PAPER Special Section on Data Engineering and Information Management

NFRR: A Novel Family Relationship Recognition Algorithm Based on Telecom Social Network Spectrum

Kun NIU[†a)], ***Nonmember***, **Haizhen JIAO**[†], ***Student Member***, **Cheng CHENG**[†], **Huiyang ZHANG**[†], ***and*** **Xiao XU**[†], ***Nonmembers***

SUMMARY There are different types of social ties among people, and recognizing specialized types of relationship, such as family or friend, has important significance. It can be applied to personal credit, criminal investigation, anti-terrorism and many other business scenarios. So far, some machine learning algorithms have been used to establish social relationship inferencing models, such as Decision Tree, Support Vector Machine, Naive Bayesian and so on. Although these algorithms discover family members in some context, they still suffer from low accuracy, parameter sensitive, and weak robustness. In this work, we develop a Novel Family Relationship Recognition (NFRR) algorithm on telecom dataset for identifying one's family members from its contact list. In telecom dataset, all attributes are divided into three series, temporal, spatial and behavioral. First, we discover the most probable places of residence and workplace by statistical models, then we aggregate data and select the top-ranked contacts as the user's intimate contacts. Next, we establish Relational Spectrum Matrix (RSM) of each user and its intimate contacts to form communication feature. Then we search the user's nearest neighbors in labelled training set and generate its Specialized Family Spectrum (SFS). Finally, we decide family relationship by comparing the similarity between RSM of intimate contacts and the SFS. We conduct complete experiments to exhibit effectiveness of the proposed algorithm, and experimental results also show that it has a lower complexity.

key words: *telecom social networks, relational spectrum matrix, specialized family spectrum*

1. Introduction

With the increasing use of cellphone, people contact with each other more frequently and these communications form a reflection of real-life social network. How to identify different semantics of social associations has been a research hotspot in recent years. A series of algorithms have been proposed for identifying diverse relationships, such as advisor-advisee [1], [2], co-workers [2]–[6], friends [2]–[4], [7], [8], families [6], [9], and so on [3], [6]. Among various social ties, identifying family relationship accurately holds a key to retain existing customers as well as developing new customers in telecom industry. However, current solutions still have many limitations. The researches focus on applying standard classification methods of data mining, whose performances are limited to the effect of classifiers. On the other side, the rapid growth of data also poses a huge challenge on algorithm usability. Therefore, designing an effective algorithm for family relationship recognition in telecom social network is a valuable and challenging work.

1.1 Related Work

Previous researches, which recognized relationships among users, usually obtain labeled data from Over The Top (OTT) application or online social networks, such as Facebook, Twitter, Hotmail and so on. For example, Chi Wang et al. [1] proposed a time-constrained probabilistic factor graph model (TPFG) to mine advisor-advisee relationship based on a computer science bibliographic network. Combined with several basic social psychological theories, Jie Tang et al. [10] established a transfer-based factor graph (TranFG) model to predict social relationships in a network by borrowing knowledge from a different network. Xiao Han et al. [7] also built a social P2P network model based on the empirical analysis.

With the popularity of smartphones, the temporal and spatial information of user is easy to be collected. More and More studies recognize relationships via this. Lauw et al. [11] focused on the behavior that two or more users are collocating around the same time. They noticed these behaviors imply that there might be some association among them. Zhou et al. [12] proposed a multi-context trajectory embedding model to conduct social link prediction. Hongjian Wang et al. [13] proposed a framework which unify personal, global and temporal factors to measure the relationship between two given mobile users. Hsun-Ping Hsieh et al. [14] propose a two-phase prediction method for the social inference using mobile sensor data.

However, above studies ignored the data about call and SMS which the telecom operators are very concerned about. Jinhyuk Choi [5] and Jun-Ki Min et al. [6] mined social relationship types by co-location data and instant messenger data. But in these researches, the data set is too small. Ronghui Liu et al. [9] divided users into several family groups by using logistic regression algorithm to recognize family relationship. But its success rate is not high enough. Even though Jie Tang et al. [10] analyzed the unlabeled relationship in the mobile social network, the method proposed still require to get the knowledge from other labeled network.

Today, there are more and more researches that study the relationship among users in social network. But they

Manuscript received June 20, 2018.
Manuscript revised October 30, 2018.
Manuscript publicized January 11, 2019.
[†]The authors are with the School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China.
a) E-mail: niukun@bupt.edu.cn
DOI: 10.1587/transinf.2018DAP0008

didn't pay much attention to mobile network because telecom operators always hold their customer information as both secret and estate and never share it out. Thus many researches, which perform mobile network analysis, recruit countable volunteers to get their communication data. Their results are hard to prove trustworthy on existing complicated telecom network because of the high possibility of overgeneralization.

1.2 Our Contributions

In this paper, we propose a new algorithm named NFRR and our contributions in this research can be summarized as follows.

1. We propose the concept of **Relational Spectrum Matrix (RSM)** and **Specialized Family Spectrum (SFS)**. For correlating the communication behavior and relationships effectively, we set up a 3-D feature space, a matrix as the relational spectrum, which based on the temporal and spatial features of communication records and the corresponding patterns with contacts, to reflect the relationship between users.
2. We propose a NFRR algorithm based on the relational spectrum. The algorithm identifies family relationships accurately from the users' intimate contacts.
3. Our algorithms have been extensively tested through experiments on integrated datasets with satisfactory results. The proposed algorithm exhibit both the effectiveness and robustness, and has a lower complexity.

The rest of this paper is organized as follows. In Sect. 2, we state the main idea of NFRR, and give notions and definitions. Section 3 shows the whole process of NFRR with pseudo-code implementations. Experimental results are presented in Sect. 4. Finally, we conclude our work in Sect. 5.

2. Family Recognition by Relational Spectrum Matrix

2.1 Notions

Given a graph $G = \{U, E, F\}$, $U = \{u_1, \ldots, u_n\}$ is the set of vertexes, both $E = \{e_{ij}\}$ and $F = \{f_{ij}\}$ are sets of edges. Each edge e_{ij} connects the vertexes pair $<u_i, u_j>$ if and only if they have communicated in definite time period.

$$e_{ij} = \begin{cases} 1, & \text{when } u_i \text{ and } u_j \text{ have communicated} \\ 0, & \text{when } u_i \text{ and } u_j \text{ have not communicated} \end{cases} \tag{1}$$

There is another kind of edge f_{ij} connects vertexes pair $<u_i, u_j>$ if and only if they are regarded as family in a definite time period. Of course, vertexes refer to customers of mobile network in business way.

$$f_{ij} = \begin{cases} 1, & \text{when } u_i \text{ and } u_j \text{ are family} \\ 0, & \text{when } u_i \text{ and } u_j \text{ are not family} \end{cases} \tag{2}$$

For any u_i, its metadata can be represented as a 3-tuple set:

$$u_i = \{T, S, C\} \tag{3}$$

- T is temporal feature of all telecom records those u_i produced, which is filled with timestamps.
- S is spatial feature of all telecom records about u_i, which contains Cell ID in mobile network.
- C represents and describes communication behavior u_i. This subset has more than one attribute.

Definition 1. ***Time Division.***

$$t = \begin{cases} ST(\textit{Sleep Time}), & (22 \sim 24] \,\&\, (0 \sim 6]\,; \\ WT(\textit{Work Time}), & (9 \sim 12] \,\&\, (14 \sim 18] \\ & \textit{in work day;} \\ LT(\textit{Leisure Time}), & \textit{all the other time.} \end{cases} \tag{4}$$

Here, we divide each week into three time segments: sleep time, work time and leisure time. Thus, we can identify a user's locations of home and workplace based on the happen of communications.

Definition 2. ***Home Location.***

$\forall u_i \in U$, home location $S_{i,home}$ is defined as follows:

- $S_{i,home} \in u_i.S$, that is, $S_{i,home}$ is a Cell ID of base station, not real GPS position in a sense of geography.
- $S_{i,home} = u_i.S.mode\,(ST)$, where $mode(ST)$ is the base station ID which appears most frequently in ST.

Definition 3. ***Workplace.***

$\forall u_i \in U$, workplace $S_{i,work}$ is defined as follows:

- $S_{i,work} \in u_i.S$, $S_{i,home}$ is also a Cell ID.
- $S_{i,work} = u_i.S.mode\,(WT)$, where $mode(WT)$ is the base station ID which appears most frequently in WT.

Definition 4. ***Family Relationship.***

Any user pair $<u_i, u_j> \in U$ are taken as having family relationship when meeting all the following conditions:

- $e_{ij} = 1$ in WT and LT;
- $e_{ij} = 0$ in ST;
- $S_{i,home} = S_{j,home}$.

To this extent, families refer to people who live together. Concretely, roommates would be alse regarded as a family relationship, but family members who live separately would not.

Definition 5. ***Relational Spectrum Matrix (RSM).***

For any user pair $<u_i, u_j> \in U$, we define the RSM M_{ij} as a $t \times p$ matrix shown in Eq. (5), where t is the number of time segments, and p is the number of key positions. The key positions are also denoted as Cell ID, showing one's POI (points of interest). POIs are the most popular location who like to be, including home and workplace.

$$M_{ij} = \begin{pmatrix} m_{11} & \cdots & m_{1p} \\ \vdots & \ddots & \vdots \\ m_{t1} & \cdots & m_{tp} \end{pmatrix} \tag{5}$$

Specially, RSM has two types, call duration and call times. In call duration RSM, each value m_{xy} means the total call durations between $<u_i, u_j>$ in time segment x and position y. For call times RSM, m_{xy} represents the frequency they call in the same condition.

Definition 6. ***Specialized Family Spectrum (SFS).***

For each user $u_i \in U$, we define the SFS $\mathcal{M}_i$ of u_i as the weighted matrixes of its K most similar neighbors, which can be calculated by Eq. (6).

$$\mathcal{M}_i = \sum_{j=1}^{K} W_{ij} \times \mathcal{M}'_j \tag{6}$$

Here, W_{ij} is defined as the similarity between u_i and its j^{th} similar neighbor, higher similarity brings higher weight. We use function $sim(u_i, u_j)$ to measure similarities and its detailed introduction is given in 2.2. A transformation is given in Eq. (7) to make sure $\sum_{j=1}^{K} W_{ij} = 1$.

$$W_{ij} = \frac{sim(u_i, u_j)}{\sum_{j=1}^{K} sim(u_i, u_j)} \tag{7}$$

For any similar users u_j, $\mathcal{M}'_j$ represents the average RSM of u_j and its l labelled family, which can be calculated by Eq. (8). Here L is the total number of u_j's family.

$$\mathcal{M}'_j = \frac{\sum_{l=1}^{L} M_{jl}}{L} \tag{8}$$

2.2 The NFRR Algorithm

To judge whether two users are in a family, a general idea is to learn communication pattern from the known family members. Families always keep close connection in daytime and face to face at night. Therefore, the puzzle is separated into three sub-problems.

- First, how to mining user's communication patterns. Here, we proposed Relational Spectrum Matrix (RSM) to mining user's patterns from 3 directions: temporal, spatial and communications.
- Second, what kind of RSM can be identified as a family relationship. Here, we use several user-pairs that has been labelled as "family" to establish the Specialized Family Spectrum (SFS) matrix.
- Last, how to measure similarities between the unknown user-pairs and the family-labelled user-pairs for deducing and judging.

2.2.1 RSM - Mining Two Users' Communication Patterns

As Definition 5 mentioned, it is a $t \times p$ matrix, where t represents different time segments and p represents different positions. RSM mines users' patterns from three directions: temporal, spatial features of communication and contact patterns.

Temporal Features: In RSM, one day is divided into 3 main segments WT, LT and ST as Definition 1 shows. Easily to see that users' communications show a strong correlation in time direction. Assume that all contacts can be simply classified as $Family$, $Friend$, $Colleague$ and $Stranger$. Generally, communications with families most concentrate on WT and they definitely have few connections at ST for staying together at night. Besides, families also have a contact peak on their way home from work. Colleagues work at the same place and most contacts may appear in non-work time, however colleagues in different offices may also have contacts in WT. Friends usually contact for going out and call most centralize on LT.

Spatial Features: In RSM, two key positions are used, $Home$ and $Workplace$. Generally, families have less contacts at ST and colleagues in the same office have less contacts at WT because they can exchange idea face to face. For identifying home or workplace locations, we first select most frequent base stations where the target user appears in ST and WT, and then use Eq. (9) to select the final credible locations.

$$f(s) = \frac{\sum_{i=1}^{N} \frac{c_{si}}{C_i}}{N} \tag{9}$$

where $N \in \{28, 29, 30, 31\}$ represents the days of each month.

For home location identifying, C_i counts the total times user appeared at sleep time in i^{th} day, c_{si} represents the times user appeared at base station at sleep time in i^{th} day. Base station with the biggest $f(s)$ will be user's home. Similarly, when identifying workplace, C_i and c_{si} represents the records produced at WT in i^{th} day.

Communications: Each user-pair has two types of matrixes: call duration and call times. Communications with different contacts also poses different features. Time spent on families are significantly higher than other relationships, total call times among them always keep a higher level, but average length of each call is quite short. Oppositely, calls between colleagues are mainly relative to their work, and always take longer durations to discuss detail of business.

Above all, we can present each user pair's communication pattern via Relational Spectrum Matrixes. Different relationships lead to different matrixes, and a typical example is given in Fig. 1. Figure 1 (a) represents a family matrix. Calls rarely appear on ST or when they are detected at HL. Figure 1 (b) shows a friend matrix, which contacts always happen in LT. Figure 1 (c) is from colleagues, which have less contacts at WT or when they are detected at WL. Then, we only need to find RSM which shows family characteristics for judging families.

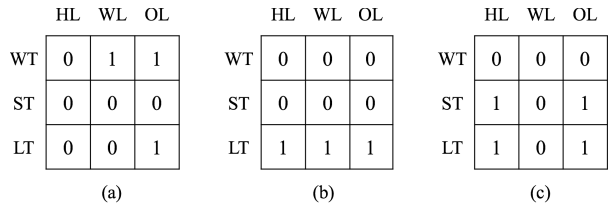

Fig. 1 An example of RSMs for different relationships. (a) represents a family matrix, (b) represents a friend matrix, and (c) represents a colleague matrix. Value '0' means that very few records appear at the corresponding time and location, '1' means more records are produced.

2.2.2 SFS - Describe User's Standard Family Patterns

In real life, it's quite difficult to define a unified standard matrix to represents the family communication pattern because people have very different ways in communicating with families. For example, roommates in school, who usually have similar schedules, are regarded as a family relationship according to Definition 4. Therefore, communications among them mainly focus on lunch or dinner time. That's quite different for conditions that people have full-time jobs, which means there are rare contacts during lunch. However, similar users have similar behaviors and show a mutual special spectrum, for example, student group shows a similar pattern and full-time job group shows another similar pattern. For these reasons, we can find similar users and design a specialized standard matrix for them. Therefore, for each user, we select K most similar users from the last billing cycle as its 'nearest neighbors', and then weight RSM of the K users to approximate that user's SFS to represent its standard communication patterns with its family.

To pick up K most similar users, we select several attributes and calculate the similarities between any two users by Eq. (10), where R marks the total number of attributes. Here, the Manhattan distance are used, which possesses a faster real-time response in practice than other distance measure functions. Larger distance means less similarity between two customers. The list of attributes used in our algorithms is shown in Table 1.

$$sim\left(u_i, u_j\right) = \frac{1}{\sum_{r=1}^{R} \left|u_{ir} - u_{jr}\right|} \tag{10}$$

Additionally, all these attributes have to be normalized before calculation.

Above all, we obtain user's RSM with its contacts and its SFS. We can easily make the judgment by comparing the two matrixes. The probability of being families depends on how similar the two matrixes are.

2.2.3 Similarities Measurement and Relationship Judgment

Through above work, we finally get two matrixes, RSM M_{ij} and SFS $\mathcal{M}_i$. Then, the puzzle of judging families is perfectly transformed to measuring similarity between two matrixes, as shown in Eq. (11).

Table 1 Attributes list for similarity measurement.

No.	Attributes	Title	Description
1	Total Duration	TD	Sum of call duration in a month
2	Total Calling Times	TCT	Frequency of calls in a month
3	Long-Distance Calling Times	LDCT	Frequency of long-distance calls in a month
4	WT Calling Times	WCT	Frequency of calls in work time
5	ST Calling Times	SCT	Frequency of calls in sleep time
6	LT Calling Times	LCT	Frequency of calls in leisure time
7	Workday Calling Times Proportion	WCTP	The proportion of calls in workdays
8	Workday Calling Duration Proportion	WCDP	The proportion of call duration in workdays
9	Number of Points	NP	Number of base stations in a month
10	Top-3 Call Times Concentration	TCTC	Call times concentration of top 3 contacts
11	Top-3 Call Duration Concentration	TCDC	Call duration concentration of top 3 contacts
12	Monthly Fee	MF	Average Fee Paid
13	Workplace Location	WL	GPS location of workplace
14	Home Location	HL	GPS location of home
15	Commute Distance	CD	Distance between workplace and home

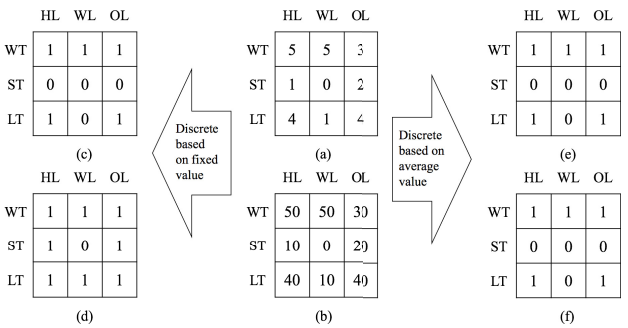

Fig. 2 Different discretization methods.

$$sim\left(\mathcal{M}_i, M_{ij}\right) = \frac{1}{\sum_{x=1}^{p} \sum_{y=1}^{t} \left|\mathcal{M}_{i\,xy} \oplus M_{ij\,xy}\right|} \tag{11}$$

For a given threshold α, if it satisfied $sim\left(\mathcal{M}_i, M_{ij}\right) \geq \alpha$, NFRR will label the user pair $<u_i, u_j>$ with families and assign $f_{ij}.value = 1$.

To reduce the calculation complexity and balance comparability and accuracy, all matrixes are discretized into normalization matrixes. Instead of the traditional discretization based on fixed threshold, we use average-based discretization to reduce the effect caused by different absolute usage amount. As shown in Fig. 2, (a) and (b) are two RSMs, discretize (a) based on 3 and then we obtain (c), discretize (b) with 15 and then we obtain (d). On the other hand, (e) and (f) are average-based results. Clearly, (a) and (b) have a similar behavior which can be inferred by (e) and (f), however (c) and (d) reflect a big bias that may affect the performance of NFRR.

3. Algorithm Descriptions

The whole flow of NFRR is shown in Fig. 3, which mainly includes four phases: data prepare, RSM generating, SFS generating, and family relationship judging. The current results will be used in the third phase in the next time cycle.

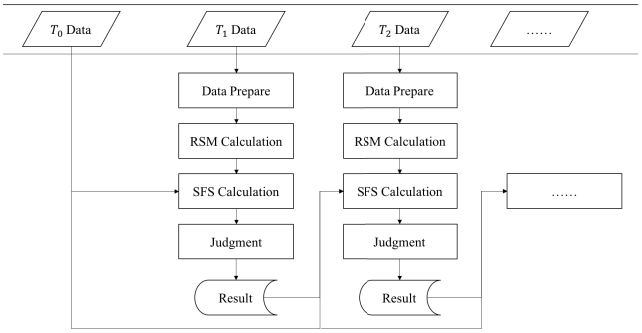

Fig. 3 Flow of NFRR.

3.1 Data Prepare

NFRR deals with detail data of telecom network. These details describe user's each consuming process, which consist of IMSI, IMEI, in-bound or out-bound identifier, timestamp, duration, fee and so on. They also contain location information. Generally, each cellphone sends a message of GPS automatically every half an hour when in idle state, and sends continuously when busy.

Certainly, data pre-processing is necessary. It transforms available data into a suitable and processable format. Specific techniques such as data cleaning, data integration, missing value imputation are executed in sequence. Here, we delete records less than three seconds because they make no sense. Besides, duplicated records are also cleaned.

3.2 NFRR Details

Firstly, aggregating process via map-reduce functions divide the dataset into two parts, Telecom Records (TR) and Location Based Records (LBR). For each user u_i, TR data is its detailed correspondence bills consists of IMSI/IMEI, in-bound or out-bound identifier, call-time, total fee and so on, LBR data is users' location information consists of Cell-ID, GPS and so on. NFRR selects top high frequency base stations at ST/WT as candidates. Then, calculates $f(s)$ using Eq. (9), base station with the biggest $f(s)$ is regarded as home/work location. TR data is used to select top call duration and top call frequency contacts, other unselected contacts are labeled with not families. These top contacts are family candidates. We suggest top five call duration and call frequency contacts independently, but the total ten people may overlap each other and only generate five persons totally in extreme case.

RSM Generating: After getting user u_i and its intimate contacts, NFRR generates call duration relational spectrum matrixes for top five contacts, and also generates call frequency relational spectrum matrixes for another top five contacts. All matrixes are discretized into normalized matrix based on average value.

SFS Generating: For any user u_i, NFRR selects top K similar users using Eq. (8) from last cycle, and generates the SFS $\mathcal{M}_i$ using Eqs. (6), (7), (8) and (10). Finally, get $sim\left(\mathcal{M}_i, M_{ij}\right)$ using Eq. (11).

Family Relationship Judging: After comparing similarities between $\mathcal{M}_i$ and M_{ij}, the user pair $<u_i, u_j>$ which satisfies $\left|sim.\left(\mathcal{M}_i, M_{ij}\right) - \alpha\right| < \epsilon$ will be labeled as family relationship and assign $f_{ij}.value = 1$.

NFRR Initialization: For the first NFRR, it needs a known family data set to calculate the SFS. The data set is available by two ways: 1. Questionnaire. Questionnaires are randomly sent to customers to ask whether two persons are families. 2. Family package records. There are many family packages, products and activities that the operator produced especially for family members, which records customers' information including the family members.

Dynamic Updating: Since the relationships between K neighbors and their own contacts are given, we can easily pick out their families and take RSM to calculate their SFS individually. Considered that user's communication behavior may change as time goes on, the K neighbors list is also dynamic. It is quite possible that the user itself is also selected. In NFRR, any user is different from itself in other billing cycles. Owing to the data produced periodically, NFRR runs once in a billing cycle. The known family relationship from last billing cycle are used as labels for current process.

Algorithm 1 NFRR

```
DataPrepare (data);
for u_i in U = {u_1, ..., u_n} do
    /** RSM Generating*/
    identifyHomeLocation ();
    identifyWorkLocation ();
    CDUList = selectTopCallDurationContact ();
    CTUList = selectTopCallFrequencyContact ();
    for u_j in {CDUlist, CTUlist} do
        matrixes. Add (generate Matrix ());
    end for;
    discretize(matrixes);
    /** SFS Generating*/
    SMList = selectSimilarUser(u_i, K);
    M_i = calculateMI (SMList);
    /** Relationship Judgement */
    for M_ij in in matrixes do
        sim = similarityCalculate(M_i, M_ij);
        if sim ≥ α then
            labelEdge (1);
        else
            labelEdge (0);
        end if
    end for;
end for;
```

4. Experiments and Results

In this parts, we introduced the experiments and results. Firstly, we present the dataset, environment configurations, and evaluation measurements. To show the effectiveness of NFRR algorithm, we conducted comparison experiments with two baseline algorithms, Support Vector Machine

(SVM) and Random Forest (RF). Since NFRR has two parameters, we conducted several experiments to discuss results with different parameter settings. At last, we analyzed the time complexity of NFRR.

4.1 Empirical Evaluation

4.1.1 Data Sets

Friend and Family datasets: The dataset is from the Human Dynamics Lab at Massachusetts Institute of Technology (MIT), and many experiments have been done based on it [18]–[20]. It can be easily accessed at `http://realitycommons.media.mit.edu/friendsdataset4.html`. The dataset contains 57 couples, more than 50,000 friends, and contains their daily call records, SMS records, and location records. Detailed information is shown in Table 2.

4.1.2 Environments

Experiments were performed on efficient distributed computing platform, Spark 2.1.0 and HDFS 2.6.0, constituted by three servers with Intel Xeon CPU E5-2640 v2 @ 2.00GHz, 64GB memory, 1200GB HD@15000rpm and CentOS 6.5. All codes were implemented in Scala with Scala-Eclipse2.0.2 IDE.

4.1.3 Evaluation Measures

F1-Score: F1-score is a commonly used two-class classification evaluation indicator, which can be calculated by Eq. (12). A higher F1-score means a better result. F1-score is the combination of two indicators: Precision and Recall, which are calculated based on Confusion Matrix. In our experiments, "Family" and "Not Family" are different classes, and the confusion matrix is shown in Table 3. The calculation of Precision and Recall shows in Eqs. (13) and (14).

$$F1 = \frac{2 \times P \times R}{P + R} \tag{12}$$

Table 2 Information of friend and family datasets.

Log Title	Size
Call Log	164,906
SMS Log	88,656
Location Log	970,607
Survey of Couples	57
Survey of Friendship	50,915

Table 3 Confusion matrix of family identifying.

Actual Class \ Predicted	Family	Not Family
Family	TP (True Positive)	FN (False Negative)
Not Family	FP (False Positive)	TN (True Negative)

$$p = \frac{TP}{TP + FP} \tag{13}$$

$$R = \frac{TP}{TP + FN} \tag{14}$$

4.2 Result Analysis on Different Algorithms

Here, we selected 7 continuous months' data which from Nov-2010 to June-2011 to do the experiments. T0-data (2010.11) was used as training dataset, and other 6 months' data were test dataset. To reduce the impact of abnormal users, users-pairs with less than 5 calls were deleted in the dataset. Besides, we chose two baseline algorithms: Support Vector Machine (SVM) and Random Forest (RF).

Support Vector Machine (SVM) [16]: proposed by Corinna Cortes and Valdimir Vapnik, is a kind of supervised learning model that can be applied to classification tasks. In our experiments, the L1-Regulatization was applied to avoid over-fitting.

Random Forest (RF) [15]: firstly proposed by Tin Kam Ho in 1995. It is an ensemble learning method that operate by constructing a multitude of decision trees. In our experiments, the L1-Regulatization was used to avoid over-fitting. The tree's maximum depth was set as 3, and the criterion of tree generation was Gini indicator.

Results are shown in Table 4, and the best results are in bold italic style. Results show that the algorithm outperformed than other two baseline algorithms, which shows the highest F1-score among 6 datasets. It is easy to see that F1-score in T_2 data is the worst compared to others. That is because T_2 data produced in January, when there were many holidays, especially for students. For families with children, the communication mode of parents and children changed a lot, which shows a great effect on the accuracy of results. However, it can still be seen that the NFRR is stable enough and shows the highest score when the users' communication patterns fluctuated.

4.3 Parameters Discussion

In this part, two parameters as follows are discussed in detail:

1. Number of the nearest neighbors: K.
2. Family relationship judgment threshold value: α.

4.3.1 Parameter K

We assigned the parameter K range as 2 to 10. Experimental results are shown in Table 5. The best results are in bold style, and the best results with lowest K are in bold and italic

Table 4 Results of F1 under the different models.

Model	T_0	T_1	T_2	T_3	T_4	T_5	T_6
NFRR	-	***0.875***	***0.813***	***0.866***	***0.862***	***0.864***	***0.847***
SVM	-	0.853	0.810	0.839	0.828	0.850	0.844
RF	-	0.860	0.762	0.830	0.827	0.805	0.823

Table 5 Results with difference K.

K Value	T_0	T_1	T_2	T_3	T_4	T_5	T_6
2	-	0.867	0.813	0.832	0.829	0.827	0.817
3	-	***0.875***	0.802	***0.866***	0.851	0.853	0.843
4	-	**0.875**	0.813	**0.866**	***0.862***	***0.864***	***0.847***
5	-	**0.875**	0.813	0.863	0.861	0.857	0.839
6	-	**0.875**	0.813	0.863	0.855	0.857	0.839
7	-	**0.875**	0.813	**0.866**	0.861	0.857	0.839
8	-	**0.875**	***0.819***	**0.866**	0.861	0.857	0.839
9	-	**0.875**	**0.819**	**0.866**	0.861	0.857	0.839
10	-	**0.875**	0.813	**0.866**	0.861	0.857	0.839

Table 6 Results with difference α.

α	Friend and Family Dataset					
	T_1	T_2	T_3	T_4	T_5	T_6
0.1	0.568	0.537	0.703	0.505	0.505	0.671
0.2	0.704	0.624	0.679	0.553	0.596	0.696
0.3	0.781	0.731	0.703	0.612	0.672	0.677
0.4	0.747	***0.787***	0.754	0.783	0.727	0.785
0.5	0.779	0.775	0.789	0.754	0.782	0.763
0.6	0.786	***0.788***	0.776	0.780	0.790	0.792
0.7	***0.836***	0.777	***0.811***	***0.814***	***0.841***	***0.821***
0.8	***0.809***	0.771	***0.794***	***0.837***	0.790	***0.819***
0.9	0.620	0.694	0.701	0.727	***0.797***	0.808

style. We can find that the best results on T_1 and T_3 appeared first when $K = 3$. For data set T_4, T_5 and T_6, the best results appeared first when $K = 4$.

When K is less than 4, we can see that as the K value increases, the accuracy of the result gradually increases. That is due to more users are selected to calculate the SFS matrix, the patterns they concluded are more accurate. Fewer similar users also cause the over-fitting problem, which has a great effect on the experimental results. When K is greater than 4, the accuracy of most experiments is decreasing, owing to that more similar users will blur the communication patterns. Moreover, a larger K means more calculations on similar calculation and sorting, which could increase the overall computation time.

Therefore, it is really important to choose the best K to balance the accuracy and computational complexity. For this dataset, we choose 4 as the best setting.

In addition, results on T_2 do not show similar patterns as others, that the optimal result is obtained at $K = 8$ and $K = 9$. Clearly, the result is heavily affected by the holiday, and we can achieve a sub-optimal result when $K = 4$.

4.3.2 Parameter α

α means the threshold value of the final judgment, whose value range from 0 to 1. For example, $\alpha = 0.9$ means for any users, they can be labelled with "Family" only when the possibility between two users is higher than 0.9.

Firstly, we split the interval from 0 to 1 by 0.1 and conducted 10 experiments to compare the results with different α. Results are shown in Table 6, and the Top-2 best results are in bold and italic style. When α increases among 0.1 to 0.7, F1-Score increases, which means the accuracy of NFRR is gradually improved. When α further increases to 0.9, the F1-score of each data set decreases. This is because a larger α will cause more misjudgments on real home users which did not show clearly and specifically communication patterns.

Table 7 Results with difference α.

α	Friend and Family Dataset					
	T_1	T_2	T_3	T_4	T_5	T_6
0.7	0.836	0.777	0.811	0.814	0.841	0.821
0.72	0.836	0.782	0.807	0.815	0.832	0.821
0.74	0.875	***0.813***	***0.866***	***0.862***	***0.864***	**0.847**
0.76	***0.890***	0.812	0.841	0.853	0.797	0.829
0.78	0.817	0.749	0.778	0.838	0.737	0.808
0.80	0.809	0.771	0.794	0.837	0.790	0.819

According to Table 6, we can find that the best value is located on interval $[0.7, 0.8]$. Therefore, to additionally find the best α, we further divided it by 0.02. Results are shown in Table 7, and the best result is in bold and italic style. The data shows that the best result obtained when K assigned with 0.74. Therefore, we choose 0.74 as the best setting.

4.4 Time Complexity Analysis

In NFRR, most of time is used for distance calculating, and the time complexity of our algorithm is $\Theta(M \times N)$, where M represents the number of unlabeled users in current cycle, N represents the number of labeled user in the past cycles. In order to fasten the total calculation, we implemented the distributed version of NFRR based on Spark. It reduces the total time cost to $\Theta(\frac{1}{C} \times M \times N)$, where C represents the number of partitions calculated in parallel.

5. Conclusion and Future Work

In this paper, we study the problem of recognizing family relationship in telecom social networks. We propose Relation Spectrum Matrix (RSM) based on spatial, temporal and communication behavior to mine patterns between users, and propose a novel family relation recognition algorithm NFRR to discover families in telecom network. Experimental results on datasets show both the effectiveness and robustness of NFRR. However, there are still many potential directions of this work in the future. On the one hand, to improve accuracy with gathering more application information, such as price, permissions and so on. On the other hand, to expand the idea of NFRR to other social ties, such as colleagues and friends.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (2016YFE0204500), the National Natural Science Foundation of China (61671081), the Beijing Natural Science Foundation (4172042), and the Fundamental Research Funds for the Central Universities.

References

[1] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo,

"Mining advisor-advisee relationships from research publication networks," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.203–212, 2010.

[2] H. Zhuang, J. Tang, W. Tang, T. Lou, A. Chin, and X. Wang, "Actively learning to infer social ties," Data Mining and Knowledge Discovery, vol.25, no.2, pp.270–297, 2012.

[3] J. Tang, T. Lou, and J. Kleinberg, "Inferring social ties across heterogenous networks," ACM International Conference on Web Search and Data Mining, pp.743–752, ACM, 2012.

[4] W. Tang, H. Zhuang, and J. Tang, "Learning to infer social ties in large networks," Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, vol.6913, pp.381–397, Springer, 2011.

[5] J. Choi, S. Heo, J. Han, G. Lee, and J. Song,"Mining social relationship types in an organization using communication patterns," Conference on Computer Supported Cooperative Work, pp.295–302, 2013.

[6] J.-K. Min, J. Wiese, J.I. Hong, and J. Zimmerman, "Mining smartphone data to classify life-facets of social relationships," Conference on Computer Supported Cooperative Work, pp.285–294, 2013.

[7] X. Han, Á. Cuevas, N. Crespi, R. Cuevas, and X. Huang, "On exploiting social relationship and personal background for content discovery in P2P networks," Future Generation Computer Systems, vol.40, pp.17–19, 2014.

[8] C. Yu, N. Wang, L.T. Yang, D. Yao, C.-H. Hsu, and H. Jin, "A semi-supervised social relationships inferred model based on mobile phone data," Future Generation Computer Systems, vol.76, pp.458–467, 2016.

[9] R. Liu, Y. Gao, and Z. Zhang, "Family group recognition and application in complex telecom social network," Industrial Engineering and Management, vol.21, no.5, pp.105–110, 2016.

[10] J. Tang, T. Lou, J. Kleinberg, and S. Wu, "Transfer learning to infer social ties across heterogeneous networks," ACM Trans. Information Systems, vol.34, no.2, Article No.7, 2016.

[11] H.W. Lauw, E.-P. Lim, H. Pang, and T.-T. Tan, "STEvent: Spatio-temporal event model for social network discovery," ACM Trans. Information Systems, vol.28, no.3, Article No.15, 2010.

[12] N. Zhou, W.X. Zhao, X. Zhang, J.-R. Wen, and S. Wang, "A general multi-context embedding model for mining human trajectory data," IEEE Trans. Knowl. Data Eng., vol.28, no.8, pp.1945–1958, 2016.

[13] H. Wang, Z. Li, and W.-C. Lee, "PGT: Measuring mobility relationship using personal, global and temporal factors," IEEE International Conference on Data Mining, pp.570–579, 2014.

[14] H.-P. Hsieh and C.-T. Li, "Inferring social relationships from mobile sensor data," International Conference on World Wide Web, pp.293–294, 2014.

[15] T.K. Ho, "Random decision forests," Proc. Third International Conference on Document Analysis and Recognition, pp.278–282, IEEE, 1995.

[16] C. Cortes and V. Vapnik, "Support-vector networks," Machine Llearning, vol.20, no.3, pp.273–297, 1995.

[17] A. Davidson and A. Or, "Optimizing shuffle performance in spark," Tech. Rep., Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 2013.

[18] Y. Altshuler, N. Aharony, M. Fire, Y. Elovici, and A. Pentland, "Incremental learning with accuracy prediction of social and individual properties from mobile-phone data," 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Confernece on Social Computing (SocialCom), pp.969–974, IEEE, 2012.

[19] J. Staiano, B. Lepri, N. Aharony, F. Pianesi, N. Sebe, and A. Pentland, "Friends don't lie: Inferring personality traits from social network structure," Proc. 2012 ACM Conference on Ubiquitous Computing, pp.321–330, ACM, 2012.

[20] Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A.S. Pentland, "Predicting personality using novel mobile phone-based metrics," International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction, Lecture Notes in Computer Science, vol.7812, pp.48–55, Springer, Berlin, Heidelberg, 2013.

[21] B. Hu, M. Jamali, and M. Ester, "Spatio-temporal topic modeling in mobile social media for location recommendation," IEEE International Conference on Data Mining, pp.1073–1078, 2013.

[22] S. Liu, G. Li, and J. Feng, "A prefix-filter based method for spatio-textual similarity join," IEEE Trans. Knowl. Data Eng., vol.26, no.10, pp.2354–2367, 2014.

[23] X. Bao, J.Yang, Z. Yan, L. Luo, Y. Jiang, E.M. Tapia, and E. Welbourne, "CommSense: Identify social relationship with phone contacts via mining communications," IEEE International Conference on Mobile Data Management, pp.227–234, 2015.

[24] L. Backstrom and J. Kleinberg, "Romantic partnerships and the dispersion of social ties: A network analysis of relationship status on facebook," Proc. 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, pp.831–841, 2014.

[25] H. Liang, K. Wang, and F. Zhu, "Mining social ties beyond homophily," IEEE International Conference on Data Engineering, pp.421–432, 2016.

[26] W. Xu, M. Rezvani, W. Liang, J.X. Yu, and C. Liu, "Efficient algorithms for the identification of top-k structural hole spanners in large social networks," IEEE Trans. Knowl. Data Eng., vol.29, no.5, pp.1017–1030, 2017.

[27] A.A. Nanavati, R. Singh, D. Chakraborty, K. Dasgupta, S. Mukherjea, G. Das, S. Gurumurthy, and A. Joshi, "Analyzing the structure and evolution of massive telecom graphs," IEEE Trans. Knowl. Data Eng., vol.20, no.5, pp.703–718, 2008.

[28] N. Eagle, A. Pentland, and D. Lazer, "Infering social network structure using mobile phone data," Proc. National Academy of Sciences, pp.79–88, 2013.

Kun Niu is an associate professor in School of Software Engineering, Beijing University of Posts and Telecommunications. Her research interests include big data analysis, data mining, intelligent information process and industry application.

Haizhen Jiao received her B.E. in 2016 from Beijing University of Posts and Telecommunications. She is a Grade 3 Master student at School of Software Engineering, Beijing University of Posts and Telecommunicaitons. Her research interests include data mining and knowledge discovery.

Cheng Cheng graduated from Beijing University of Posts and Telecommunications in 2017 and now is working for master of Software Engineering in Beijing University of Posts and Telecommunicaitons. He has always been interested in big data mining and maching learning.

Huiyang Zhang received his B.E. in 2017 from Beijing University of Posts and Telecommunications. He is a Grade 2 Master student at School of Software Engineering, Beijing University of Posts and Telecommunicaitons. His research interests include data mining and knowledge discovery.

Xiao Xu received his B.E. in 2017 from Beijing University of Posts and Telecommunications. He was admitted to School of Software Engineering, Beijing University of Posts and Telecommunications for a postgraduate degree. His research interests include data mining, relationship recognition and machine learning.