]. Sun et al. presented a suite of detection techniques to identify fraudulent usage of mobile telecommunication services by exploiting regularities demonstrated in users' behaviors [22]. Song et al. indicated the limits of predictability in human mobility [23]. Kakiuchi et al. proposed an automatic human tracking system to overcome the limitation of existing video surveillance systems by applying the competence that a mobile agent normally does not lose tracking in automatic human tracking systems [24].

These researches give us inspiration to explore new applications or analyze human characteristics based on user behaviors. However, existing studies lack further segmentation and analysis by considering both value and behaviors. None of them concern the three questions raised by us. Thus they cannot provide valuable suggestions to the mobile operators in terms of the communication behavior, movement behavior and consumption behavior.

3. Communication behavior indicator and FCM algorithm

To answer the aforementioned three questions in Section 1, we study the users behaviors and try to cluster the users into different groups. By comparing the behavior features of the users in different groups, we hope to find some relations among the features, such as ARPU, mobility or calling habit. Therefore, we first define some quantitative indicator to describe the users communication behavior and then introduce a user segmentation algorithm to accomplish the clustering.

3.1. Communication behavior indicator

We propose a communication behavior indicator to reflect the affordability, contact relationship and calling habits of a user. From the operators' point of view, they are interested in the calling behavior. Thus, referring to the concept of traffic we select the duration (in minutes) of outgoing/incoming call per hour as the user's communication behavior indicator.

From h o'clock to $h+1$ o'clock, the total duration of outgoing call is:

$$x_h = \sum_{i=1}^{n} t_{hi} \tag{1}$$

where n is the number of outgoing calls in this hour, t_{hi} represents the duration of the ith call. Hence we get user j's outgoing call vector in a day as:

$$X_j = \{x_1, x_2, \ldots, x_{24}\}. \tag{2}$$

The outgoing call vector provides a detailed statement about the tendency of initiative communication request of a user. Considering that many operators take one-way charge, the outgoing call vector also reflects the affordability of the user.

Similarly, we define an incoming call vector as:

$$Y_j = \{y_1, y_2, \ldots, y_{24}\} \tag{3}$$

where y_i is the total incoming call duration (in minutes) at the ith hour of a day.

X_j and Y_j not only indicate the total traffic but also imply the characteristic of communication behaviors. Fig. 1 illustrates a certain user's outgoing and incoming call durations in a day by averaging and normalizing his 16-weeks calling records. The duration of outgoing calls is drawn in positive y axis while incoming ones are oppositely shown in negative y axis.

The outgoing and incoming call vectors give us an insight of communication difference among users. This cannot be indicated by the pyramidal model commonly used by the operators. However, even two users with similar ARPU, may behave quite differently in communications, e.g., one would like making calls in the morning while the other prefers at night.

3.2. User segmentation with fuzzy c-means clustering

A user's behavior is random, nonlinear and multi-attribute associated with many factors. In order to investigate user behaviors we must segment users based on different behavior patterns. Fuzzy clustering is a common approach for this purpose. Fuzzy clustering analysis allows one piece of data to be loaded in two or more clusters and finally decides the most suitable one. The clustering result depends on the whole data set that can keep all characteristics needed in latter analysis. The most popular fuzzy clustering algorithm is fuzzy c-means (FCM) clustering, which is also applied in this paper. This method (developed by Dunn in 1973 and improved by Bezdek in 1981 [25,26]) is frequently used in pattern recognition.

The FCM clustering is based on the minimization of following objective function:

$$J_m(U, C) = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \|x_i - c_j\|^2, \quad 1 \le m \le \infty \tag{4}$$

where J_m is an objective function, m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith d-dimensional measured data ($d = 2 \times 24$ in our case, where 24 is the total number of hours per day, 2 indicates the types of calling: outgoing call and incoming call), c_j is the center of cluster j in d-dimension, N is the total number of measured data, C is the total number of clusters, and $\| * \|$ is the norm expressing the similarity between two data.

Fuzzy partitioning is carried out through iterative optimization of the objective function as shown in Eq. (4), by updating the degree

of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{5}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}. \tag{6}$$

This iteration stops when $\max_{ij}\{|u_{ij}^{(k+1)} - u_{ij}^{(k)}|\} < \varepsilon$, where ε is a termination criterion between 0 and 1, and k is the iteration step. This procedure converges at a local minimum or saddle point J_m.

The FCM algorithm is described in Algorithm 1. The convergence of the algorithm has been proved in [27].

Algorithm 1 FCM algorithm:

1: Step 1. Initialize $U = [u_{ij}]$ matrix, $U(0)$, $k = 0$;
2: Step 2. At the kth step: calculate center vector $Ct^{(k)} = [c_j]$ with $U^{(k)}$ based on Eq. (6);
3: Step 3. $k++$, update $U^{(k)}$, and get $U^{(k+1)}$ based on Eq. (5) using c_j got in Step 2;
4: Step 4. Calculate the new partition matrix $\Delta = \|U^{(k+1)} - U^{(k)}\| = \max_{ij}\{|u_{ij}^{(k+1)} - u_{ij}^{(k)}|\}$
5: **if** $\Delta < \varepsilon$ **then**
6: STOP;
7: **else**
8: go to Step2.
9: **end if**

The weighting exponent m is an important parameter in the FCM algorithm. When m is close to one, FCM approaches hard c-means algorithm, also known as K-means. When m approaches infinity, the only solution of FCM is the mass center of input data set. Hence, choosing a suitable m is very important when implementing FCM. Bellman and Zadeh gave an optimization tool for choosing m [28]. Many researchers studied the optimization of the m chosen. Bezdek et al. thought m should be between 1.1 and 5 [25]; Cheung and Chan considered m between 1.25 and 1.75 based on application background [29].

We choose the method introduced in [30] to calculate m. Given a fuzzy objective function G and a constraint C_f, a decision is produced by the intersection of G and C_f, $D = G \cap C_f$ [25]. The membership function $u_G(m)$ of data set D is defined as:

$$u_G(m) = \exp\left\{ -\alpha \cdot \frac{J_m(U, C)}{\max_{\forall m}(J_m(U, C))} \right\} \tag{7}$$

where α is a positive constant larger than 1, typically, $\alpha = 1.5$. $J_m(U, C)$ is the object function described in Eq. (4), $U = [u_{ij}]$ is the membership metric, C is the number of clusters. FCM algorithm wants to get a minimal value of J_m.

The membership function of fuzzy constraint C_f is:

$$u_{C_f}(m) = \frac{1}{1 + \beta \cdot \left(\frac{H_m(U,C)}{\max_{\forall m}(H_m(U,C))} \right)} \tag{8}$$

where β is a positive constant, typically, $\beta = 10$. H_m is the entropy of a partition defined as:

$$H_m(U, C) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij} \cdot \log_a(u_{ij}) \tag{9}$$

a is the base of logarithm that $a \in (1, +\infty)$. Other notions have the same meanings as described before. Particularly, when $u_{ij} = 0$, $u_{ij} \cdot \log_a(u_{ij}) = 0$. The smaller the H_m, the better the partition is.

Furthermore, we get an optimized weighting exponent m^*:

$$m^* = \arg_{\forall m} \left\{ \max\{\min\{u_G(m), u_{C_f}(m)\}\} \right\}. \tag{10}$$

As mentioned before, we get three levels of users based on their ARPU. We further calculate m^* based on Eqs. (7)–(10), as depicted in Fig. 2.

Referring to the outgoing and incoming call vectors, X_j and Y_j, we get matrixes X and Y for n users as:

$$X = \{X_1, X_2, X_3, \ldots, X_n\} \tag{11}$$

$$Y = \{Y_1, Y_2, Y_3, \ldots, Y_n\}. \tag{12}$$

We use X and Y as the inputs of FCM algorithm and normalize them before processing in order to get a comparable result. Herein, we use Min–Max normalization. That is, data values are scaled such that the smallest value in the two arrays (i.e., X and Y) becomes 0 and the largest one becomes 1. Our proposed algorithm for user segmentation is described in Algorithm 2.

Algorithm 2 Algorithm of user segmentation:

1: Step 1. Divide all sample users into three levels based on their ARPU (from high to low, the thresholds are top 20% ARPU, top 20% − 50% ARPU and others);
2: Step 2. Normalize user data as the input of FCM algorithm;
3: Step 3. Implement the FCM algorithm for each level of users, cluster them into 3 classes ($C = 3$);
4: Step 4. Average the communication behavior indicators of each class and calculate user proportion for each corresponding level.

4. User behavior patterns

Based on the algorithm presented in Section 3, in this section, we investigate user behavior patterns and extract salient characteristics that indicate the relationships among users' consumption capacity, communication time, mobility, and locations.

4.1. Communication behavior clustering

In order to study the communication behaviors of users, we firstly divide the one million sample users into 3 levels based on their ARPU (from high to low), then we cluster each level of users by applying the FCM algorithm.

The proportion and total ARPU contribution of each level is shown in Fig. 3. We know that the famous 20−80 rule in economics indicates that 20% of customers contribute 80% of profits. However, from Fig. 3 we can see that due to fierce competition, the gap of total revenue between each level is narrowed. This indicates that we cannot underrate any level of users. Applying the FCM algorithm for each level of users, the clustering result is shown in Fig. 4. The subgraphs in Fig. 4 present the results of different ARPU levels, from high to low. For the sake of convenience, we define the clusters as $c1$, $c2$ and $c3$ in a certain ARPU level, respectively. The proportion of $c1$, $c2$ and $c3$ is illustrated on the left of each subgraph.

From Fig. 4, we can see that for the top ARPU level, $c1$ has longer outgoing and incoming calling durations in the morning than afternoon or evening and night. This situation is reversed for $c2$. The cluster $c3$ has sharp peak calling durations in the evening around 20:00 h. Generally, the top ARPU level users have longer outgoing call duration than the incoming one, which indicates that they have strong affordability.

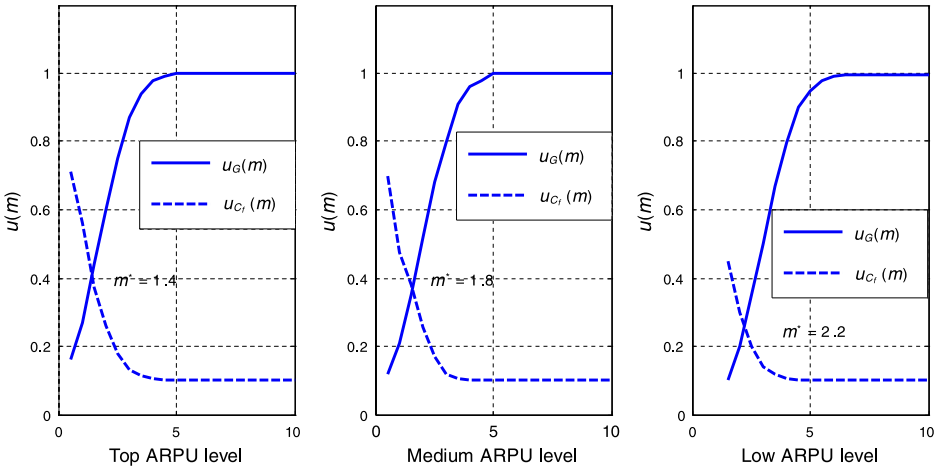

Fig. 2. The optimized weighting exponent m for each level of users.

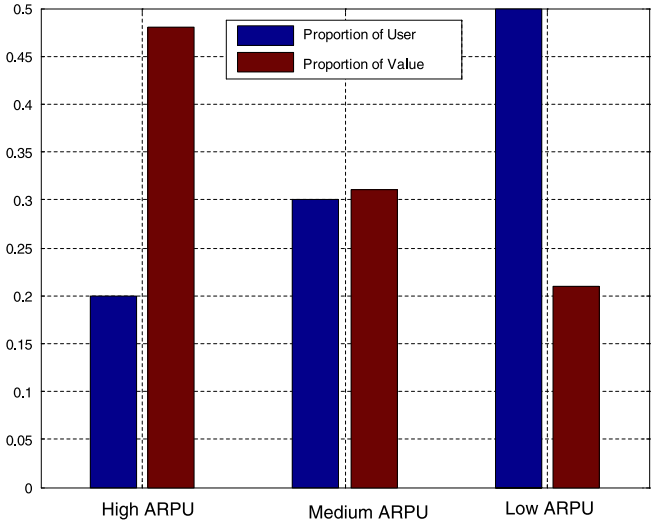

Fig. 3. The proportion and total ARPU contribution of each level.

For the medium ARPU level, $c1$, $c2$, $c3$ have similar characteristics to the top one, except that the outgoing calling duration is approximately equal to the incoming one.

For the low ARPU level, the characteristics of $c1$, $c2$, $c3$ are not so as distinct as the top and medium levels. We get the largest $m^* = 2.2$ in the implementation of FCM algorithm for this level. But based on the incoming call duration, we can still draw the same conclusion as the other two ARPU levels. Besides, this level obviously has longer incoming call duration than the outgoing one. In summary, people's communication behaviors can be divided into 3 patterns, as $c1$, $c2$, $c3$ described above.

4.2. Communication behavior and mobility

As introduced before, our data set records the user's closest base station ID when a call is generated. In urban area, a base station's coverage is about 1–3 km. Thus we can get the approximate location of the user based on the base station ID. We use scope and regularity to characterize a user's mobility. The scope is the number of different base stations visited by a user in a month and can represent the range of the user's movement. The regularity is denoted by the entropy of a user's movement. In information theory, entropy is a measure of uncertainty associated with a random variable. Kontoyiannis et al. measure the entropy of a user's trajectory to explore the user's mobility [17]. The smaller the entropy of trajectory is, the more steady the user's mobility is.

Table 1
The scope of different ARPU level users with different communication behaviors.

Scope ARPU	$c1$ (41%)	$c2$ (22%)	$c3$ (37%)	Average
High (20%)	53.76	51.27	50.29	51.93
Medium (30%)	34.33	32.81	30.09	32.43
Low (50%)	17.59	16.27	16.39	16.85
Average	29.84	28.23	27.28	28.54

The distribution of scope of all sample users is shown in Fig. 5.

The average scope is 28.54. In other words, a user's trajectory on average crosses nearly 28 base stations' service areas in a month. With the result presented above, we calculate the scope of different ARPU level users with different communication behaviors, as shown in Table 1.

From Table 1, we can see that the scope has strong relationship with ARPU: the higher the ARPU, the higher the scope. As for the communication behavior, $c1$ has the highest scope value, but not much higher than $c2$ and $c3$. $c2$ has a similar scope value to $c3$. Thus, we can draw a conclusion that the scope has tight relationship with ARPU and people who like calling in the daytime have more scope.

We further investigate the regularity of user mobility.

Definition: let X_i be a random variable representing a user's location at time i. Movement entropy is defined as:

$$S(X) = -\sum_{x \in X} p(x) \log_2 p(X), \tag{13}$$

where $p(x) = P(X_i = x)$ is the probability that $X_i = x$. For a stationary stochastic process, the user movement entropy could be written as:

$$S \equiv \lim_{n \to \infty} \frac{1}{n} S(X_1, X_2, \ldots, X_n). \tag{14}$$

People who have high-entropy of trajectory tend to live very variably and it is hard to predict their locations. While low-entropy of trajectory is characterized by solid patterns across all time scales, by calculating the entropy of users, we can provide insight into their lives in order to lay a valuable foundation for CRM.

We segment the 16-week detailed calling records into hour-long intervals for all sample users. For each interval, each user is assigned a base station ID to indicate its location. If the user passes more than one base station in a given interval, we choose one of them randomly. If the user has no call in this hour interval, we fill it with a symbol "?". We use a string to express the user's movement trajectory, which contains a number of symbols that are the IDs of the base stations visited by the user and "?"s. Thus, we obtain

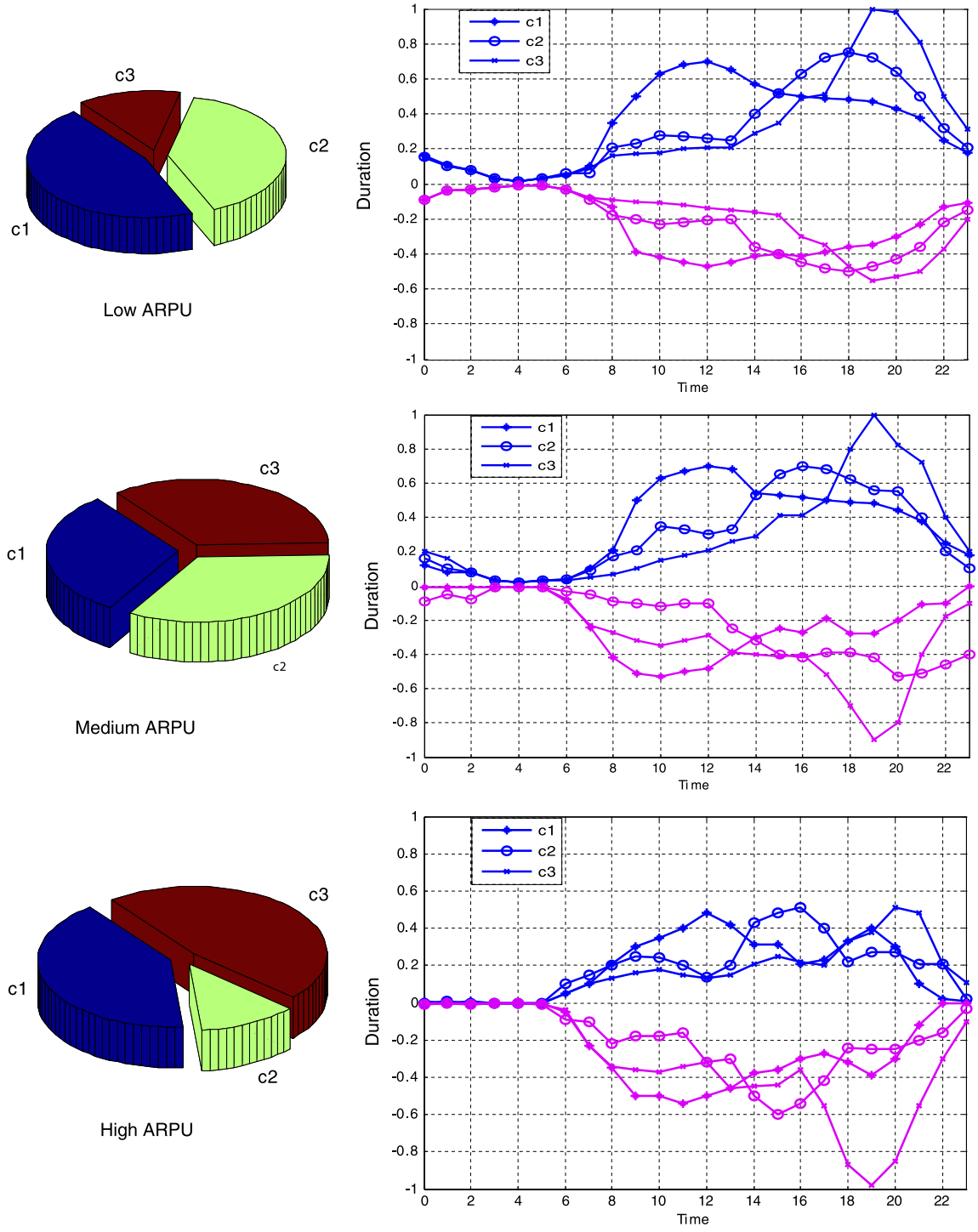

Fig. 4. Clustering result for each ARPU level.

a string with the length $L = 24 \times 7 \times 16 = 2688$ for each user (24: the number of hour-long intervals per day; 7: the number of days in a week; 16: the total number of weeks during which our user data were collected). The string presents the trajectory of the user's movement. Thereby, the entropy of the string can indicate the stability of the trajectory. Obviously, the less the entropy, the stronger the stability of the trajectory of user movement is.

To calculate the entropy based on the user's past location history, we use an estimator based on Lempel–Ziv data compression [31] to calculate the real entropy of the string used to express the trajectory of personal movement. For a string with n symbols, the entropy is estimated by

$$S^{\text{est}} = \left(\frac{1}{n}\sum_i \Lambda_i\right)^{-1} \ln n \tag{15}$$

where Λ_i is the length of the shortest substring starting at position i that does not previously appear from position 1 to $i-1$. It has been proven that S^{est} converges to the actual entropy when n approaches infinity [31]. Based on Eq. (15), we get the probability distribution of entropy, as shown in Fig. 6.

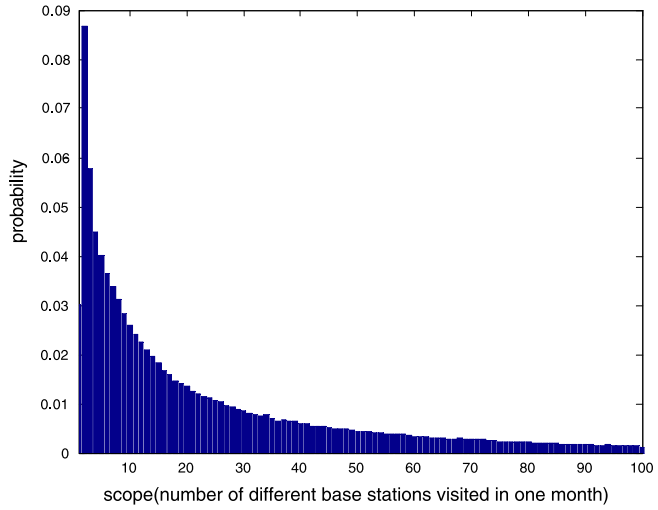

Fig. 5. The scope of sample users.

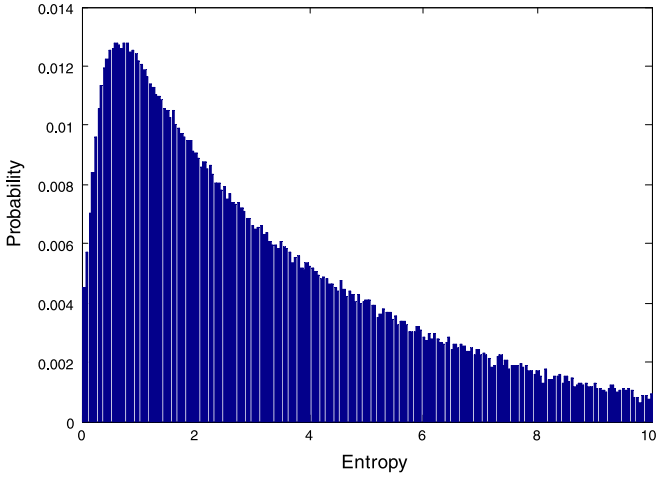

Fig. 6. The probability distribution of entropy.

From Fig. 6, we can see that most people's entropy is less than 2, which indicates that most users' whereabouts at any time is less than $2^2 = 4$ locations. The average of all users' entropy is about 3.07. We calculate the regularity of different ARPU level users with different communication behaviors, as shown in Table 2.

Table 2 implies that the people who make less calling or like making calls at night have more regularity than other clustered users (smaller entropy indicates more regularity). We also find an interesting result: suppose R_{c^*} (c^* would be $c1, c2$, or $c3$) denotes the regularity of c^* and R_{A^*} (A^* would be AH, AM, or AL) denotes the regularity of high (AH), medium (AM) and low (AL) ARPU level users, there are $2^{R_{c2}} \approx \left(2^{R_{c1}} + 2^{R_{c3}}\right)/2$ and $2^{R_{AM}} \approx \left(2^{R_{AH}} + 2^{R_{AL}}\right)/2$. This means that if we divide the users based on their regularity, we will get the same division as the result based on ARPU. This result also indirectly proves that the user behavior analysis method proposed by us is an appropriate one, like the pyramidal model based on ARPU.

4.3. Group clustering

A movement group (in short, group) is a number of persons who move together from one place to another. As shown in Fig. 7, the trajectory of group movement may pass through many base stations' coverage. In our daily life, a bus or a metro is an instance of movement group. People who go to the same company may have a segment of same trajectory to go to work and get off work. They have the greatest possibility being a group. There are many other reasons people move together. A method to find movement group is proposed in [1]. Based on our data set, we find 200 groups to investigate the group members' characteristics. Each group's size is from 100 to 500 persons. The total number of group users is 27 468. The users of each group exhibit similar movement trajectory despite that they may be not aware of it. In addition, the movement group's trajectory is very steady. This implies that they live in a life with "low-entropy" and they "like" to be with each other. No matter moving or staying, they have changeless "close friends" around them, although they do not even know this fact.

Table 2
Regularity of different ARPU level users with different communication behaviors.

Regularity	$c1$ (41%)	$c2$ (22%)	$c3$ (37%)	Average
High (20%)	6.51	5.75	3.69	5.30
Medium (30%)	5.63	4.73	2.16	4.15
Low (50%)	1.90	1.61	1.07	1.53
Average	4.44	3.63	1.22	3.07

Table 3
The proportion of group users in different ARPU level users with different communication behaviors.

Proportion	$c1$	$c2$	$c3$	Total
High ARPU	2.3%	0.6%	0.5%	3.4%
Medium ARPU	23.2%	20.4%	14.9%	58.5%
Low ARPU	21.2%	14.4%	2.5%	38.1%
Total	46.7%	35.4%	17.9%	100%

We calculate the proportion of all 27 468 group users in different ARPU level users with different communication behaviors, as shown in Table 3. We can see that the users who belong to the high level ARPU are the most "lonely" ones—they are hard to be "found" by others. They may have their own vehicles and private work place. They would like to come and leave all alone. The most gregarious ones are the medium ARPU level users. In addition, $c1$ and $c2$ contain much more groups than $c3$, which indicates that people who like calling at night may be lonelier than those calling in the daytime.

4.4. Results and discussions

We found differences among the three clusters of each ARPU level. This differences cannot be seen from the ARPU division. It is interesting to note that the $c3$ cluster has peak calling traffic at night. In high ARPU level, the proportion of $c3$ is small but in medium and low ARPU levels, it is large (about 45%). This group of users (about 40% of total population) should be paid attention by the operators because they are sensitive on communication cost but have strong communication demands.

We also studied the relationship among movement behaviors, consumption capacity and communication behaviors. We gave an insight into user mobility and aggregation. It is easy to understand the result about scope: the stronger the user consumption capacity, the larger the range the user would travel. People who can afford high telephone charges could possibly afford international travel. We also find that people who like calling in the daytime have more mobility than those who like calling at night. However, this difference is not very significant.

We investigated the regularity of users and found that low level ARPU users and $c3$ users have low entropy. The reason may be that the work or life of low level ARPU users is simple and repeatable, thus these users move more regular (i.e., have lower movement entropy) than others. And those who like calling at night may have fewer activities during the daytime than others. In other words,

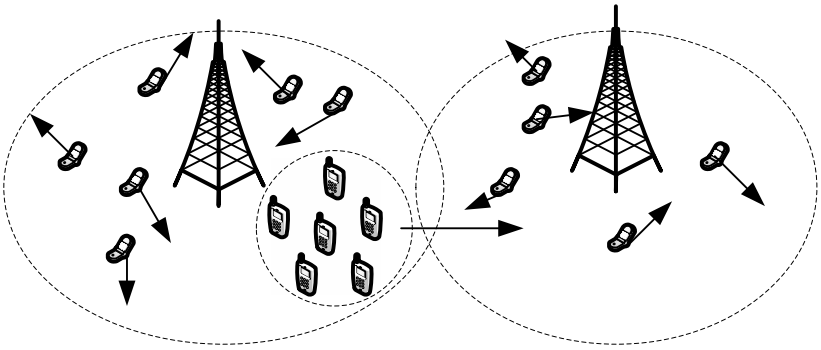

Fig. 7. An illustration of movement group.

they are more inclined to stay at home. What is more, we find that the regularity of medium ARPU users is approximate in the middle of high and low level ARPU users'. The same conclusion can be made for the $c1$, $c2$, and $c3$ clusters. This result can be seen as a good proof of the effectiveness of our proposed algorithm.

At last, we studied the behavior pattern of user aggregation. We found that the high ARPU users are the most "lonely" ones. They occupy the smallest proportion (3.4%) of the whole group population. We could explain this phenomenon that most of the high ARPU users are the elite of society, so it is hard to find many people move together with them. Medium ARPU users occupy nearly 60% of group population, which means middle class are the most gregarious. This finding could provide a significant justification for the operators to design and develop valuable mobile services for mobile users, e.g., mobile social network applications. We found that $c3$ is the "lonely" ones, too. This result can be explained as: people who feel lonely are willing to call at night. Based on this, corresponding services can be provided to satisfy the need of this cluster of users, e.g., psychological consultant services, social agent systems, etc.

5. Conclusions and future work

Due to the popularity of mobile phone usage and the fast development of computer technology, we can study human behaviors based on telecommunication records. Comparing with questionnaires, the telecommunication records can accurately reflect people's real life, thus the data analysis based on these records is more convinced. Our work described in this paper is significant to instruct ISP CRM and help anthropology study. We proposed an algorithm of user segmentation based on FCM. Using this algorithm, we divide mobile users into 3 clusters considering their communication behaviors. Integrating with the traditional pyramidal model based on ARPU, we got in-depth behavior patterns of mobile users.

However, limited by the supporting projects purpose and privacy concerns, we can only use the 16-week long and a little bit outdated data. Nevertheless, the proposed methods in the paper have already been submitted to the service provider and used in their system. According to their feedback, the methods work fine for other later CDR data. Therefore, we believe the 16-week long data is enough for the research itself and suitable for ISP's CRM purpose.

For future work, we will work on the technologies for privacy preservation in order to appropriately make use of user data for better mobile services. And we hope to model users' long term behaviors and the behavior dynamics with the support of data of longer period, which can be even more valuable to the service providers.

Acknowledgments

This work is supported by the Young Scientist Fund of National Natural Science Foundation of China (Grant No. 61202303) and the National Science and Technology Major Project of China (Grant No. 2012ZX03001035-003).

References

[1] N. Eagle, A. Sandy Pentland, Reality mining: sensing complex social systems, Pers. Ubiquitous Comput. 10 (2006) 255–268.
[2] Engineering Social Systems, 2002..
[3] D.L. Nelson, J.C. Quick, Organizational Behavior: Foundations, Realities, & Challenges, Thomson Press, 2005.
[4] S. Strogatz, Exploring complex networks, Nature 410 (2001) 268–276.
[5] L. Freeman, The Development of Social Network Analysis, Empirical Press, Vancouver, 2006.
[6] C.B. Bhattacharya, S. Sen, Consumer-company identification: a framework for understanding consumers' relationships with companies, J. Mark. 67 (2003) 76–88.
[7] J.W.C. Rygielski, C. David, Data mining techniques for customer relationship management, Technol. Soc. 24 (2002) 483–502.
[8] S. Soper, Practice papers: the evolution of segmentation methods in financial services: where next? J. Financ. Serv. Mark. 7 (2002) 67–74.
[9] F. Bushman, Systematic life styles for new product segmentation, J. Acad. Mark. Sci. 10 (1982) 377–394. http://dx.doi.org/10.1007/BF02729342.
[10] J.T. Plummer, The concept and application of life style segmentation, J. Mark. 38 (1974) 33–37.
[11] E.W. Lazer, Stephen Greyser, Life style concept and marketing, in: Toward Scientific Marketing: Proceedings of the Winter Conference, American Marketing Assn., Chicago, 1963.
[12] D.T.W. Wells, Activities, interests, and opinion, J. Adv. Res. 11 (1971).
[13] A. Hughes, Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable, Customer-Based Marketing Program, Irwin Professional, 1994.
[14] R.I. Haley, Benefit segmentation: a decision-oriented research tool, J. Mark. 32 (1968) 30–35.
[15] H. Nemati, C. Barko, Enhancing enterprise decisions through organizational data mining, J. Comput. Inf. Syst. (2002) 21–28.
[16] S.C.U. Fayyad, P. Bradley, Data mining and its role in database systems, in: VLBD 2000 Tutorial, 2000, pp. 27–35.
[17] T.J.H. Hwang, E. Suh, An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry, Expert Syst. Appl. 26 (2004) 181–188.
[18] D.L. Paris, A. Brazalez, A new autonomous agent approach for the simulation of pedestrians in urban environments, Integr. Comput.-Aided Eng. 16 (2009) 283–297.
[19] T. Sohn, A. Varshavsky, A. LaMarca, M.Y. Chen, T. Choudhury, I. Smith, S. Consolvo, J. Hightower, W.G. Griswold, E. De Lara, Mobility detection using everyday GSM traces, in: Proceedings of the 8th International Conference on Ubiquitous Computing, UbiComp'06, Springer-Verlag, Berlin, Heidelberg, 2006, pp. 212–224.
[20] H. Pan, C. Jon, Human mobility models and opportunistic communications system design, Philos. Trans. R. Soc. Lond. Ser. A 366 (2008) 2005–2016.
[21] E.-C. Lu, V. Tseng, Mining cluster-based mobile sequential patterns in location-based service environments, in: Tenth International Conference on Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09, pp. 273–278.
[22] B. Sun, Y. Xiao, R. Wang, Detection of fraudulent usage in wireless networks, IEEE Trans. Veh. Technol. 56 (2007) 3912–3923.
[23] C. Song, Z. Qu, N. Blumm, A.-L. Barabasi, Limits of predictability in human mobility, Science 327 (2010) 1018–1021.

[24] T.S.H. Kakiuchi, T. Kawamura, K. Sugahara, Bypass methods for constructing robust automatic human tracking system, Integr. Comput.-Aided Eng. 17 (2010) 41–58.
[25] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Academic Publishers, Norwell, MA, USA, 1981.
[26] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. 3 (1973) 32–57.
[27] J.C. Bezdek, A convergence theorem for the fuzzy isodata clustering algorithms, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-2 (1980) 1–8.
[28] R.E. Bellman, L.A. Zadeh, Decision-making in a fuzzy environment, Manag. Sci. 17 (1970) B-141–B-164.
[29] Y.S. Cheung, K.P. Chan, Modified fuzzy isodata for the classification of handwriting Chinese characters, in: Proc. of Int. Conf. Chinese Comput., Singapore, pp. 361–364.
[30] X. Gao, J. Pei, W. Xie, A study of weighting exponent m in a fuzzy c-means algorithm, Acta Electron. Sin. 28 (2014) 80–83.
[31] I. Kontoyiannis, P.H. Algoet, Y.M. Suhov, A.J. Wyner, Nonparametric entropy estimation for stationary processes and random fields, with applications to english text, IEEE Trans. Inf. Theory 44 (2006) 1319–1327.

Zhenhua Wang is a Ph.D. student in Information and Communication Engineering from Huazhong University of Science and Technology, China. His research areas include big data, social computing, and information management.

Lai Tu received the B.S. in Communication Engineering and Ph.D. degree in Information and Communication Engineering from Huazhong University of Science and Technology, China, in 2002 and 2007 respectively. From 2007/7 to 2008/12, he worked as a postdoc fellow in the Department of Electronics and Information Engineering in Huazhong University of Science and Technology. From 2009/1 to 2010/10, he worked as a postdoc researcher in the Department of CSIE in National Cheng Kung University, Taiwan. Currently, he is an Associate Professor of the Department of Electronics and Information Engineering in Huazhong University of Science and Technology. His research areas include social computing, human behavior study, mobile computing and networking.

Zhe Guo, was born in 1982 and received his B.S. in Communication Engineering and Ph.D. degree in Information and Communication Engineering from Huazhong University of Science and Technology in 2004 and 2011 respectively. His research areas include human behavior in mobile network, mobile network optimization, cloud computing.

Laurence T. Yang graduated from Tsinghua University, China and got his Ph.D. in Computer Science from University of Victoria, Canada. He joined St. Francis Xavier University in 1999. He has published many papers in various refereed journals, conference proceedings and book chapters in these areas (including around 100 international journal papers such as IEEE Transactions on Computers, IEEE Journal on Selected Areas in Communications, IEEE Transactions on Parallel and Distributed Systems, IEEE Intelligent Systems, etc). He has been involved actively in conferences and workshops as a program/general/steering conference chair (mainly as the steering co-chair of IEEE UIC/ATC, IEEE CSE, IEEE HPCC, IEEE/IFIP EUC, IEEE ISPA, IEEE PiCom, IEEE EmbeddedCom, IEEE iThings, IEEE GreenCom, etc) and numerous conference and workshops as a program committee member. His current research includes parallel and distributed computing, embedded and ubiquitous/pervasive computing.

Benxiong Huang, is Doctoral Supervisor, professor in Department of Electronic and Information Engineering in Huazhong University of Science of Technology, and vice director of National Engineering Laboratory of Next generation Internet access system, and secretary-general of Innovation Institute of internet of things. His research interests cover communication system, next generation mobile internet, signal processing and social computing. He is the laureate of Fund for Distinguished Young Scholar of Hubei province, and member of specialist group of the integration of telecommunications networks, cable TV networks and the internet, senior member of China computer association, Commissary of Young Computer Scientists & Engineers Forum (YOCSEF), Chairman of the third YOCSEF in Hubei province, special councilor of Yichang local government, executive member of the council of Wuhan Institute of Electrical Technology, member of specialist group of the National Basic Research Program of China (973 Program)—"research on ultra-speed, ultra-capacity, and ultra-long distance optical transmission".