On Structural Features, User Social Behavior, and Kinship Discrimination in Communication Social Networks

Shu-Sen Zhang, Xun Liang, *Senior Member, IEEE*, Yu-Dang Wei, and Xuan Zhang

***Abstract*— In the research of social networks, their structural characteristics, user social behaviors, and user relationships are elements that are important to understand social networks, to predict user behaviors, and to manage social networks. In this article, we took as the research object of social networks the mobile communication network, which is closely connected with people's real lives. We studied the structural characteristics using the complex network analysis methods and derived the laws that exist in the structure of communication social networks. We analyzed users' social behaviors or social patterns from the perspective of age, gender, social scope, age differences, and time. In addition, we extracted various salient features of user's calling behavior, and used the XGBoost and logistic regression (LR) fusion method to establish the Kinship-XL model, which is able to improve the performance and speed. Through the experimental verification, the Kinship-XL model can determine whether there is a kinship between users or not.**

***Index Terms*— Kinship, network structure, social behavior, social networks.**

I. Introduction

IN RECENT years, social networks have received wide attention from computer science, sociology, psychology, complex systems, and other disciplines, and have become a focus of many scientific researchers [1]. Social networking is a relational structure formed by connections among users and is an extension of the users' real-life social relationships. Among the types of social networks, the communication network breaks the limitation of the traditional social network of individual time and space and is an "online" process in effect at any time. In this research, we chose the cellular telephone call networks and message networks in communication networks as a social network research object, namely the communication social networks. Compared with other social networks, such as Facebook, Twitter, micro blogging, and other online social networks, the communication social networks can reflect the users' social relationships in real life, and the users are all real rather than computer generated. Call and message networks are more authentic and can reflect actual social situations. Therefore, from the perspective of social network research and human social behavior analysis, these social networks have research value and significance.

In the past, social network research based on cell phone data mainly focused on the network structure, such as its topological properties [2], [3], and the prediction of social relationships between links or users [4], [5]. Communication social networks are formed by the social relationships in people's real lives, and the network structure shows certain rules and characteristics. From this, we believe that the analysis of network structures is of value and is feasible. In fact, studying the structure of social networks will help people deepen their understanding of real networks and predict or monitor the spread of news.

In social networks, much data is being generated through social interactions. This social data can serve as an important resource for understanding users' social behaviors. Previous studies have examined user behavior laws. For instance, recent studies have shown that, despite a clear dissimilarity among the methods of the mobile communication by specific individuals, there is marked regularity in human mobility behaviors, suggesting that most individuals follow a simple and reproducible pattern [6], [7]. Analyzing users' social behavior can have a great practical and commercial value. Telecom operators, for instance, can precisely target different types of users to improve their experience of using cell phones, thereby improving customer retention. A range of services can be envisioned here, including personalized recommendations, way-finding advice, and targeted advertising [8]. In addition, current research on users' social laws focuses mainly on the users of virtual social networks, such as Twitter and microblogs, and research on users in real social networks uses more traditional methods, such as questionnaires.

Another topic in social network research is user relations. Usually, researchers collect and analyze social data and use the relevance that users and their contacts in terms of attributes and behaviors to identify user relationships, or identification of anonymous identical users of cross platforms [9]. Indeed, the study of user relationships has important practical significance. For example, user relationships can greatly influence corporate marketing and sales, especially online sales. In social networks, kinship is often reflected in the

Manuscript received June 5, 2018; revised September 21, 2018, January 20, 2019, May 28, 2019, and December 7, 2019; accepted December 21, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 71531012, in part by the National Social Science Foundation of China under Grant 18ZDA309, in part by the Opening Project of State Key Laboratory of Digital Publishing Technology of the Founder Group, in part by the Natural Science Foundation of Beijing under Grant 4172032, and in part by the Jingdong Mall E-commerce Research Project under Grant 413313012. *(Corresponding author: Xun Liang.)*

The authors are with the School of Information, Renmin University of China, Beijing 100872, China (e-mail: zss2446@ruc.edu.cn; xliang@ruc.edu.cn; wdang@ruc.edu.cn; zhangxuanalex@ruc.edu.cn).

Digital Object Identifier 10.1109/TCSS.2019.2962231

strength of the connections between certain users. However, because of the openness of social networks, some inherently strong connections between users are inevitably lost. Currently, research on user kinship is usually based on an online social network and image data. Research using cell phone user data is relatively scant. In fact, there are many challenges in dealing or inferring kinships among communication social networks, such as the social networks complexity analysis problem, and due to the complexity of human behavior and the complexity of people's various reactions and interactions in the process of communication, it becomes more difficult to identify users' kinship. The massive data problem, with the rapid development of communication technology, the amount of user data in communication network is increasing, which brings us great challenges to distinguish the relatives in the communication network. For example, as the capacity of network nodes (users) continues to increase, it becomes more and more complicated when using a graph theory to analyze the communication networks topology. When the number of nodes is millions, it is more difficult to get valid results in a short time. In addition, communication data collection issues, privacy, and security issues are also challenges in studying the kinship of communication social networks.

In this article, our contributions are as follows.

1) We analyzed the structural characteristics of communication social networks through the complex network analysis methods. Previous studies have been more about the analysis of the structure and relationship of the telephone network. It lacked the comparison with other social networks, the mutual influence, and the analysis of the differences, especially the popular online social networks. In this article, we studied the differences with other networks and found the differences between the communication network and other social networks, and explained the conclusions and laws we obtained.

2) We analyzed the social behaviors of cell phone users from different angles as well as the differences in social behaviors of different users, and summarized users' social behavior laws. In this article, we also expressed the differences between users' social laws and social behaviors in a more intuitive form, and analyzed the causes of the differences.

3) We extracted various salient features regarding user call behavior and used a fusion of the XGBoost and logistic regression (LR) methods to establish the Kinship-XL model. The method has the advantages of efficiency and effectiveness, and can quickly obtain the results of relative discrimination, which is suitable for a large-scale social network analysis. In the previous study of kinship, due to the limitations of research conditions, there are often problems with small data sizes and small samples. The Kinship-XL model is based on big data, and its accuracy rate is higher than that of small-scale data.

II. Related Work

A. Related Research

The concept of social networks was first articulated by the German sociologist Georg Simmel to describe social interactions among people [10]. Research on social networks is done mainly using network analysis methods. The study of social networks is closely related to that of complex networks [1] and is the result of combining theoretical complex networks with those in real society. In 1957, Anatol Rapoport's mathematical model emphasized the importance of the distribution of degrees in social networks [11]. In 1969, Milgram, a professor of sociology at Harvard University, found that a connection between any two people in the real world can be formed by only six people on average, known as "six degrees of separation" [12]. In the 1990s, Watts and Strogatz [13] published a "small-world" network model that revealed the prevalence of small-world properties in the network. In 1999, Barabási and Albert [14] revealed the prevalence of power law distributions in networks, finding that nodes with many links in networks are usually fewer than those with fewer links. Thus, the Poisson distribution, which is conventionally considered in networks, may not be applicable. In addition, in the 1990s, the British anthropologist Dunbar [15], basing his work on the apes' intelligence and related social networks, inferred that the maximum number of people that an individual can maintain a stable social relationship with is 150. This is the famous "Dunbar's number," or the 150 rule. In 2002, Cohen *et al.* [16] theoretically analyzed the small-world network and its scale-free structural characteristics. The analysis of social network features is an important way to research and analyze social networks. However, the sample size in previous studies is often not large enough, and the conclusions are often inaccurate [1].

In social networks, analyzing the social behavior of users has important research and application value. In this vein, Fei *et al.* [17] designed a novel multitask learning algorithm that can predict a user's response (e.g., a comment or the like) to their friends' posts (e.g., on blogs) with respect to the message content. Zhou *et al.* [18] proposed a network-aware method for identifying four specific kinds of users. Their method is based on a set of attributes and measures derived from analysis of users' influence-related social behaviors and their dynamic connections. Han and Yan [19] put forward a novel approach to friend recommendation, where three aspects of social behavior—social standing, social content, and social relations—are extracted to represent the relationship of each user pair in large-scale microblog data. Nemiche *et al.* [20] presented a theoretical model of the transmission and evolution of social behaviors using the agent-based modeling technology. In addition, to analyze human behaviors in a social context, Roudposhti *et al.* [21] proposed a new approach that explores interrelations between body part motions in scenarios with people participating in a conversation.

Currently, research on user kinship is usually based on online social network and image data. Recent advances in kinship discrimination have shown that kinship between users can usually be authenticated or discriminated by feature descriptions and metric learning methods [22]. For example, Yan *et al.* [23] extracted multiple features using different face descriptors to characterize facial images from different angles. Incorporating this, they proposed a discriminative multimetric learning method that can describe facial images for the purposes of kinship verification. Zhang *et al.* [24]

TABLE I
EXPERIMENTAL DATA SETS

Dataset	Phone users (million)	Data records (ten million)
MPC	2.2	1.7
CDRs	3.9	48.9
SMS	2.4	5.6

extracted kinship features by using a deep convolutional neural network that was also based on user facial images. Puthenputhussery *et al.* [25], meanwhile, proposed a scale-invariant feature transform (SIFT) flow-based genetic Fisher vector feature for kinship verification. Outside of this, researchers have analyzed facial images to identify kinship by other methods, such as scalable similarity learning [26], robust similarity learning [27], and neighborhood repulsed correlation metric learning [28]. In addition, Boutellaa *et al.* [29] investigated the kinship verification problem from face video sequences, whose video information outperforms those using still images. Patel *et al.* [30] explored the effectiveness of the periocular region in verifying kinship from images captured in the wild and proposed a block-based neighborhood repulsed metric learning framework.

B. Related Data sets

In this article, the communication network data set is the China Telecom operator data set, which consisted of three parts: mobile phone contact (MPC) data, call records (CDRs) data, and short messaging service (SMS) data. The three data sets were collected during 2017. The data sets analyzed during our study were not publicly available for privacy reasons, their details are in Table I.

The data sets used in this article can truly reflect the structure of social networks and the characteristics of user behavior. The communication social network used in this article is composed of the call and SMS records between cell phone users, and it is analyzed with user's own attribute information. This social network is formed according to the interaction process between users, in which each point represents a mobile phone user. Because of the privacy of the data and the need to keep the information secure, the data sets have been desensitized and anonymous-replacement processed.

III. STUDY OF NETWORK STRUCTURE CHARACTERISTICS USING COMPLEX NETWORK ANALYSIS

Many complex systems in real life can be modeled as complex networks for research. Complex networks can also be seen as a form of data, but can also be used in a scientific research method. Social networks are complex networks formed during user interaction. Using the graph theory, we abstracted the social network into a graph consisting of point and edge sets, i.e., $G = (V, E)$, where V is the set of user nodes in the social network, the number of which is expressed as $N = |V|$, and E is the set of edges, i.e., the links between users in the social network, the number of which is expressed as $M = |E|$.

TABLE II
ADS OF NETWORKS

Network	ADD	ADU
MPC	6.72	6.85
CDRs	5.53	5.61
SMS	2.42	2.53

TABLE III
ADS IN DIFFERENT SOCIAL NETWORKS

Social network	Average degree(s)	Social network	Average degree(s)
Ten_m	51.56	MPC	6.72/6.85
Sin_m	11.98	CDRs	5.53/5.61
Twitter	18.6/46.3	SMS	2.42/2.53
Facebook	25.3/3.7	Other SMS	4.3

The measurement of the characteristics of social networks in this article is included mainly the average degree (AD) of directed networks (ADD), the AD of undirected networks (ADU), the clustering coefficient (CC), the assortativity coefficient (ac), and the average path length (APL).

A. Measure of AD

In social networks, the degree of one node (the number of connections to other nodes) can reflect the status and influence of the node (or user): the greater the degree of the node, the more important the node is in the network. In this article, the degree can be characterized as the users' popularity, influence, or activity. In a cell phone network, which is directed, a higher in-degree (number of incoming edges) indicates that a user's popularity is greater, and a higher out-degree (number of outgoing edges) indicates that the user's activity is greater.

As shown in Table II, the CDRs network is twice the SMS network in ADD and ADU. This shows that in real life, people communicate more by telephone than with SMS communication.

The ADs of social networks in this article were from 2 to 7, which is similar to those of other communication social networks, such as the average of SMS networks in [31], which is 4.3. Compared with online social networks, the ADs of social networks in our study are significantly lower than those of online social networks, such as the Tencent microblog (Ten_m) [32], the Sina microblog (Sin_m) [32], Facebook [33], and Twitter [34], [35]. The ADs of specific information of different social networks are shown in Table III.

We analyzed the reasons for these differences, the social network in this article is based on the user's real relationship, and the social breadth will be subject to certain constraints. The degrees of the online social network are attributable to its own open, virtual features, and the social scope of the users is unlimited. Therefore, the AD of online social networks is greater than that of the social network we studied. We also considered ADs from the standpoint of user activity level. The higher AD of online social networks suggests that people are

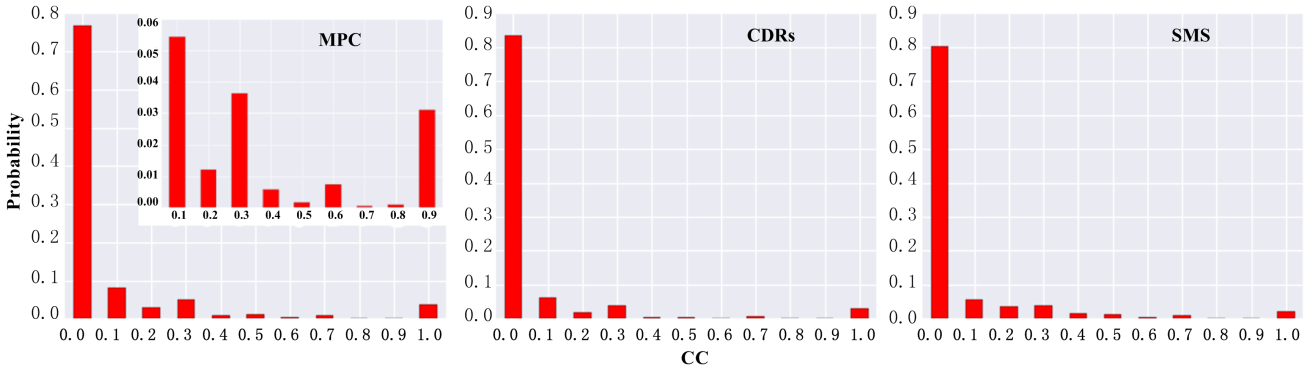

Fig. 1. CCs of the CDRs, MPC, and SMS networks. The top-right graph in the MPC network is the distribution after removing the extreme values of the CC in that network. The x-axes are the CC values, and the y-axes are the probabilities of the values.

TABLE IV
CC IN DIFFERENT SOCIAL NETWORKS

Social network	CC	Social network	CC
Ten_m	0.132 [36]	MPC	0.043
Sin_m	0.024 [36]	CDRs	0.003
Twitter	0.262 [37]	SMS	0.001

more motivated in virtual networks and more active in social activity.

B. Clustering Coefficient

In a social network, the CC is used to describe the close relation degree of neighbors, i.e., the probability of a friend's friend also being a friend. The CC is defined as

$$\mathrm{CC} = \frac{2E_i}{k_i(k_i - 1)} \tag{1}$$

where k_i is the degree of node i, i.e., the number of neighbors, and E_i represents the actual number of connecting edges between adjacent nodes of node i.

The greater the CC of the network, the greater the cohesion of the whole network, and the higher the stability of the group in the network. The distribution of CCs in our study is shown in Fig. 1.

From the analysis shown in Fig. 1, we found that the CC of most of the users in this social network is 0, which indicates that most users' friends had no friend relationships with their neighbors (friends). When $C > 0$, the CC shows a bipolar trend, i.e., the proportion of users with lower CCs was similar to that of users with higher CCs, indicating that the relationships between the user's friends were either alienated or a close social circle. Similarly, we compared the CCs with those of online social networks, as shown in Table IV.

As shown in Table IV, compared with online social networks, the CCs of the communication social network are significantly lower. These communication networks, especially the CDRs network and the SMS network, are established based on the real connections among the users, and the establishment of the social relations in the networks is based on at least one call or short message. The establishment of online social networking is based on users followed by or following other users and on other online social activities, and its social relations do not mean that there are real social events, for example, the automated false "zombie fan" users and the many one-time registered users. Thus, the CC of online social networking is generally higher than that of the CDRs and SMS networks.

At the same time, the CDRs network's CC is higher than that of the SMS network, and both networks are based on the fact that social activity takes place. Thus, we can conclude that the CDRs' network is more likely to form a more intensive community. With the rapid development of online social networking, such as WeChat, Facebook, and microblogging, users can communicate more conveniently with each other. SMS networks use text information as the main carrier and have great substitutability compared with online social networks. In addition, in China, the 2012 SMS network aggregation coefficient [36] was in the range of 0.010–0.033, and in 2016, it was approximately 0.001. The SMS network's CC is decreasing. As a result, there are more structural holes in the SMS network than ever before, making users less likely to form dense communities.

In addition, the MPC network is established by the users' address books. In that MPC network, there are automated zombie users who have social relations but no social activities. That is, some of the contacts in a user's address book, although the user has a link with them, never engage in practical social activities. Therefore, the CC of the MPC network (0.043) is similar to that of online social networks, such as the Sin_m.

C. Assortativity Coefficient

In social networks, the ac measures the matching level between nodes. It is a Pearson correlation coefficient based on degree, that is

$$\mathrm{ac} = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i{}^2 + k_i{}^2) - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2} \tag{2}$$

where j_i and k_i, respectively, denote the degree of nodes connected by the ith edge.

In the complex network, if nodes with large degrees tend to connect with other nodes with large degrees, the network is

TABLE V
ACs OF SOCIAL NETWORKS

Social network	AC	Social network	AC
MPC	0.130	Ten_m	-0.609
CDRs	0.016	Twitter	–0.055
SMS	0.006	Facebook	0.266 [39]
Flicker	0.202 [38]	Sin_m	0.104

TABLE VI
AVERAGE LENGTHS OF SOCIAL NETWORKS

Social network	APL	Social network	APL
Ten_m	2.75	MPC	6.18
Sin_m	4.14	CDRs	6.59
Twitter	4.12	SMS	6.92
Facebook	4.7		

assortative. On the contrary, if nodes with large degrees tend to connect with nodes with small degrees, the network is disassortative. The ac values are generally between -1 and 1, and 0 indicates that there is no correlation in the network structure. A positive value indicates a synergic relation among nodes with the same node degree, and a negative value represents some kind of association between nodes with different node degrees. Most social networks are almost assortative, whereas other types of networks, such as technology networks, biological networks, and the Internet of Things, are disassortative. The ACs of different social networks are shown in Table V.

Table V shows that for these general social networks, the ac values are greater than 0, but for the Ten_m and Twitter, the ac values are less than 0. We analyzed the data set they used and found that the number of users was 1 940 000 and 150 000, respectively, and the number of records was 5.056 million and 1.314 million, respectively. There are many large "V" users (influential users whose accounts are stamped with the letter V to indicate their verified status) in online social networks, and V user nodes normally have high degrees. When the network is small, these V users tend to make the network disassortative. At the same time, online social networking also has many zombie and inactive users. These users attract much attention to increase their reputation; they also tend to make the network disassortative. We observed the degree distribution of the Ten_m and Twitter and found that there were more serious "fat tail" characteristics. This indicated that there were more V or zombie users in their research data sets, resulting in a lower ac, which makes the network appear disassortative.

From Table V, we can see that the communication social network in this article is assortative, which is in accordance with the law that most social networks are almost assortative.

D. Average Path Length

The distance between any two nodes in a network is defined as the path length, and the APL is the average path length, that is

$$\mathrm{APL} = \frac{2}{N(N-1)} \sum_{i \neq j} d_{ij} \tag{3}$$

where d_{ij} represents the number of edges contained in the shortest path from node i to node j.

In social networks, the shortest path characterizes the efficiency of information transmission between users. In addition, the shorter the shortest path, the higher the efficiency of the information transmission between users, and vice versa. The APL is an important indicator of the overall information dissemination efficiency of a social network. In fact, it is of great practical significance to study the APL of a network. For example, it can effectively reveal the necessity of controlling the spread of rumors and reducing their negative effect. It can also reveal information dissemination and exchange between organizations, such as associations and enterprises.

Studies have shown that although there are many nodes in an actual network, the APL of the network is very small [1] and is positively related to the network's number of nodes. We compared the average lengths of different social networks, as shown in Table VI.

In Table VI, we can see that the APL of the communication network is higher than that of the online social networks, such as Ten_m and Twitter. From this, we can conclude that the overall information transmission efficiency of online social networks is higher than that of the communication social network.

We further analyzed the reasons for the differences in APL between online social networks and communication social networks. Currently, the rise of online social networks has changed people's traditional lifestyles, becoming a platform for sharing life, values, experiences, and interests. Its dissemination efficiency far exceeds that of other media. The APL of online social networks is higher than that of communication social networks because of the following.

1) The number of users in online social networks is large, and the networks are open. The possibility of establishing relationships among users is high. However, factors, such as space time and social fear in communication social networks, affect the possibility of establishing social relationships between users.

2) In online social networks, users, such as V users and zombies, have an important effect on network connectivity, resulting in shortening of the shortest path between users. In communication social networks, such users are relatively few.

In this section, using the complex network analysis methods, we analyzed the structural characteristics of communication social networks and we derived the laws that exist in the structure of communication social networks.

IV. ANALYSIS OF USERS' SOCIAL BEHAVIORS BASED ON STATISTICS

In social networks, people's social behavior plays an important role in user research. In this article, we analyzed the users' social behaviors from aspects of age, gender, and time based on statistical methods, and summarized users' social behavior laws.

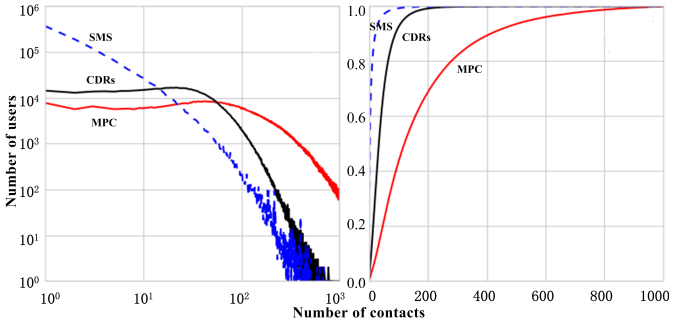

Fig. 2. Distribution of the number of contacts in the MPC, CDRs, and SMS networks. The x-axes represent the number of contacts and the y-axes are the number of users with the same contact size (or number). The right graph is the cumulative probability density of the number of contacts.

A. Users' Call Behaviors and Contacts Distribution

1) Users' Call Behaviors: In the CDRs network data, we defined the ratio of the number of calls a user has made to the total number of calls made and received as the user's callout ratio r_{out}, which can be expressed as $\mathrm{r_{out}} = (\mathrm{n_{out}}(i)/(\mathrm{n_{out}}(i) + \mathrm{n_{in}}(i)))$, where $n_{\text{out}}(i)$ and $n_{\text{in}}(i)$ are the number of calls made and the number of calls received in all CDRs for user i. The r_{out} value is in the range from 0 to 1, and if $r_{\text{out}} = 1$, this means that the user only makes calls and does not receive them, whereas $r_{\text{out}} = 0$ means that the user only receives calls, but does not make them.

In the CDRs networks, after removing the extreme r_{out} values of 0 and 1, the user callout ratio had a Gaussian distribution with mean μ, which, in this case, was approximately 0.5. Therefore, disregarding extreme values for call duration, the frequency of interactions among the social network's users was generally balanced. At the same time, there were some unbalanced social user interactions, with approximately 10% of users (i.e., those where r_{out} ⟨0.1 and r_{out}⟩ 0.9) having a callout ratio that was close to 0 or 1. This indicates that some users rarely made calls or rarely received them. We can conclude that there was an overall balance among users in the network for making and receiving calls. Therefore, for most users who use a cell phone, the number of calls made and received is close to equal.

2) User's Contacts: We analyzed and counted the number of user contacts in communication social networks, and the results are shown in Fig. 2.

As shown in Fig. 2, the MPC and CDRs networks of a certain contact size maintain a balance, and at greater than a certain scale (approximately 75–80) show a power law distribution, whereas the SMS network obeyed the power law distribution as a whole. The MPC was stable within 128 people, indicating that the total amount of social relationships that met the basic social requirements of users was approximately 128. Call contacts (or contacts in the CDRs network) reflected the actual interactions of users, and the number of call contacts remained constant within 55. This indicates that the size of a closely linked group was within 55 people, i.e., a certain size of the population constituted the user's core network members. As in the Marsden's core network theory [40], the main increase in social relations is not composed of core network members, but of the number of people who make the occasional contact. Individual users have many social relationships, but they interact with only some of them. Thus, the number of contacts in the MPC and CDRs networks remained stable at a certain scale, and then followed the power law distribution.

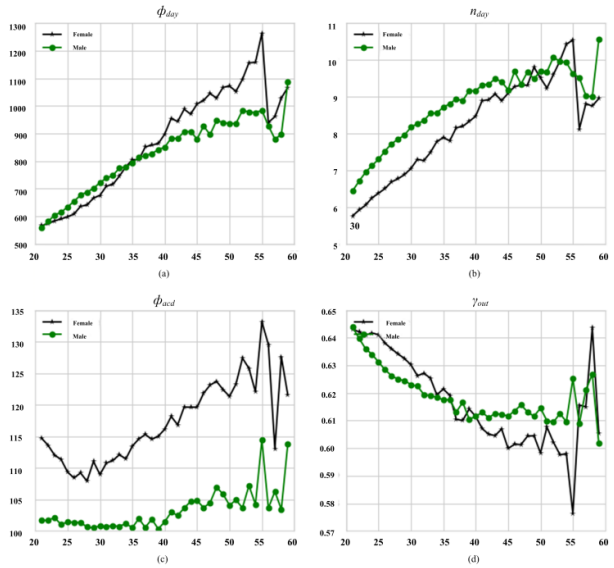

Fig. 3. Users' social behaviors in relation to age, gender, and call metrics. The x-axis represents the user's age and the y-axis represents the call metrics. (a) Average daily call duration Φ_{day} (in seconds). (b) Average number of daily calls n_{day}. (c) Average duration of each call Φ_{acd} (in seconds). (d) Callout ratio γ_{out}.

Compared with the CDRs network and the MPC network, the SMS network obeyed the power law distribution as a whole, and there was no stable stage in the early period. The reason is that with the continual development of instant messaging tools, such as WeChat and Facebook, the number of users who use SMS to communicate with each other and the frequency of SMS usage become less and less.

B. Differences in Age and Gender

As users get older, gender differences begin to become clear in their social behaviors with regard to making calls and using SMS. When we analyzed users' social behaviors, we examined the duration and frequency of user calls and obtained user changes and laws relating to them. The statistical analysis of the CDRs and SMS network data produced the results shown in Figs. 3 and 4.

From Figs. 3 and 4, we can see that the average daily call duration and frequency, and the monthly mean for short messages, were positively related to users' ages in these social networks. This shows that the frequency and quality of interaction among users increased with age. However, a user's callout ratio was negatively correlated with age. In other words, as a user's age increased, the proportional length of calls in relation to overall time spent in calls (whether making or receiving them) decreased. From the gender point of view, users with an average duration and number of messages

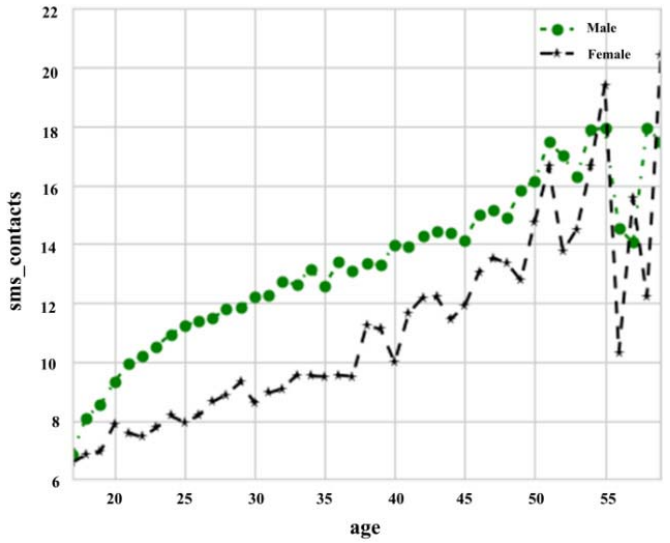

Fig. 4. Number of messages in relation to age and gender in the SMS network. The x-axis represents user ages and the y-axis represents the monthly mean for short messages.

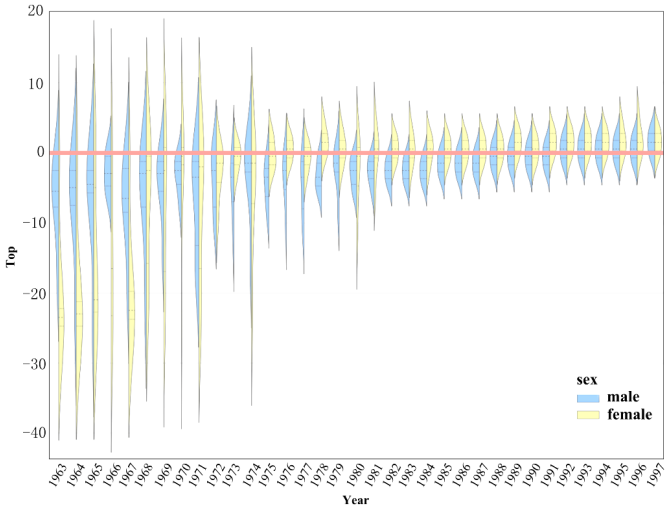

Fig. 5. Top distribution of the SAD for male (left side) and female (right side) users. The x-axis represents the year of birth for male and female users and the y-axis indicates the top distribution of the SAD.

are more heterogeneous, but women talk longer than men, and men send and receive more messages.

C. Social Patterns of Different Gender Groups

Although some researchers try to study people's real-world social patterns, because of information security, privacy protection, and other restrictions, data acquisition can be difficult. Accordingly, research on users' real-world social patterns often has the problem of small-data size and inaccurate conclusions. In addition, research on actual social patterns based on the big data environment is still basically nonexistent. We based our study of users' social patterns on mass cell phone users, so we were able to research users' real-world social patterns.

In communication social networks, there are usually age differences among connected users. In this article, we defined these age differences as social age difference (SAD). We analyzed the users' SADs in communication networks, and then studied social patterns of different user groups.

To analyze the social patterns for male and female groups, we analyzed user call data and obtained the SAD distribution for male (left side) and female (right side) groups at different ages, as shown in Fig. 5.

Fig. 5 shows that the top distribution approximately reflects the overall range of social interaction for users in that age group. It also shows that the peak for the top distribution approximately reflects the SAD for the people who have the most calls with users, so we found the following patterns.

1) As the age of users increased, the overall range of the SAD became larger, and the range of social contact also grew. At the same time, it is evident that users tend to talk to people younger than they are as they get older. Groups born between 1988 and 1998, i.e., aged between 19 and 29, were generally students or had just entered a more socially active phase. At this stage of life, most of the group contact was with classmates, colleagues, girlfriends or boyfriends, and other people of a similar age. Thus, the change in the top distribution at this stage is not very large, with the peak position fluctuating around 0. Groups born between 1963 and 1980 (between 37 and 54 years old) are generally at a critical point in their careers and generally have a wider range of social contacts. As a result, the range of the top distribution and changes within it are relatively large. From this, we conclude that as people get older (but are not yet elderly), and there is an increase in their life experience, the range of their social contacts will also increase.
2) We analyzed whether gender had an effect and found that the top distribution of the SAD for male and female users across different age groups was not the same. As their ages increased, the overall change in the SAD for female users (including the range and peak of top distribution) was greater than it was for male users. The peak of top distribution for the older female group (1963–1967) was significantly different from the peak for male users and young women, with a value of approximately -25. One explanation for this is that the main type of other users contacted by females in this age group was their children.

Therefore, we can see overall here that using differences in social patterns and the top distributions of the SAD, it is possible to infer a user's gender and age.

D. Day Differences in Users' Social Behaviors

To analyze differences in users' call behavior based on their age and time of day, we analyzed four different aspects of user call behavior, and the results are shown in Fig. 6.

From Fig. 6, we derived the following conclusions.

1) User call duration, frequency, and average duration of each call at various times in the day increase with age (x-axes), and have obvious differences at different times

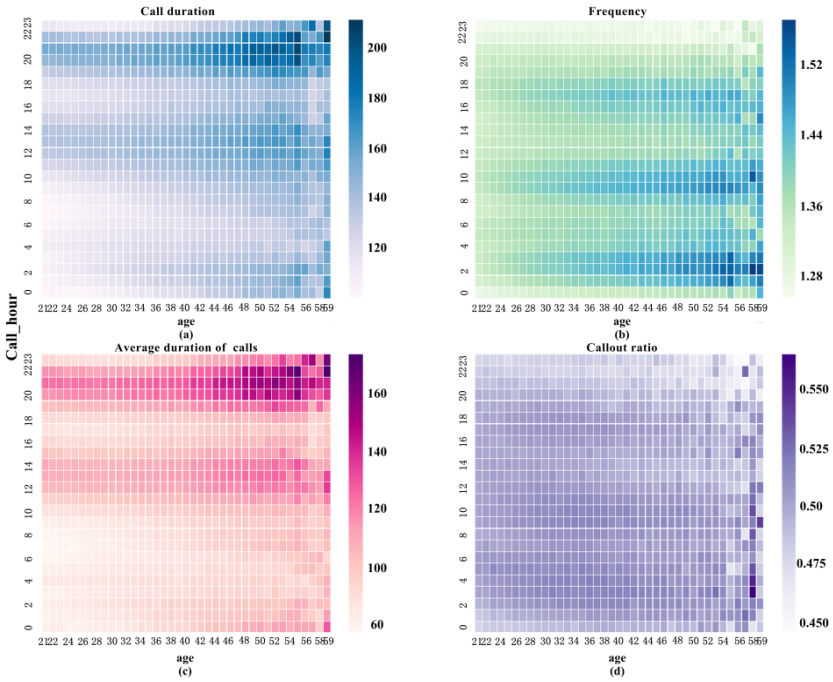

Fig. 6. User call behavior according to age, time of day, and call metrics. (a) Call duration. (b) Call frequency. (c) Average duration of each call. (d) Callout ratio. The x-axes represent the user's ages (ranging from 21 to 59) and the y-axes represent the hours in one day.

of day (y-axes). However, there was no major change in callout ratios with age.

2) User call duration and the average duration of each call were obviously higher in the 12:00–4:00 and 20:00–22:00 h, which indicates that users tended to have longer calls during nonworking hours. The user call frequency was significantly higher during three periods: 1:00–3:00, 9:00–11:00, and 16:00–18:00. All of those periods were working hours. Thus, we can conclude that people's call behaviors during working hours have high-frequency, low-duration, high-callout ratios, and during the rest time have low-frequency, high-duration, and low-callout ratios.
3) Analysis of the CDRs network shows that calls among core network members have long durations and low frequencies, and are usually at rest time. Calls with other members, such as colleagues or friends, show high frequency, short duration, and are usually during working hours.

In normal social activities, the groups that people socialize with during breaks are usually members of their core network. The core network is a relatively high-density network formed by individuals and neighbors in the network. McPherson *et al.* [41] suggest that some things may be talked about only with people who are more closely related. Although social actors have great differences in social attributes, they always hope to get help or comfort from those who have a close relationship with them, and these people form the core network.

V. Kinship Discrimination

To study the kinship user social patterns, we first identified users' kinship by the machine learning method. In this article, user kinship discrimination is mainly based on the CDRs data set and the SMS data set. The total amount of CDRs data set and users is much larger than the SMS data set. The kinship is mainly pulled from the SMS data set in this article. The CDRs data set and SMS data set are aggregated through impala tool in this article to obtain the common users.

We divided kinship into six categories, as well couples, paternity and siblings, cousins, sister-in-law, brother-in-law, and other relations (such as grandparents and grandchildren). Based on the SMS data, we determined whether the users have kinship relations through the information defined by kinship relations, such as appellation and identity. In the process of kinship user extraction, there is one of the relationships between users, and we think that the users have a kinship relationship. As a result, there are 147.4 million users with kinship. Moreover, we preprocessed these data sets according to the actual situation, consisting of outlier processing, data transformation, and class imbalance processing. Then, the experimental data set is obtained, including data sets with kinship and data sets without kinship.

A. Representation and Extension of User Call Behavior Characteristics

User's call behaviors can be represented by the ordered set Θ, and $\Theta = (u, v, t, d)$, where u and v are two users with their relative order indicating that user u calls user v, t is the start time of the call, and d is the call duration.

In fact, good data characteristics in machine learning can better train the model and thus obtain better classification results. In this article, feature extraction is based on call duration and call type. We extended the user behavior features to include the number of calls, when the calls were made, and

other dimensions. The various typical features are described below.

(1) *Total Number of Calls* $\mathbf{\Phi}_{\mathrm{nc}}$*:* The total number of calls between users can indicate the depth of the users' connections. We defined the total number of calls between user u and user v during one year as

$$\Phi_{\mathrm{nc}} = |\Theta(u, v)|. \tag{4}$$

(2) *Average Call Duration* $\mathbf{\Phi}_{\mathrm{acd}}(u, v)$*:*

The average call duration can indicate the average quality level of the connections between users. As for the call behaviors during one year, the greater the $\Phi_{\mathrm{acd}}(u, v)$, and the higher the quality of the connections between users. $\Phi_{\mathrm{acd}}(u, v)$ is defined as

$$\Phi_{\mathrm{acd}}(u, v) = \frac{\sum_{(u,v,t,d)\in\theta} d}{\Phi_{\mathrm{nc}}}. \tag{5}$$

(3) *Average Call Interval* $\mathbf{\Phi}_{\mathrm{aii}}(u, v)$*:*

Given users u and v, we can sort by call time. We assume here that (u, v, t_i, and d_i) and (u, v, t_{i+1}, and d_{i+1}) are two consecutive call behaviors, where $i \geq 1$, and we defined n as $n = |\Theta(u, v)|$. The average call interval Φaii (u, v) can be defined as

$$\Phi_{\mathrm{aii}}(u, v) = \frac{\sum_{i=1}^{n-1} (t_{i+1} - t_i)}{n1}. \tag{6}$$

The following definition gives the statistics for $\Theta(u, v)$ according to day, weekends (Saturday, Sunday), month, weekdays (Monday to Friday), and daily cycles.

(4) *Asymmetry of Call Behaviors* $\Phi_{\mathrm{ww}}(u, v)$*:*

$$\Phi_{\mathrm{ww}}(u, v) = \left| \frac{|\theta_{\mathrm{work}}(u, v)| - |\theta_{\mathrm{week}}(u, v)|}{|\theta(u, v)|} \right|. \tag{7}$$

To measure asymmetry between users, we standardized it by using $|\Theta(u, v)|$. If $\Phi_{\mathrm{ww}}(u, v) = 1$, it means that the call behaviors between users were extremely asymmetric, and calls occurred only on weekends or workdays. If $\Phi_{\mathrm{ww}}(u, v) = 0$, it means the call behaviors between users were symmetrical, and users' calls were evenly distributed among weekends and workdays.

(5) *Interaction Frequency* $\Phi_{\mathrm{eg}}(u, v)$*:*

The interaction frequency is used to measure the tendency between users to make a call, i.e., the frequency of calls between two users. It is defined as

$$\Phi_{\mathrm{eg}}(u, v) = \left| \frac{\Phi_{\mathrm{nc}}(T_i)}{T_i} \right|. \tag{8}$$

The interaction frequency measures the frequency of calls between two users in a certain duration T_i, and T_i can be days, weeks, or months.

(6) *Interaction Mode* $\Phi_{\mathrm{md}}(u, v)$*:*

The interaction mode measures the propensity between two users to engage in a certain call behavior. In other words, it is the law of calls between users. It is defined as

$$\Phi_{\mathrm{md}}(u, v) = 1 - \left| \frac{|\theta u(u, v)| - |\theta v(u, v)|}{\Phi_{\mathrm{nc}}} \right|. \tag{9}$$

The interaction mode measures the connection pattern between user u and user v. If $\Phi_{\mathrm{md}}(u, v) = 1$, the interaction between the users is symmetrical. If $\Phi_{\mathrm{md}}(u, v) = 0$, the interaction between them is highly asymmetric, i.e., just one of the two users always makes a call, and the other always receives it.

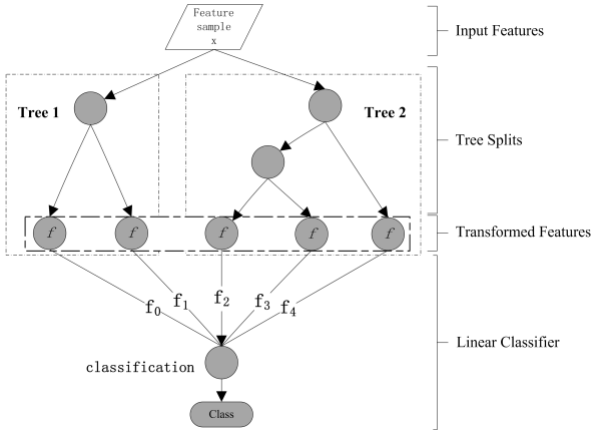

Fig. 7. Fusion method.

In this section, we have described six key features of user call behavior that were extracted from our CDRs data. Other features amount to extensions of these six features. We finally derived 75 user call features by extending and processing the limited call behaviors in CDRs.

B. Kinship Discrimination

LR is a common machine learning method, which is usually used to estimate the possibility of something. The problems that how to realize feature discovery and combination automatically to compensate for the lack of artificial experience are the problem encountered during the LR analysis. XGBoost is short for eXtreme Gradient Boosting package and a supervised learning algorithm that implements a process called boosting to yield accurate models. In this article, we can understand each tree of XGBoost as a feature conversion method with supervised cases, mapping the original feature to the leaf node in a certain logical discrimination. Each of the lifting trees in XGBoost can be regarded as a method of constructing features, and the characteristics generated by the supervised method have a distinction.

We regard the leaf nodes of boost tree as the characteristic variables of LR. As a result, the time to manually find features and combinations of features is greatly reduced. And, we got the Kinship-XL model to determine whether there is a kinship between users or not. The Kinship-XL model is established by the XGBoost and the LR fusion method. The processing process of the XGBoost and the LR fusion method, as shown in Fig. 7.

The XGBoost model gets Tree1 and Tree2 through learning, and the input sample x is traversed by two trees and forms the value on the leaf nodes of the two trees, respectively. If x falls on the second leaf node and the first leaf node of Tree1 and Tree2, respectively, the corresponding feature is [0, 1, 0, 1, 0]. The tree is a simple prediction model that represents a mapping relationship between sample data and target variables. Each node in the tree represents an object, and

TABLE VII
COMPARISON OF EIS

EI	DT	NB	SVM	LR	RF	XGBoost	Kinship-XL
Precision	0.6521	0.6897	0.7373	0.7236	0.7269	0.7332	0.8101
Recall	0.6562	0.6202	0.6853	0.7258	0.718	0.7116	0.7624
Accuracy	0.6554	0.6829	0.725	0.7264	0.7276	0.7314	0.7918
F1 score	0.6542	0.6527	0.7102	0.7247	0.7224	0.7222	0.7855
AUC	0.6554	0.6825	0.7251	0.7264	0.7276	0.7313	0.7918

each path represents some possible attribute value. The leaf node corresponds to the object value represented by the path from the root node. Each path is a differentiated path obtained by a method, such as minimization of the loss function. The features processed through this path are features that have been judged and integrated.

We need to find the raw feature and feature combination manually before the integration. After the integration of XGBoost and LR, we directly realized the automatic discovery of features and feature combinations through the XGBoost model. We choosed the XGBoost and the LR fusion method to identify the kinship, which is mainly to comprehensively consider the speed and accuracy of the kinship discrimination, and the feasibility of running large-scale data. This article adds the XGBoost model to the LR model, because it is necessary to solve the problem of nonlinear discriminating of LR. After the XGBoost model, the LR is to solve the speed of the algorithm or to apply to large scale. The combination of the two model is able to improve the performance and speed.

In the following process, we conducted experiments in the CDRs data set. We compared the output of our model with other traditional algorithms, and verified the effectiveness of the kinship discrimination model.

C. Experiment

To verify the effectiveness of the Kinship-XL model, we used a cross-validation approach, and $k = 10$ in this experiment. We used a variety of other evaluation indicators (EIs) to evaluate the kinship discrimination model and contrasted algorithms, including precision, recall, accuracy, F1 score, and area under the curve (AUC).

In addition, we used a variety of traditional classification algorithms for comparison. These consisted of decision tree (DT), naive Bayes (NB), support vector machine (SVM), LR, random forest (RF), and XGBoost. We compared the experimental results for each algorithm with the experimental results for our own Kinship-XL model and derived the evaluation index for every algorithm, as shown in Table VII and Fig. 8, sequenced according to their AUC values, which provides a good indication of their overall performance.

In Fig. 8, we can see that the Kinship-XL model has achieved good results, and the precision value is 81.01% and recall value is 76.24%. In addition, we can more intuitively observe the differences between the different methods. Among them, the precision of LR is 72.36%, which is 1.04% and 8.65% lower than that of the XGBoost model and Kinship-XL model, respectively. The precision of SVM is relatively high, which is 73.73%, but the recall rate is relatively low,

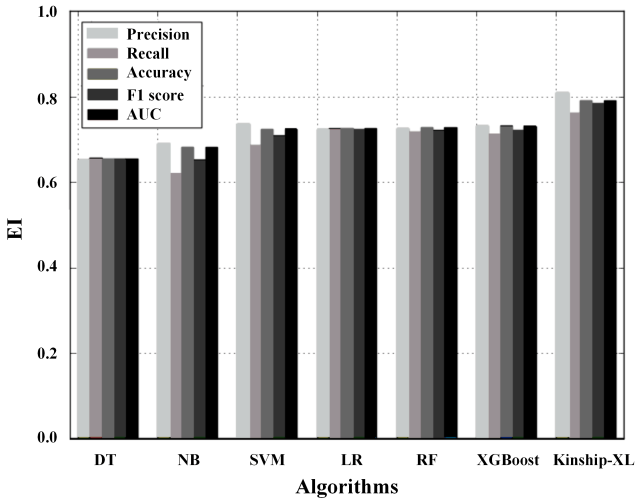

Fig. 8. Comparison of the kinship discrimination model in relation to other algorithms.

resulting in low AUC value. And, the Kinship-XL model is obviously better than these comparison algorithms. The Kinship-XL model enables the discrimination of the users' kinship in communication social network. Meanwhile, the kinship model in this article is based on big data, and its accuracy rate is higher than that of small-scale data.

In social networks, kinship is the most closely related social relationship with users. In the previous study of kinship, due to the limitations of research conditions, researchers often collected data in the form of questionnaires for research, and the sample size was not large enough. As a result, there are often problems with small data sizes and small samples. Currently, due to the rapid development of computers and information technology, the acquisition and analysis of large-scale social relationship data have become possible, which provides conditions for us to study and analyze kinship. The cell phone data adopted in this article has the characteristics of large amount of data and wide range. Therefore, it is more practical than the study of the small sample size.

VI. CONCLUSION

In the communication social networks, cell phone users gradually form the social networks with different intensities by means of mutual communication and exchange. We researched the communication social networks prevalent in people's lives based on the communication data of cell phone users. With the help of a large data analysis platform, we researched the communication social networks from the perspectives of

network structure, social behaviors, and user relationships by processing the data of large-scale cell phone users.

In the study of structural characteristics, we further analyzed the causes of differences between different social networks by measuring different structural characteristics and comparing current popular online social networks, and explored the impact of new social modes on traditional social modes. In terms of user behaviors, we analyzed the users' social behaviors from different angles, and explored the interesting social behavior patterns and the social differences between different groups, while it provides help to take a step forward to identify different groups. In the kinship discrimination, due to the limited basic characteristics of cell phone users, we constructed user characteristics and proposed a solution to quickly obtain discrimination results in large-scale data sets, namely Kinship-XL model. Which can get the result of kinship only through the user's calling behaviors, and also help to identify other relationships among cell phone users. In addition, our future work will include recognition of user attributes, the laws of information dissemination, public opinion analysis, and application in real life.

REFERENCES

[1] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003.

[2] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove, "Mobile call graphs: Beyond power-law and lognormal distributions," *ACM SIGKDD*, Las Vegas, NV, USA, Aug. 2008, pp. 596–604.

[3] H. Zhang and R. Dantu, "Predicting social ties in mobile phone networks," in *Proc. IEEE Int. Conf. Intell. Secur. Informat.*, May 2010, pp. 25–30.

[4] S. Motahari *et al.*, "The impact of social affinity on phone calling patterns: Categorizing social ties from call data records," in *Proc. SNA KDD Workshop*, 2012.

[5] L. Rokach, M. Kalech, I. Blank, and R. Stern, "Who is going to win the next Association for the Advancement of Artificial Intelligence Fellowship Award? Evaluating researchers by mining bibliographic data," *J. Amer. Soc. Inf. Sci.*, vol. 62, no. 12, pp. 2456–2470, Dec. 2011.

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

[7] M. Lin and W.-J. Hsu, "Mining GPS data for mobility patterns: A survey," *Pervasive Mobile Comput.*, vol. 12, pp. 1–16, Jun. 2014.

[8] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, "Computer assisted customer churn management: State-of-the-art and future trends," *Comput. Oper. Res.*, vol. 34, no. 10, pp. 2902–2917, Oct. 2007.

[9] X. Zhou, X. Liang, X. Du, and J. Zhao, "Structure based user identification across social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1178–1191, Jun. 2018.

[10] L. Rainie and B. Wellman, *Networked: The New Social Operating System*, vol. 53, no. 1. Cambridge, MA, USA: MIT Press, 2012, pp. 203–204.

[11] A. Rapoport, "Contribution to the theory of random and biased nets," *Bull. Math. Biophys.*, vol. 19, no. 4, pp. 257–277, 1957.

[12] J. Travers and S. Milgram, "An experimental study of the small world problem," *Sociometry*, vol. 32, no. 4, pp. 425–443 1969.

[13] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442 1998.

[14] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.

[15] R. Dunbar, "Grooming, gossip and the evolution of language," *J. Hist. Behav. Sci.*, vol. 34, no. 4, pp. 398–399, 1998.

[16] R. Cohen, S. Havlin, and D. Ben-Avraham, "Structural properties of scale-free networks," in *Handbook Graphs Networks*. 2002, pp. 85–110.

[17] H. Fei, R. Jiang, Y. Yang, B. Luo, and J. Huan, "Content based social behavior prediction: A multi-task learning approach," in *Proc. Conf. Inf. Knowl. Manage. (CIKM)*, Glasgow, U.K., 2011, pp. 995–1000.

[18] X. Zhou, B. Wu, and Q. Jin, "User role identification based on social behavior and networking analysis for information dissemination," *Future Gener. Comput. Syst.*, vol. 96, pp. 639–648, Jul. 2019, doi: 10.1016/j.future.2017.04.043.

[19] S. Han and X. Yan, "Friend recommendation of microblog in classification framework: Using multiple social behavior features," in *Proc. Int. Conf. Behav., Econ. Social Comput.*, Nanjing, China, Oct. 2015, pp. 1–6.

[20] M. Nemiche, V. Cavero, and R. P. Lopez, "Understanding social behavior evolutions through agent-based modeling," in *Proc. Int. Conf. Multimedia Comput. Syst. (ICMCS)*, Tangiers, Morocco, 2012, pp. 980–988.

[21] K. K. Roudposhti, U. Nunes, and J. Dias, "Probabilistic social behavior analysis by exploring body motion-based patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1679–1691, Aug. 2016.

[22] Y.-G. Zhao, Z. Song, F. Zheng, and L. Shao, "Learning a multiple kernel similarity metric for kinship verification," *Inf. Sci.*, vols. 430–431, pp. 247–260 Mar. 2018.

[23] H. Yan, J. Lu, W. Deng, and X. Zhou, "Discriminative multimetric learning for kinship verification," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 7, pp. 1169–1178, Jul. 2014.

[24] K. Zhang *et al.*, "Kinship verification with deep convolutional neural networks," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Swansea, U.K., 2015, pp. 148.1–148.12.

[25] A. Puthenputhussery, L. Qingfeng, and L. Chengjun, "SIFT flow based genetic Fisher vector feature for kinship verification," in *Proc. Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, 2016, pp. 2921–2925.

[26] X. Zhou, H. Yan, and Y. Shang, "Kinship verification from facial images by scalable similarity fusion," *Neurocomputing*, vol. 197, pp. 136–142, Jul. 2016.

[27] M. Xu and Y. Shang, "Kinship verification using facial images by robust similarity learning," *Math. Problems Eng.*, vol. 2016, pp. 1–8, Jan. 2016.

[28] H. Yan, "Kinship verification using neighborhood repulsed correlation metric learning," *Image Vis. Comput.*, vol. 60, pp. 91–97, Apr. 2017.

[29] E. Boutellaa *et al.*, "Kinship verification from videos using spatio-temporal texture features and deep learning," in *Proc. Int. Conf. Biometrics*, Halmstad, Sweden, 2016, pp. 1–7.

[30] B. Patel, R. Maheshwari, and B. Raman, "Evaluation of periocular features for kinship verification in the wild," *Comput. Vis. Image Understand.*, vol. 160, pp. 24–35, Jul. 2017.

[31] R. Lambiotte *et al.*, "Geographical dispersal of mobile communication networks," *Phys. A, Stat. Mech. Appl.*, vol. 387, no. 21, pp. 5317–5325, 2008.

[32] H. Xia, "Research on mining structure and behaviors in mobile social networks," (in Chinese). Ph.D. dissertation, Univ. Electron. Sci. Technol., Chengdu, China, 2012.

[33] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in Facebook," in *Proc. 2nd ACM Workshop Online Social Netw.* Barcelona, Spain, 2009, pp. 37–42.

[34] A. Java, S. Xiaodan, T. Finin, and B. Tseng, "Why we Twitter: Understanding microblogging usage and communities," in *Proc. 9th WebKDD 1st SNA-KDD Workshop Web Mining Social Netw. Anal.*, San Jose, CA, USA, 2007, pp. 56–65.

[35] A. A. Nanavati *et al.*, "On the structural properties of massive telecom call graphs: Findings and implications," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, Arlington, VA, USA, 2006, pp. 435–444.

[36] J. Bollen, B. Gonçalves, G. Ruan, and H. Mao, "Happiness is assortative in online social networks," *Artif. Life*, vol. 17, no. 3, pp. 237–251, Jul. 2011.

[37] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. 19th Int. Conf. World Wide Web* Raleigh, NC, USA, 2010, pp. 591–600.

[38] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. Internet Meas. Conf. (IMC)*, San Diego, CA, USA, 2007, pp. 29–42.

[39] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the Facebook social graph," 2011, *arXiv:1111.4503*. [Online]. Available: https://arxiv.org/abs/1111.4503

[40] P. V. Marsden, "Core discussion networks of Americans," *Amer. Sociol. Rev.*, vol. 52, no. 1, p. 122, Feb. 1987.

[41] M. McPherson, M. E. Brashears, and L. Smithlovin, "Social isolation in America: Changes in core discussion networks over two decades," *Amer. Sociol. Rev.*, vol. 71, no. 3, pp. 353–375 2006.

Shu-Sen Zhang is currently pursuing the Ph.D. degree in computer engineering with the School of Information, Renmin University of China, Beijing, China.

His current research interests include data mining and social computing.

Xun Liang (Senior Member, IEEE) received the Ph.D. degree.

From 1993 to 1995, he was a Post-Doctoral Fellow with the Institute of Computer Science, Peking University, Beijing, China. He is currently a Professor of economic information management with the Renmin University of China, Beijing. His current research interests include Internet information analysis, data mining, business intelligence, and social computing.

Yu-Dang Wei received the master's degree in computer engineering from the School of Information, Renmin University of China, Beijing, China.

His current research interests include data mining and social computing.

Xuan Zhang is currently pursuing the Ph.D. degree in computer engineering with the School of Information, Renmin University of China, Beijing, China.

His current research interests include data mining and social computing.