Wan HY, Wang ZW, Lin YF *et al.* Discovering family groups in passenger social networks. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 30(5): 1141–1153 Sept. 2015. DOI 10.1007/s11390-015-1589-z

Discovering Family Groups in Passenger Social Networks

Huai-Yu Wan[1] (万怀宇), *Member, CCF*, Zhi-Wei Wang[1] (王志伟)
You-Fang Lin[1] (林友芳), Xu-Guang Jia[2] (贾旭光), and Yuan-Wei Zhou[2] (周元炜)

[1] *Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology*
Beijing Jiaotong University, Beijing 100044, China
[2] *TravelSky Technology Limited, Beijing 100010, China*

E-mail: {hywan, 12120461, yflin}@bjtu.edu.cn; {xgjia, zhouyw}@travelsky.com

Received September 16, 2014; revised June 12, 2015.

Abstract People usually travel together with others in groups for different purposes, such as family members for visiting relatives, colleagues for business, friends for sightseeing and so on. Especially, the family groups, as a kind of the most common consumer units, have a considerable scale in the field of passenger transportation market. Accurately identifying family groups can help the carriers to provide passengers with personalized travel services and precise product recommendation. This paper studies the problem of finding family groups in the field of civil aviation and proposes a family group detection method based on passenger social networks. First of all, we construct passenger social networks based on their co-travel behaviors extracted from the historical travel records; secondly, we use a collective classification algorithm to classify the social relationships between passengers into family or non-family relationship groups; finally, we employ a weighted community detection algorithm to find family groups, which takes the relationship classification results as the weights of edges. Experimental results on a real dataset of passenger travel records in the field of civil aviation demonstrate that our method can effectively find family groups from historical travel records.

Keywords passenger social network, family group, collective classification, community detection

1 Introduction

People usually travel to the same destinations and for the same purposes together with other people in groups, such as family members for taking a vacation or visiting relatives, colleagues for business, friends for sightseeing or attending major events and so on. Especially, family groups as one of the most common travel consumer units, have a considerable scale in the field of civil aviation market. If we can figure out a method to accurately detect family groups in the massive amount of passengers, it will greatly help carriers to provide personalized travel services and precise product recommendation to passengers[1]. For example, family members usually like to sit together with each other on airplanes, so airlines can reserve adjacent seats for family groups to improve their satisfaction; airports can provide convenient check-in channels for family groups with children or elders to increase their convenience; travel agencies can recommend suitable travel lines according to different family patterns (such as couples, nuclear families, three-generation families); and airlines can provide specialized services for high-value family groups. At the same time, the statistics of family groups will greatly help the governments or related organizations to make decisions. For example, analyzing the travel patterns of different types of families can assist airlines to optimize their product portfolios and help local governments to improve city infrastructure and adjust the destination image[2-3].

With the rapid development of information technology, airlines accumulate sufficient information about

Regular Paper

Special Section on Social Media Processing

This work was partially supported by the Fundamental Research Funds for the Central Universities of China, the National Natural Science Foundation of China under Grant No. 61403023, the Beijing Committee of Science and Technology under Grant No. Z131110002813118, and the Program for Changjiang Scholars and Innovative Research Team in University of China under Grant No. IRT201206.

©2015 Springer Science + Business Media, LLC & Science Press, China

passengers and their historical travel behaviors in the passenger information systems (PIS), which provides us a potential way to effectively discover family groups from the massive amount of passengers. The greatest challenge of finding family passenger groups is to discover family relations between the massive amount of passengers. Specifically, one may travel together with many others in many trips, but how could we know which ones are his/her family members? In addition, many real-world family relations may miss in the historical travel records because some family members have never flew together, and this situation will hinder us from discovering intact families.

In this paper, we propose a novel method to discover family groups based on the personal information and historical travel records of passengers, from the perspective of detecting specific community structures in social networks.

First, we construct a specific kind of large-scale passenger social networks, called co-travel networks, by extracting social relations between passengers from their historical travel records. In the field of civil aviation, the social relations in co-travel networks can be obtained from the information of booking tickets together which is recorded as passenger name records (PNRs) in PIS, and such networks can to some extent reflect the real social relations between passengers.

Fig.1 gives a simple example of a family group in a co-travel network consisting of nine passengers. The letters and the numbers in a bracket indicate the gender, the age, and the historical travel times of a passenger respectively, whereas the number beside an edge indicates the historical co-travel times between two passengers. The red solid lines in the figure represent family relations, while the blue dotted lines represent college relations and the green dashed lines represent friend relations. According to the types of social relations, we can easily detect the family group (indicated by the red dashed ellipse) which is composed of family relations.

Next, we will discover family groups based on the co-travel networks of civil aviation. Intuitively, this problem can be straightforwardly treated as a community detection problem. However, general community detection algorithms[4] cannot generate typed communities, because they do not take into account the community types. Actually, in social networks, the types (i.e., the topics) of communities are mainly determined by the labels of relations between people. For example, family relations form family communities, friend relations form friend communities, and colleague relations form colleague communities. In this paper, discovering family groups is a specific problem of typed community detection.

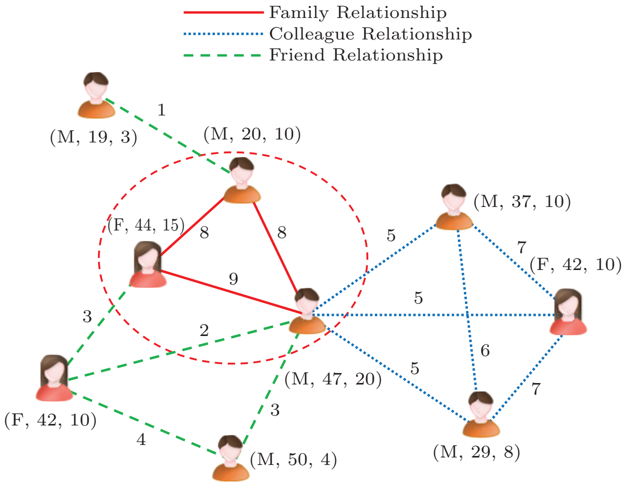


Fig.1. Simple example of a family group in a co-travel network.

In order to discover family groups, we need to identify the types of relations in co-travel networks at the first place, i.e., to determine which ones are family relations and which ones are not. The identification of family relations is really a classification problem and the most straightforward method is employing traditional classifiers. We can generate a series of basic features for relation classification, including historical co-travel characteristics (such as co-travel times, co-travel mileages, and co-travel times on holidays), demographic characteristics (such as gender, age, and family name) and network-based characteristics (such as the number of common neighbors). Traditional classification is based on an assumption of independent and identical distribution (IID); however, the labels of relations are not completely independent. For example, if passengers A and B are family members, so are passengers B and C, then the probability that passengers A and C are also family members becomes higher. Consequently, we need to simultaneously decide on the labels of all the passenger relations together, rather than to classify each relation separately, which is so called collective inference[5]. In this paper, we employ conditional random fields[6] to execute the collective classification of passenger relations.

Finally, we discover family groups based on the results of relation classification. For each relation, we take the probability that it belongs to the "Family" relations as the weight; and then we employ weighted community detection algorithms to discover family communities.

Our proposed framework utilizes not only the link structure but also the types of links in social networks to discover the specific type of communities, i.e., family groups, which overcomes the shortcomings of general community detection methods. Therefore, it will expectedly resolve the family discovery problem very well.

We experiment on a real dataset in the field of civil aviation. Experimental results demonstrate that our method which discovers family groups based on relation classification in passenger co-travel networks can effectively discover family groups from the passenger information data and their historical travel records. In addition, the results also indicate that in the phase of relation classification, the collective inference strategy is much superior to the traditional classifiers, while in the phase of community detection, the weighted community detection outperforms other general methods.

The rest of this paper is organized as follows. The next section provides a brief introduction of related work. Section 3 formally defines the problem and Section 4 presents typed community detection methods based on relation classification, followed by the experimental evaluations in Section 5. Finally, we give the conclusions of our work in Section 6.

2 Related Work

In recent years, there are many researches related to family behaviors in travel and tourism management. For example, Lehto *et al.*[7] studied the correlation between vacation activities and family cohesion, and found that spending leisure time with family on holidays has a great influence on family function and thus family travelling becomes a very important part of tourism marketing. Prayag *et al.*[8] studied the travel motivation of elders, and illustrated that entertainment with family is one of the most important reasons. Barlés-Arizón *et al.*[9] also studied the travel patterns of different family structures. These researches have been successfully applied to many fields, such as the improvement of tourism product involvement[2] and airport services[10].

A similar work to our family detection problem is predicting social circles of users in their ego networks[11], which was posed as a node clustering problem on a user's ego network, and network structure as well as user profile is combined for detecting circles. However, this work also did not consider the types of social circles, such as family circles, colleague circles, classmate circles or friend circles.

In order to identify the types of social communities, we should first identify the types of social relations. Relation classification has been an important research topic in social network mining. In recent years, there have been many studies on inferring the meanings of social relations. Zhao *et al.*[12] used relational Markov networks to predict the types of relations between terrorists. Diehl *et al.*[13] introduced a method to identify the manager-subordinate relations by learning a ranking function. Eagle *et al.*[14] presented using communication patterns to infer friendship relations in mobile social networks. Wang *et al.*[15] proposed an unsupervised probabilistic model to mine the advisor-advisee relations from publication networks. Crandall *et al.*[16] investigated the problem of inferring friendship between people from co-occurrence in time and space. Tang *et al.*[17] proposed a partially-labeled pairwise factor graph model (PLP-FGM) to infer the types of social ties within a semi-supervised framework. Wan *et al.*[18] presented a community-based conditional random field model to label the relationships in social networks. Tang *et al.*[19] studied the problem of inferring social ties across heterogeneous networks. The major focus of these studies is how to make use of the attributes of relations and the structural information of social networks to collectively infer the types of social relations.

The problem of discovering family groups can be essentially treated as a community detection problem. Community structure[20-21] is a very hot research issue in the field of social networks. A general community detection problem can be modeled as an unsupervised learning problem. A large number of community detection algorithms[4] have been developed. For example, the most original and famous one is the GN algorithm[22], which is based on modularity function and followed by numerous variations, such as BGLL[23]; the Infomap algorithm[24] based on information theory is widely recognized as one of the most accurate and stable community detection algorithms; Label Propagation algorithm (LPA)[25] is one of the most efficient community detection algorithms. In recent years, the researches on overlapping community detection have received more and more attentions and many relevant algorithms are emerging, such as clique percolation[26] based on complete sub-graph, hierarchical clustering[27] based on division of edge, LFM[28] based on local optimization and COPRA[29] based on label propagation. However, the general community detection algorithms do not consider the types of communities, so they cannot produce typed communities (such as family communities, colleagues communities, and so on), and the discovered communities are not interpreted. Some researchers study the general community detection problem on signed networks where both positive and negative links are present. For example, Traag and Bruggeman[30] extended an existing Potts[31] model to incorporate negative links by adapting the concept of modularity to detect communities, which solves a long-standing problem in the theory of social balance, namely the clustering of signed graphs.

In this paper, we discover family groups from passenger information and historical travel records, based on relation classification and community detection. It is very interesting to study the relations between passengers from the perspective of social networks. Lin *et al.*[32] used historical co-travel records to construct passenger social networks and infer the travel purposes of passenger groups, that is, to identify whether a passenger group is a tourist group or a business group. This research is very useful for airlines and corresponding agencies to better understand the passengers and provide them targeted travel services or product recommendation, as well as to adjust the product portfolios and marketing strategies. Our work in this paper further studies how to discover family groups from massive passengers, which is believable to be applied to the precision marketing and decision making.

3 Definitions

In this section, we give some necessary definitions first and then formulate the research problems.

Definition 1 (Passenger Group). *A passenger group* $g = (P_g, ori_g, dst_g, dis_g, d\_date_g, r\_date_g)$ *is a collection of passengers who book tickets together for a certain travel, where* $P_g = \{p_i\}$ *is a set of passengers,* ori_g *and* dst_g *are the common origin and the common destination of the travel,* dis_g *is the mileage of the travel, and* d_date_g *and* r_date_g *are the departure date and the return date of the travel, respectively.*

Obviously, we have $|P_g| \geqslant 1$, where $|P_g| = 1$ means a passenger travels alone.

In practice, airlines record the data of all their passenger groups in their information systems. Given a large set of passenger groups $S = \{g_1, g_2 \dots g_n\}$ in a period of time, we can construct co-travel networks by extracting co-travel relations from S.

Definition 2 (Co-Travel Networks). *A co-travel network is a graph* $G = (V, E)$*, where* V *is a node set, and each node* $p_i \in V$ *represents a passenger;* E *is an edge set, and each edge* $e_{ij} = (p_i, p_j) \in E$ *indicates that passengers* p_i *and* p_j *have booked tickets and traveled together at least once.*

The relations in co-travel networks represent the passengers' pairwise behaviors of booking tickets and travelling together, which reflect the real social relations between passengers to some extent. For training and evaluating our proposed method, we need to annotate a sample dataset. However, because of the difficulty of observing the social relations in the real world, we can only annotate very little part of relations in large-scale co-travel networks. We call such networks as sparsely labeled networks.

Definition 3 (Sparsely Labeled Co-Travel Networks). *A sparsely labeled co-travel network is an augmented co-travel network denoted as* $G = (V, E^L, E^U)$*.* V *is a node set;* E^L *is a labeled edge set, and each edge* $e_{ij} \in E^L$ *indicates a relation whose label is known (such as a family or colleague relation);* E^U *is an unlabeled edge set, and each edge* $e_{ij} \in E^U$ *indicates a relation whose label is unknown. Obviously, we have* $E^L \cup E^U = E$*.*

In sparsely labeled co-travel networks, we can take the labeled relations as training data to learn a classifier to infer other unlabeled relations. We can define the problem as follows:

Problem 1 (Relation Classification). *Given a sparsely labeled co-travel network* $G = (V, E^L, E^U)$*, the objective is to learn a function:*

$$f : G = (V, E^L, E^U) \rightarrow L,$$

where L *is the label space of the problem, i.e., all the possible types of social relations.*

Our goal in this paper is to identify whether an unlabeled relation is a family relation or not, and in this case, the problem can be treated as a binary classification problem.

Definition 4 (Weighted Co-Travel Network). *In a sparsely labeled co-travel network* $G = (V, E^L, E^U)$*, after the relation classification, we will get the probability that an edge* $e_{ij} \in E$ *belongs to a certain relation type; then we treat the probability as the weight of the relation and form a weighted co-travel network* $G = (V, E^W)$*.*

Definition 5 (Family Group). *We employ* $C^F = (V^F, E^F)$ *to represent a family group, where* V^F *is a set of family members and* $V^F \in V$*,* E^F *is a set of relations between passengers and* $E^F \in E^W$*.*

Problem 2 (Discovering Family Groups in Co-Travel Networks). *Given a weighted co-travel network $G = (V, E^W)$, the objective is to cluster all the nodes and get a set $C^F = \{c_k^F\}$ which makes each node only belong to a single community $c_k^F \in C^F$, where c_k^F represents a family group.*

For clarity, the notations used in this paper are summarized in Table 1.

Table 1. List of Notations

Notation	Description
$G = (V, E)$	Co-travel network
$G = (V, E^L, E^U)$	Sparsely labeled co-travel network
$G = (V, E^W)$	Weighted co-travel network
$c_k^F = (V_k^F, E_k^F)$	Family group
p_i, p_j	Passenger
e_{ij}	Relation between passengers p_i and p_j
g	Passenger group
P_g	Member set of passenger group g
R_{ij}	Set of co-travel records between passengers p_i and p_j
r_{ij}^t	The t-th co-travel record in R_{ij}
$seat_i^t$	Seat number of passenger p_i on his/her t-th co-travel
$checkin_i^t$	Check-in number of passenger p_i on his/her t-th co-travel
$mile_{ij}^t$	Mileage of the t-th co-travel in R_{ij}
$d\_date_{ij}^t$	Departure date of the t-th co-travel in R_{ij}
$r\_date_{ij}^t$	Return date of the t-th co-travel in R_{ij}
ori_{ij}^t	Origin of the t-th co-travel in R_{ij}
dst_{ij}^t	Destination of the t-th co-travel in R_{ij}
dis_{ij}^t	Distance between ori_{ij}^t and dst_{ij}^t
age_i	Age of passenger p_i
$gender_i$	Gender of passenger p_i
$birth\_place_i$	Birth place of passenger p_i
$surname_i$	Surname of passenger p_i
W_{ij}	Weight of relation e_{ij}
d_i	Degree of passenger p_i in G
Γ_i	Set of neighbor nodes of p_i
SP_{ij}	Set of shortest paths passing through e_{ij}

4 Proposed Approach

In this section, we will introduce the approach of discovering family groups in civil aviation passenger social networks in details. Overall, our proposed framework includes four steps as follows:

• step 1: construct co-travel networks from historical travel records;

• step 2: generate features for passenger relations, including historical co-travel statistics, demographic characteristics, and network-based features;

• step 3: classify passenger relations in co-travel networks;

• step 4: discover family groups in co-travel networks via different kinds of community detection methods, based on the results of relation classification.

4.1 Constructing Co-Travel Networks

As mentioned before, airlines record historical behaviors of all the passengers in their information systems, and each record is actually a passenger group. For example, in the field of civil aviation, a PNR contains the travel information of a group of passengers who book air tickets together. From such travel records, we can construct co-travel networks in which the edges to some extent reveal the social relations between passengers in the real world.

Given a set of passenger groups $S = \{g_n\}$, we simply construct a co-travel network $G = (V, E)$ by extracting relations from each passenger group g_n, as outlined in Algorithm 1.

Algorithm 1. Constructing Co-Travel Networks

```
Input: $S = \{g_1, g_2, g_3, \ldots, g_n\}$
Output: $G = \{V, E\}$
  for each passenger group $g \in S$
    for each passenger pair $(p_i, p_j)$ $(p_i, p_j \in P_g, p_i \neq p_j)$
      if $p_i \notin V$ then
        $V \leftarrow V \cup \{p_i\}$;
      end if;
      if $p_j \notin V$ then
        $V \leftarrow V \cup \{p_j\}$;
      end if;
      if $e_{ij} \notin E$ then
        $w_{ij} = 1$;
        $E \leftarrow E \cup \{e_{ij}\}$;
      else
        $w_{ij} = w_{ij} + 1$;
      end if;
    end for;
  end for;
```

4.2 Relation Classification

In partially labeled co-travel networks, we have known a small part of relation labels, which can be used as the training set to infer the unknown labels. The basic idea for relation labeling is employing traditional classifiers to classify each relation separately, where all the labels are assumed to be independent and identically distributed (IID). But in fact, some dependencies may exist among the labels of relations in co-travel networks, for instance, the transitivity dependencies exist among the family relations. Suppose that passengers A and B are family members, so are passengers B and C, and then the passengers A and C should also be family members. Considering such dependencies, we employ a collective classification algorithm to classify the social relations between passengers in co-travel networks.

4.2.1 Generating Features

According to different generation mechanisms, we divide the features into three categories: historical co-travel statistics, demographic characteristics, and network-based features.

Historical Co-Travel Statistics. Historical co-travel statistics are based on the co-travel behaviors between passengers. For example, different types of relations may have different travel patterns, e.g., different co-travel times and mileages; family members usually check in together and sit adjacently while colleagues may not, so the check-in number and seat difference may be very important to the relation classification; in addition, from the aspect of travel time, family members usually choose to travel on weekends or holidays, while colleagues may frequently travel on workdays, so the ratios of co-travel times on weekends, holidays, and workdays may also play an important role. Table 2 lists all the historical co-travel statistics and their formulizations.

Demographic Characteristics. The demographic characteristics (e.g., age, gender, birthplace, and surname) are another type of important features. Different types of relation may have different demographic characteristics. For example, the age difference is usually great between parents and children, and small among children, while it is not sure among colleagues. The gender composition may also be helpful for relationship classification. For example, the gender composition must be male and female between parents, while it is not sure among colleagues. In addition, children almost have the same surname with their fathers, while colleagues usually have different surnames. The demographic characteristics we consider in our work and their formulizations are listed in Table 3.

Network-Based Features. All the above features of passenger relations are directly generated from the relations themselves, without considering the structural features of relations in the whole co-travel network. In fact, the rich link information of networks can also generate features which are helpful to infer the labels of relations. This paper considers several kinds of common network structure features of edge. Table 4 lists all the network-based attributes and their formulizations.

4.2.2 Conditional Random Fields

Conditional random fields (CRF) are undirected graphical models developed for labeling sequence data, which represent the conditional distribution over a set of hidden random variables given the observed ones. We use $\mathcal{F} = (X, Y, \mathcal{E})$ to represent a CRF, where X is the set of observed random variables, Y is the set of hidden random variables, and these variables are con-

Table 2. Historical Co-Travel Statistics

Feature	Formulization
Co-travel times	$co\_travel\_cnt_{ij} = \lvert R_{ij}\rvert$
Average co-travel mileage	$avg\_dist_{ij} = \frac{\sum_{r_{ij}^t \in R_{ij}} mile_{ij}^t}{\lvert R_{ij}\rvert}$
Maximum co-travel mileage	$max\_dist_{ij} = max_{r_{ij}^t \in R_{ij}} mile_{ij}^t$
Minimum co-travel mileage	$min\_dist_{ij} = min_{r_{ij}^t \in R_{ij}} mile_{ij}^t$
Average seat number difference	$avg\_seat_{ij} = \frac{\sum_{r_{ij}^t \in R_{ij}} \lvert seat_i^t - seat_j^t\rvert}{\lvert R_{ij}\rvert}$
Maximum seat number difference	$max\_seat_{ij} = max_{r_{ij}^t \in R_{ij}} \lvert seat_i^t - seat_j^t\rvert$
Minimum seat number difference	$min\_seat_{ij} = min_{r_{ij}^t \in R_{ij}} \lvert seat_i^t - seat_j^t\rvert$
Average check-in number difference	$avg\_checkin_{ij} = \frac{\sum_{r_{ij}^t \in R_{ij}} \lvert checkin_i^t - checkin_j^t\rvert}{\lvert R_{ij}\rvert}$
Count of co-travel on workdays	$workday\_cnt_{ij}$
Ratio of co-travel on workdays	$workday\_ratio_{ij} = \frac{workday\_cnt_{ij}}{\lvert R_{ij}\rvert}$
Count of co-travel on weekends	$weekend\_cnt_{ij}$
Ratio of co-travel on weekends	$weekend\_ratio_{ij} = \frac{weekend\_cnt_{ij}}{\lvert R_{ij}\rvert}$
Count of co-travel on holidays	$holiday\_cnt_{ij}$
Ratio of co-travel on holidays	$holiday\_ratio_{ij} = \frac{holiday\_cnt_{ij}}{\lvert R_{ij}\rvert}$

Table 3. Demographic Features

Feature	Formulization
Age difference	$age\_diff_{ij} = \lvert age_i - age_j \rvert$
Whether same birthplace	$birth\_place_{ij} = \begin{cases} 1, \text{if } birth\_place_i = birth\_place_j \\ 0, \text{otherwise} \end{cases}$
Gender combination 1	$gender1_{ij} = \begin{cases} 1, \text{if } gender_i = gender_j = \mathrm{F} \\ 0, \text{otherwise} \end{cases}$
Gender combination 2	$gender2_{ij} = \begin{cases} 1, \text{if } gender_i = gender_j = \mathrm{M} \\ 0, \text{otherwise} \end{cases}$
Gender combination 3	$gender3_{ij} = \begin{cases} 1, \text{if } gender_i \neq gender_j \\ 0, \text{otherwise} \end{cases}$
Whether same surname	$surname_{ij} = \begin{cases} 1, \text{if } surname_i = surname_j \\ 0, \text{otherwise} \end{cases}$

nected by a set of undirected edges $\mathcal{E}$ which indicates the relevancies between them. Let x be an assignment of values to X and y be an assignment of values to Y. CRF $\mathcal{F}$ defines a conditional distribution $P(y|x)$ over the hidden states y conditioned on the observations x.

Table 4. Network-Based Features

Feature	Formulization
Number of common neighbors	$com\_neighbor\_cnt_{ij} = \lvert \Gamma_i \cap \Gamma_j \rvert$
Average degree of common neighbors	$avg\_com\_neighbor\_d_{ij} = \frac{\sum_{p_k \in \Gamma_i \cap \Gamma_j} d_k}{com\_neighbor\_cnt_{ij}}$
Betweenness	$betweenness_{ij} = \frac{\lvert SP_{ij} \rvert}{\sum_{e_{kl} \in E} \lvert SP_{kl} \rvert}$

For a CRF $\mathcal{F} = (X, Y, \mathcal{E})$, a clique c is a set of nodes in $\mathcal{F}$ such that each pair $u, v \in c$ is connected by an edge. Let C be the set of cliques in $\mathcal{F}$. Then, a CRF factorizes the conditional distribution into a product of clique potentials $\phi_c(x_c, y_c)$, where x_c and y_c are the conditional and the target variables in clique c respectively. Clique potential ϕ_c is a non-negative real function defined on c, which indicates the compatibility among the variables in the clique. Given an assignment $x_c \cup y_c$, the larger the potential value, the more likely the assignment. By using clique potentials, the conditional distribution over the target variables in a graph is defined as:

$$P(y \mid x) = \frac{1}{Z(x)} \prod_{c \in C(G)} \phi_c(x_c, y_c), \tag{1}$$

where $Z(x)$ is the partition function dependent on x:

$$Z(x) = \sum_{y'} \prod_{c \in C(G)} \phi_c(x_c, y'_c).$$

The potential is often represented by a log-linear combination of a set of feature functions:

$$\begin{aligned} \phi_c(x_c, y_c) &= \exp\{\sum_k w_k f_k(x_c, y_c)\} \\ &= \exp\{w_c \times f_c(x_c, y_c)\}, \end{aligned}$$

where w_k is the weight of the k-th feature function f_k. Then the log-linear representation of (1) can be abbreviated as follows:

$$\begin{aligned} \log P(y \mid x) &= \sum_{c \in C(G)} \sum_k w_k f_k(x_c, y_c) - \log Z(x) \\ &= w \times f(x, y) - \log Z(x). \end{aligned}$$

Constructing CRF. In the relation classification in co-travel networks, we employ the transitivity dependencies of family relations mentioned before to construct CRF. The relations are corresponding to the target variables Y in CRF and their content features are corresponding to the conditional variables X. An edge between any pair of target variables (Y_m, Y_n) is established if the relations m and n are neighboring (i.e., have a common node) in the co-travel network. We need to define cliques and the feature functions for the clique potentials. Two types of cliques need to be defined: evidence cliques and compatibility cliques. An evidence clique is a dyad clique that consists of a target variable and one of its content features. It indicates the direct dependency of the target variable conditioned on the feature. Compatibility cliques consist entirely of target variables which indicate the correlations among target variables.

Referring to the classification of family and non-family relations in this paper, we define triad compatibility cliques according to the transitivity regularity. If any three relations form a loop in the co-travel network, then we establish a compatibility clique for their corresponding target variables in the CRF. Then we need to define the feature functions for the clique potentials. Here we just define potentials for a binary classification model. For the dyad evidence cliques, we use the indicator functions with the form $f_k(x_k, y) = y \times x_k$, where $y = \pm 1$, $x_k \in [0, 1]$. And for the triad compatibility cliques, we simply use a single feature function to track whether the three labels are the same:

$$f_k(y_m,\ y_n,\ y_l) = \begin{cases} 1, & \text{if } y_m = y_n = y_l, \\ 0, & \text{otherwise.} \end{cases}$$

Learning and Inference. Maximum likelihood estimation (MLE) can be used to learn the parameters of CRFs[6]. In the process of parameters estimation, computing the expected feature function directly is an NP-hard problem in general. Consequently, we cannot perform an exact inference and need to use approximate inference algorithms in CRF. Belief propagation (BP)[33] and Markov chain Monte Carlo (MCMC)[34] are two kinds of approximate inference algorithms which are employed commonly.

Wan *et al.*[35] employed pseudo-likelihood to approximately describe the CRF and proposed a maximum pseudo-likelihood estimation (MPLE) to learn parameters without using approximate inference algorithms. Simultaneously, for making the inference procedure converge more quickly, the authors proposed an iterative inference algorithm. With only a little loss of accuracy, the pseudo-likelihood based CRF can greatly improve the efficiency of learning and inference. Thus we employ this approach in our experiments.

4.3 Discovering Family Groups

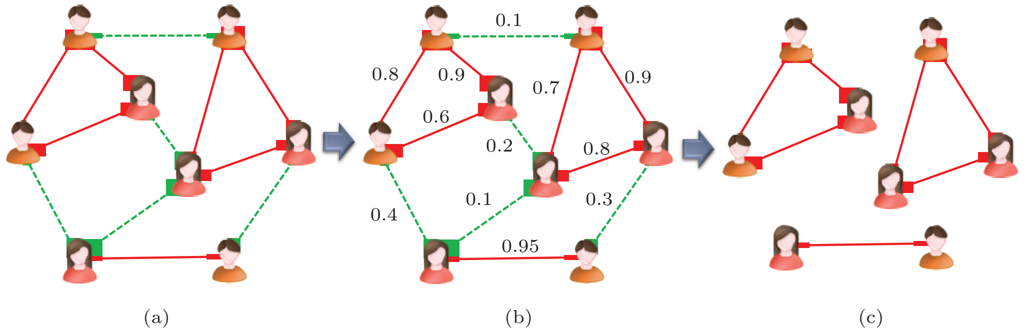
In this subsection, we will discuss how to utilize the results of relation classification to detect family groups in passenger co-travel networks. The basic idea is that the more likely two passengers have a family relationship, the more likely they belong to the same family. Thus we take the probability that a relation is labeled as "Family" to be its weight, and then use a weighted community detection method to identify family groups.

After the relation classification, we get the probability of each possible label for a relation, and then the label with the maximum probability is identified as the true label of the relation. Such a probability indicates not only the possibility that the relation belongs to a certain type of relations, but also the tightness of the relation under the type. Consequently, in this work, we can use the probabilities of the label "Family" of relations as the weights, and employ weighted community detection methods to discover family groups. Many sophisticated community detection methods, such as Infomap[24] and BGLL[23], can be used in the framework.

In order to reduce the influence of low-weight edges to the community detection algorithms, we set a threshold ξ to filter them. First of all, the relations whose weights are less than threshold ξ are removed from the network directly, and then we start to execute the weighted community detection algorithms.

Fig.2 illustrates the process of employing weighted community detection method to discover family groups in co-travel networks. Fig.2(a) is a simple co-travel network, in which a solid line indicates a family relation while a dashed one indicates a non-family relation. After relation classification, we get the weighted family relation network, as shown in Fig.2(b). Finally, we use weighted community detection algorithms to discover family groups, as shown in Fig.2(c).

General community detection algorithms assume that all the edges in a network are equal, while the weighted ones take into account the different influences of edges with different weights to the community structure of the network. Consequently, the communities discovered through the weighted community detection



(a) (b) (c)

Fig.2. Process of employing weighted community detection to discover family groups in passenger co-travel networks. (a) Passenger co-travel network. (b) Weighted family relation network. (c) Detected family groups.

algorithms are more accurate and reasonable. Using the weighted community detection algorithms based on the results of relation classification is able to utilize the link structure of the whole network to revise the deviation of the mislabeling of relations, which makes the detected family groups more accurate.

5 Experiments

In this section, we present a set of experiments on a real-world dataset to evaluate the performance of our proposed family passenger group discovering framework.

5.1 Dataset

We collected a dataset from the encrypted civil aviation PNRs provided by TravelSky Technology Ltd.[1], which is the largest civil aviation IT provider in China. The dataset contains the information of passengers (e.g., ID, gender, and age) and historical travel records (e.g., passenger ID, flight number, origin, destination, departure time, and seat number) from 2010 to 2011. It should be emphasized that the personal and travel information used in our experiments is private and must be seriously protected. In fact, all the passenger IDs, flight numbers, origins, destinations, departure time, and seat numbers appearing in our dataset had been irreversibly encrypted by the provider, so that one cannot backtrack a specific passenger from the encrypted data.

Based on the dataset, we construct a passenger co-travel network which contains 13 492 passengers, and 25 383 relationships. Some of the passenger relation types (family or non-family) in the network are previously labeled by the provider so that the structure of family groups is known and we have 378 family groups in total. Statistics of the dataset is shown in Table 5.

Table 5. Statistics of the Dataset

Number of passengers	13 492.00
Number of relations	25 383.00
Number of labeled family relationships	2 104.00
Number of labeled non-family relationships	4 835.00
Number of labeled family groups	378.00
Average number of family groups	3.87

5.2 Experimental Setup

As mentioned before, general community detection cannot effectively discover family groups since it does not consider the community types. In our experiments, we use two popular general community detection algorithms, i.e., Infomap[24] and BGLL[23], to verify this assertion. These two methods can automatically identify the number of communities and are considered to be the most efficient community detection methods up to now.

Our proposed framework includes two main phases, relation classification and community detection. In the relation classification phase, we compare the performance of our proposed collective classification algorithm (i.e., the classifier based on conditional random fields) with a traditional classifier (i.e., logistic regression), to identify family relations. In the community detection phase, we compare the performance of our proposed weighted community detection method (Infomap[24] is employed as the basic community detection algorithm) with two baseline methods:

Edge Cutting. This is a very simple method which directly uses the labels of relations in co-travel networks. For each relation, we remove it from the network if it is not a family relation (i.e., the probability that the relation is labeled as “Family” is less than a certain threshold ξ); otherwise, we remain it. And then the network is partitioned into many connected components and the family groups are generated naturally.

Signed Network Community Detection. Traag and Bruggeman[30] proposed a model to incorporate both positive and negative links to detect communities in signed networks. For our co-travel networks, based on the results of relation classification, we treat family relations as positive links and non-family relations as negative links, and then use the model proposed in [30] to detect family groups. We also use the threshold ξ to divide all the relations into positive and negative links. Specifically, we let the relations whose probabilities of the “Family” label are less than ξ be negative links; otherwise, positive links.

Evaluation Metrics. Based on the labeled data, we use precision, recall, $F1$-measure, and rand index (RI) to evaluate the results of discovering family groups by different methods, according to whether a passenger relation (i.e., a pair of passengers) is accurately divided into real families. Let TP represent the number of real family relations which are correctly divided into the

[1] http://www.travelsky.net/, June 2015.

same families, TN represent the number of real family relations which are incorrectly divided into different families, FP represent the number of non-family relations which are correctly divided into different families, and FN represent the number of non-family relations which are incorrectly divided into the same families, then we have

$$Precision = TP/(TP + FP),$$
$$Recall = TP/(TP + FN),$$
$$F1 = 2 \times Precision \times Recall/(Precision + Recall),$$
$$RI = (TP + TN)/(TP + FP + FN + TN).$$

5.3 Experimental Results

We first report the experimental results of relation classification, and then compare the performance of discovering family groups by different methods, including general community detection algorithms and the methods based on relation classification, i.e., edge cutting, signed network community detection, and weighted community detection.

5.3.1 Results of Relation Classification

In the process of relation classification, we employ logistic regression and CRF to execute a 5-fold cross validation on the labeled data respectively. The average precision of logistic regression is 81.63%, while that of CRF is 89.16% (increased by 7.53%). In order to compare the effectiveness of the two methods, Fig.3 gives their ROC curves respectively, from which we can see that the collective classification method (i.e., CRF) is obviously superior to the traditional classifier (i.e., logistic regression).

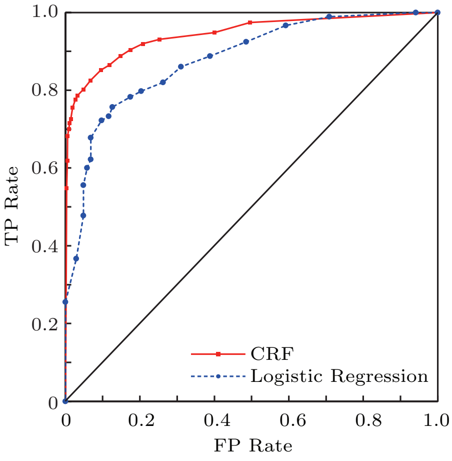


Fig.3. ROC curves of different classifiers.

The experimental results demonstrate that to classify the relations between passengers from the perspective of co-travel networks by employing collective classification (i.e., CRF) is much better than that by traditional classifiers based on the IID assumption. That is because the collective classification not only utilizes the content attributes of each instance, but also considers the correlations among the instances by modeling the probabilistic dependencies among the target variables.

To assess the importance of different features, we calculate the significance (i.e., the P-values) and contribution of each feature. For the contribution index, we in turn remove each feature from the logistic regression model and evaluate the performance degradation by each feature. The P-values (p) and contributions (c) of different features are shown in Table 6.

Table 6. Significances and Contributions of Features

	Feature	p	c(%)
Historical co-travel statistics	Co-travel times	0.543	+0.53
	Travel distance	0.436	+0.64
	Average co-travel mileage	0.611	+0.56
	Maximum co-travel mileage	0.731	+0.34
	Minimum co-travel mileage	0.669	+0.35
	Average seat distance	**0.032**	+3.93
	Maximum seat distance	0.396	+1.76
	Minimum seat distance	0.188	+2.41
	Average check-in interval	**0.045**	+4.67
	Count of co-travel on workdays	0.316	+1.28
	Ratio of co-travel on workdays	0.079	+2.95
	Count of co-travel on weekends	0.201	+1.66
	Ratio of co-travel on weekends	**0.036**	+4.36
	Count of co-travel on holidays	0.152	+1.20
	Ratio of co-travel on holidays	**0.017**	+5.78
Demographic characteristics	Age difference	0.181	+1.33
	Whether same birthplace	0.397	+0.57
	Gender combination 1	0.432	+0.62
	Gender combination 2	0.531	+0.36
	Gender combination 3	0.136	+1.60
	Whether same surname	0.089	+2.34
Network-based features	Number of common neighbors	0.217	+1.02
	Average degree of common neighbors	0.526	+0.25
	Betweenness	0.135	+1.19

From Table 6, we can see that the two measures are very consistent. The four most important features are average seat distance ($p = 0.032$, $c = +3.93\%$), average check-in interval ($p = 0.045$, $c = +4.67\%$), ratio of co-travel on weekends ($p = 0.036$, $c = +4.36\%$), and ratio of co-travel on holidays ($p = 0.017$, $c = +5.78\%$). Besides, some other features, such as ratio of co-travel on workdays, whether same surname, betweenness, gender combination 3, and age difference, are also helpful to the classification.

5.3.2 *Results of Discovering Family Groups*

The performance of general community detection to discover family groups in co-travel networks is shown in Table 7, where two popular community detection algorithms, i.e., Infomap[24] and BGLL[23], are employed. We can see that the precisions of both algorithms are very low, even though their recalls are quite high. This demonstrates that too many non-family relations are incorrectly regarded as family relations and many wrong family groups are detected.

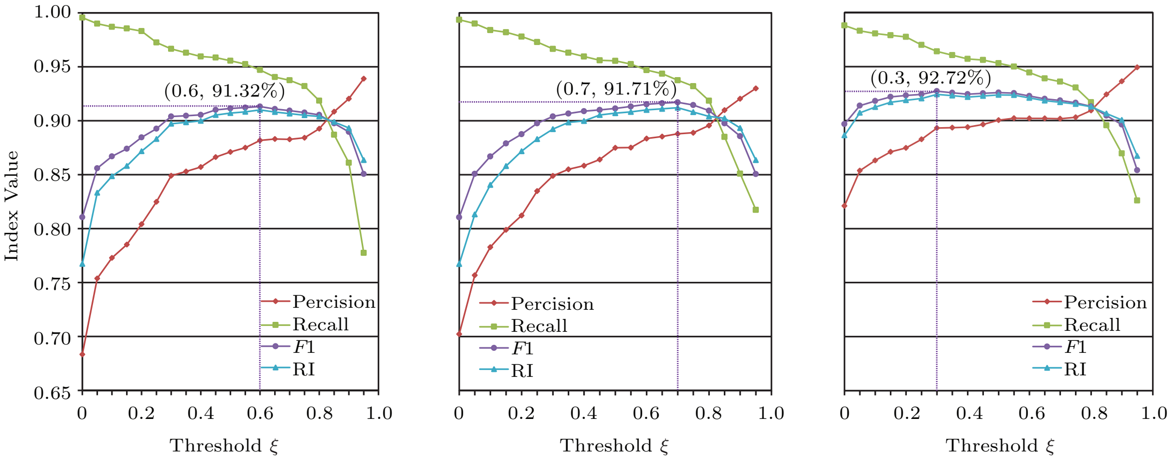
Table 7. Performance of General Community Detection

	Algorithm	
	Infomap (%)	BGLL (%)
Precision	59.58	50.39
Recall	95.90	99.85
$F1$-measure	73.50	66.98
RI	65.43	50.78

Based on the best results of classification by CRF, we employ the edge cutting, Traag's signed network community detection and weighted Infomap[24] algorithm to discover family groups respectively. The results are shown in Fig.4, in which the X-axis indicates the values of the threshold ξ, and the Y-axis indicates the values of different indices (i.e., Precision, Recall, $F1$ and RI). From Fig.4, we can see that the precision becomes better and better while the recall becomes worse and worse with the growth of the threshold. The edge cutting method achieves its best performance when the threshold equals 0.6 (where the $F1$-measure reaches a maximum value of 91.32%), and the signed network community detection method achieves its best performance when the threshold equals 0.7 (where the $F1$-measure reaches a maximum value of 91.71%), while the weighted community detection method achieves its best performance when the threshold equals 0.3 (where the $F1$-measure reaches a maximum value of 92.72%).

Overall, the weighted community detection method outperforms the edge cutting and the signed network community detection method, and it performs more stablely over all the thresholds and increases the best $F1$-measure by 1.40% and 1.01% respectively, which demonstrates that community detection based on the results of relation classification is an effective way to detect typed communities. Taking the probability-formed results of relation classification as the weights of networks, weighted community detection algorithms can easily discover communities of a specific type.

Based on the same results of relation classification, the weighted community detection method is superior to the simple edge cutting method and the signed network community detection method. The disadvantage of the latter two methods is that the community detection accuracy is completely dependent on the accuracy of relation classification, and the mislabeling of some key relations may lead to some serious errors in family group detection. While the weighted community detection not only utilizes the link information, but also considers the influence of the weights of edges to the



(a)

(b)

(c)

Fig.4. Performance of discovering family groups. (a) Edge cutting. (b) Signed network community detection. (c) Weighted community detection.

network community structure. It uses the link structure of the whole network to revise the deviation of the mislabeling of relations, and thus makes the detected family groups more accurate.

Because the edge cutting and the signed network community detection methods divide all the relations into family (positive) or non-family (negative) links according to the threshold ξ, they are more sensitive to the threshold, while the weighted community detection method is not so sensitive.

6 Conclusions

In this paper, we studied how to discover family groups based on the personal information and historical travel records of passengers in the field of civil aviation, and proposed a family group detection framework by combing relation classification and weighted community detection, in the context of passenger social networks. First of all, we constructed passenger co-travel networks based on their co-travel behaviors extracted from the historical travel records; secondly, we designed a serious of features and used a collective classification algorithm to classify the social relations between passengers into the family or non-family relations; finally, we employed weighted community detection method to discover family groups based on the results of relation classification. Experimental results on a real-world dataset demonstrated that our approach can effectively find family groups from massive passengers.

Obviously, our proposed framework can also be applied to discover specific typed groups in other network fields. For example, it may be used to identify colleague groups from email networks, find adviser-advisee groups in co-author networks, discover function groups from protein interaction networks, etc.

References

[1] Regan N R, Carlson J, Rosenberger III P J. Factors affecting group-oriented travel intention to major events. *Journal of Travel & Tourism Marking*, 2012, 29(2): 185-204.

[2] So S I, Lehto X Y. The situation influence of travel group composition: Contrasting Japanese family travelers with other travel parties. *Journal of Travel & Tourism Marketing*, 2007, 20(3/4): 79-91.

[3] Pike S, Ryan C. Destination positioning analysis through a comparison of cognitive, affective, conative perceptions. *Journal of Travel Research*, 2004, 42(4): 333-342.

[4] Fortunato S. Community detection in graphs. *Physics Reports*, 2010, 486(3/4/5): 75-174.

[5] Xiang R, Neville J. Collective inference for network data with copula latent Markov networks. In *Proc. the 6th ACM International Conference on Web Search and Data Mining*, Feb. 2013, pp.647-656.

[6] Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. the 18th International Conference on Machine Learning*, June 28-July 1, 2001, pp.282-289.

[7] Lehto X Y, Lin Y C, Chen Y, Choi S. Family vacation activities and family cohesion. *Journal of Travel & Tourism Marketing*, 2012, 29(8): 835-850.

[8] Prayag G. Senior travelers' motivations and future behavioral intentions: THE CASE OF NICE. *Journal of Travel & Tourism Marketing*, 2012, 29(7): 665-681.

[9] Barlés-Arizón M J, Fraj-Andrés E, Martínez-Salinas E. Family vacation decision making: The role of woman. *Journal of Travel & Tourism Marketing*, 2013, 30(8): 873-890.

[10] Tam M L, Lam W H K, Lo H P. Modeling air passenger travel behavior on airport ground access mode choices. *Transportmetrica*, 2008, 4(2): 135-153.

[11] Mcauley J J, Leskovec J. Learning to discover social circles in ego networks. In *Proc. the 26th Annual Conference on Neural Information Processing Systems*, Dec. 2012, pp.548-556.

[12] Zhao B, Sen P, Getoor L. Entity and relationship labeling in affiliation networks. In *Proc. the 23nd ICML Workshop on Statistical Network Analysis: Models, Issues, and New Directions*, June 2006.

[13] Diehl C P, Namata G, Getoor L. Relationship identification for social network discovery. In *Proc. the 22nd AAAI Conference on Artificial Intelligence*, July 2007, pp.546-552.

[14] Eagle N, Pentland A S, Lazer D. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 2009, 106(36): 15274-15278.

[15] Wang C, Han J, Jia Y, Tang J, Zhang D, Yu Y, Guo J. Mining advisor-advisee relationships from research publication networks. In *Proc. the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 2010, pp.203-212.

[16] Crandall D J, Backstrom L, Cosley D, Suri S, Huttenlocher D, Kleinberg J. Inferring social ties from geographic coincidences. *Proceedings of National Academy of Sciences*, 2010, 107(52): 22436-22441.

[17] Tang W, Zhuang H, Tang J. Learning to infer social ties in large networks. In *Proc. the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases*, Sept. 2011, pp.381-397.

[18] Wan H, Lin Y, Wu Z, Huang H. A community-based pseudolikelihood approach for relationship labeling in social networks. In *Proc. the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases*, Sept. 2011, pp.491-505.

[19] Tang J, Lou T, Kleinberg J. Inferring social ties across heterogeneous networks. In *Proc. the 5th ACM International Conference on Web Search and Data Mining*, Feb. 2012, pp.743-752.

[20] Girvan M, Newman M E J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 2002, 99(12): 7821-7826.

[21] Newman M E J. Detecting community structure in networks. *The European Physical Journal B*, 2004, 38(2): 321-330.

[22] Bollobás B. Random Graphs (2nd edition). Cambridge, UK: Cambridge University Press, 2001.

[23] Blondel V D, Guillaume J L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, 2008(10): Article No. 10008.

[24] Rosvall M, Bergstrom C T. Map of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 2008, 105(4): 1118-1123.

[25] Raghavan U, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 2007, 76(3): Article No. 036106.

[26] Palla G, Derényi I, Farkas I *et al.* Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005, 435: 814-818.

[27] Ahn Y, Bagrow J, Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*, 2010, 466: 761-764.

[28] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure of complex networks. *New Journal of Physics*, 2009, 11(3): Article No. 033015.

[29] Gregory S. Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 2010, 12(10): Article No. 103018.

[30] Traag V A, Bruggeman J. Community detection in networks with positive and negative links. *Physical Review E*, 2009, 80(3): Article No. 036115.

[31] Reichardt J , Bornholdt S. Statistical mechanics of community detection. *Physical Review E*, 2006, 74(1): Article No. 016110.

[32] Lin Y, Wan H, Jiang R, Wu Z, Jia X. Inferring the travel purposes of passenger groups for better understanding of passengers. *IEEE Transactions on Intelligent Transportation System*, 2015, 16(1): 235-243.

[33] Murphy K P, Weiss Y, Jordan M I. Loopy belief propagation for approximate inference: An empirical study. In *Proc. the 15th Conference on Uncertainty in Artificial Intelligence*, July 30-Aug. 1, 1999, pp.467-475.

[34] Robert C P, Casella G. Monte Carlo Statistical Methods (2nd edition). New York, NY: Springer, 2004.

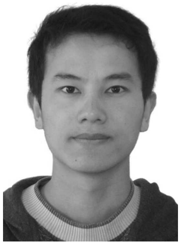
[35] Wan H, Lin Y, Wu Z, Huang H. Discovering typed communities in mobile social networks. *Journal of Computer Science and Technology*, 2012, 27(3): 480-491.

Huai-Yu Wan received his Ph.D. degree in computer science and technology from Beijing Jiaotong University, Beijing, in 2012. He is an assistant professor with the School of Computer and Information Technology, Beijing Jiaotong University. His research interests focus on data mining, social network analysis, and recommender systems.

Zhi-Wei Wang received her B.S. degree in computer science from Hebei University of Science and Technology, Shijiazhuang, in 2012. She is currently working toward her Master's degree in the School of Computer and Information Technology, Beijing Jiaotong University. Her research interests focus on traffic data analysis and mining.

You-Fang Lin received his Ph.D. degree in computer science and technology from Beijing Jiaotong University, Beijing, in 2003. He is a professor with the School of Computer and Information Technology, Beijing Jiaotong University. His main fields of expertise and current research interests include data warehousing, data mining, business intelligence, complex networks, and social network analysis.

Xu-Guang Jia received his M.S. degree in computer science and technology from Beijing Jiaotong University, Beijing, in 2003. He is a senior software engineer of TravelSky Technology Ltd., Beijing. His research and development interests focus on traffic data warehousing and mining.



Yuan-Wei Zhou received his M.S. degree in computer science and technology from Beijing Jiaotong University, Beijing, in 2013. He is a software engineer of TravelSky Technology Ltd., Beijing. His research and development interests focus on traffic data analysis.