Hindawi
Mobile Information Systems
Volume 2017, Article ID 2190310, 17 pages
https://doi.org/10.1155/2017/2190310

Research Article

Discovering Social Community Structures Based on Human Mobility Traces

Cong-Binh Nguyen,[1] Seokhoon Yoon,[1] and Jangyoung Kim[2]

[1] *Department of Electrical and Computer Engineering, University of Ulsan, Ulsan, Republic of Korea*
[2] *Department of Computer Science, University of Suwon, Gyeonggi, Republic of Korea*

Correspondence should be addressed to Seokhoon Yoon; seokhoonyoon@ulsan.ac.kr

Received 21 March 2017; Revised 15 June 2017; Accepted 16 July 2017; Published 6 September 2017

Academic Editor: Bartolomeo Montrucchio

We consider a community detection problem in a social network. A social network is composed of smaller communities; that is, a society can be partitioned into different social groups in which the members of the same group maintain stronger and denser social connections than individuals from different groups. In other words, people in the same community have substantially interdependent social characteristics, indicating that the community structure may facilitate understanding human interactions as well as individual's behaviors. We detect the social groups within a network of mobile users by analyzing the Bluetooth-based encounter history from a real-life mobility dataset. Our community detection methodology focuses on designing similarity measurements that can reflect the degree of social connections between users by considering tempospatial aspects of human interactions, followed by clustering algorithms. We also present two evaluation methods for the proposed schemes. The first method relies on the natural properties of friendship, where the longevity, frequency, and regularity characteristics of human encounters are considered. The second is a movement-prediction-based method which is used to verify the social ties between users. The evaluation results show that the proposed schemes can achieve high performance in detecting the social community structure.

1. Introduction

Social community detection allows insightful investigations into the network structure among social entities [1, 2]. Examining the organization of human networks may enable the comprehension of social features that influence individual behaviors and interactions between network nodes [3].

Social networks maintain a community structure [4, 5]; that is, human society can be segmented into different social groups, in which the intragroup connections (the links between people in the same group) are much more durable and denser than the connections between individuals of different groups. Members of the same social group tend to have a strong degree of interdependence. They interact more frequently and share correlated behavioral characteristics with each other at the community level [6]. Since unveiling the fundamental structure of human society can help to understand individuals' behavior as well as interactions between people, community detection has become an attractive research issue.

Mobile phones with powerful sensing functions have become an integral part of most people's lives. They offer the ability to record contextual information related to daily activities of users. Recently, researchers have been able to exploit the functionalities of mobile phones to track individual behavior and gather sufficient data for analysis.

There have been several studies on social structure detection [7–13]. However, those studies did not take into consideration social similarity measurement [7–11] or only considered the existence of social links between users [12, 13], rather than taking the strengths of users' social ties into account. Although some studies considered the similarity metrics between users [14–16], their main objective is to address the data routing problem in delay-tolerant networks, focusing on distributed community detection. In those works, due to reliance on locally maintained information, each node may obtain a different and partial community structure, rather than obtaining a globally agreed social community structure using the tempospatial context.

On the other hand, for social community detection, in this work, we focused on designing similarity measurements that can reflect the strength of social ties between users by considering tempospatial aspects of human interactions, followed by clustering to obtain social groups of strong-relationship users. We proposed two similarity metrics: encounter-rate-based similarity (ERS) and encounters with temporal correlations similarity (ETCS). These metrics are used to identify the degree of human relations by examining past user encounters. In ERS method, the frequency of human encounters is used to derive social intimacy. In ETCS method, both spatial and temporal factors of encounters are used to assess the degree of social connections. For clustering, the spectral algorithm and a self-organizing map (SOM) are used to obtain social communities.

We also proposed to use two evaluation methods for community detection schemes that take inspiration from properties of human relationship. The first method is based on friendship and utilizes natural characteristics of friendship such as frequency, longevity, and regularity. As an additional evaluation method, we developed a human mobility prediction model that embeds the social structure obtained from the community detection schemes. The prediction accuracy reflects the meaningfulness of the employed social factors (i.e., the level of interdependence between users in social groups), which makes it a potential method to validate the proposed community detection schemes.

In the friendship-based evaluation method, the results showed that the proposed similarity measurements, ETCS and ERS, can achieve higher performance than the eigenbehavior-based method in terms of frequency, longevity, and regularity. We observed that ETCS outperforms others when the size (the number of members within a group) of social groups is large. With a small group size, ERS can gain higher performance than other methods. According to the results of the movement-prediction-based evaluation method, the schemes employing two proposed similarity metrics also outperform the existing scheme with respect to the prediction accuracy. The evaluation results also indicate that using spectral clustering is more beneficial than using SOM in the community detection schemes.

From the mobility dataset collected by human-carried devices from real life, we extracted the necessary contextual data for our work. Bluetooth-encounter records that describe human interactions were used for social structure detection. In the movement prediction-based evaluation method, the cellular network trace, which is able to provide mobile users' position information, is additionally employed.

Our main contributions are summarized as follows:

(i) In order to detect social communities, two social similarity measurement schemes were proposed which determine the level of social closeness between individuals based on encounter history. Tempospatial aspects of human interactions were considered, and both direct and indirect human interactions were used to estimate social similarity between users.

(ii) Contextual data of human interactions and movements were extracted from a real-life mobility dataset for the analysis.

(iii) A friendship-based evaluation method was developed to validate the proposed social community detection schemes, which is rooted in the natural properties of friendship. The longevity, frequency, and regularity characteristics of human interactions are taken into account.

(iv) A movement prediction model was proposed as an additional evaluation method. The proposed human mobility prediction model uses the social community information obtained from the community detection scheme and the prediction accuracy indicates the effectiveness of the community detection schemes.

The rest of this paper is organized as follows. Section 2 presents an overview of related works and a similarity measurement method from existing studies. A description of the dataset we used is given in Section 3. In Section 4, we propose our community detection methodology. The human mobility prediction model we applied in this work will be explained in Section 5. In Section 6, the evaluation methods are described, and we also present the results and a discussion. Finally, the conclusion of this paper is given in Section 7.

2. Related Works

In this section, we first discuss the overview of social structure detection methods and compare those existing studies with ours. Then, we introduce the eigenbehavior-based similarity metric, which will be used to compare with our scheme.

2.1. Social Structure Detection. In this part, we present the existing studies on detecting the organization of networks that are related to our work. We also compare those works with our approach.

Many studies on social structure detection are based on graph clustering, which is the most common method for uncovering community structure [7–11]. In graph clustering-based methods, network is represented by a graph, which is then partitioned into communities by using clustering techniques. For example, the authors in [7] presented a divisive graph clustering method. The social graph is first formed, and the boundary edges, which are most likely to act as intercommunity connections, are removed to obtain disjoint communities. In [17], Ball et al. developed a statistical graph clustering method by employing expectation-maximization algorithm. As another example, Nguyen et al. [18] proposed a graph clustering-based method to detect and monitor the overlapping community structures in a dynamic mobile network, where the network topology frequently changes.

Those studies differ from ours in that they mainly focused on graph-partitioning algorithm, rather than measuring the degree of social closeness between individuals. The major objective of our work is to identify social groups of individuals with close social relations. Note that determining the social similarity is essential in order to effectively detect the network structure, since a community needs to be distinguished from another community based on the level of social connections between individuals. In this paper, we employ network users' interaction history to derive social intimacy.

There have been several studies that considered a social similarity based on the contact history of nodes in networks. For example, Daly and Haahr [12] determined a social similarity and betweenness centrality metric, which they used in order to detect nodes that belong to the same community and to detect nodes that can facilitate communication between different communities. Hossmann et al. [13] also examined the significance of aggregated contacts to construct a social graph. They took into account the observed encounters between nodes in the past to estimate social connections. Pandit et al. [19] studied the community detection problem in the dynamic network. In order to detect a time-varying community structure, the considered period is divided into smaller time intervals. At each time interval, a temporary social link is assumed to exist if two users are in spatial proximity of each other. However, the metrics in those studies only consider the availability (existence) of social link between users by using threshold conditions of past encounters, rather than representing the strengths of social ties between users. In addition, the objective of [19] is to capture the time-varying community structure and hence they considered the spatial proximity at each time interval to construct the social link, rather than taking into account long-term human interactions as in our work.

In [20], Boston et al. proposed an algorithm for detecting social groups based on Bluetooth traces. They employed the frequency and duration of users' meetings in order to group users. A group meeting is defined as a meeting that involves a set of more than two users, where all user pairs in the set have pairwise meetings at the same time. Using the determined group meetings, the user sets are identified. Then, threshold conditions are used to remove the negligible sets with the limited group-meeting frequency and meeting duration and obtain user groups.

This work differs from ours in that they assumed that social groups can be detected only by group meetings. In contrast, in our work, users can belong to the same social group without group meetings through indirect interactions between users. Moreover, the objective of our work is to find disjoint social groups (possibly leading to a large social group), within which users have strong social relationship, while, in their study, a user may belong to a lot of small groups depending on the group meetings the user attends.

A study that extensively uses real-life mobility traces to analyze human social networks is eigenbehavior [21]. Eagle and Pentland determined the affiliation of an individual (which community he/she may belong to) by comparing the social behavior distances (e.g., the number of Bluetooth devices encountered) of that individual to different users who are from predefined communities. However, the social closeness measurement in eigenbehavior only employed simple context data. For example, they only counted the number of past encounters of an individual, rather than considering specific users that the individual encountered; that is, each pairwise encounter has not been taken into account.

Even though the similarity methods in those studies explored interaction history, compared to our method, they do not take into full consideration the potentials of using human encounters to reflect the social relation. Note that the information of who was in contact with a specific user, and specific temporal positions of those encounters, may provide a meaningful social depiction of that individual. Thus, taking into account those contextual data allows us to enhance the estimation of social similarities between users. In our scheme, we propose a new method to assess the degree of social connections between individuals by considering both temporal and spatial aspects of human interactions. Those levels of social intimacy are represented by weighted similarity values.

There are a few community detection schemes based on each user's local information [14–16, 22]. For instance, Hui et al. [14] developed distributed community detection schemes in which an individual node detects its own local community. In order to estimate the relationships between network nodes, they considered using contact duration and the number of contacts in the past, which are correlated to familiarity and regularity characteristics of human interaction. In [15], Bulut and Szymanski presented a metric to evaluate a network node's degree of motivation to meet in order to share data with another node. This metric is based on three natural properties of friendship: longevity, frequency, and regularity in human-encounter history. Li and Wu [16] defined the pairwise similarity metrics between individuals that employed encounter frequency, the average and total contact period, and the separation period between meeting times. Williams et al. [22] examined a community detection problem in the dynamic network. Based on the time-ordered encounter graph constructed, each node focuses on determining the periodic-encounter communities (i.e., the groups in which group members periodically encounter each other). Then, locally obtained data are opportunistically exchanged between nodes in order to the identify the global community structure.

Those studies differ from our work in that each node only uses the information that it locally maintains in order to discover social communities. As a result, each node may obtain a different and partial community structure depending on the information that it has, rather than obtaining the entire community structure based on global data. Although nodes may exchange their information on the network structure, it is hard and costly to achieve the globally agreed social community structure and fine-tune the social structure using global data. We also note that using global data allows more contextual information on human interactions. Moreover, those studies did not provide a method to explicitly evaluate the performance of community detection. Instead, they used packet forwarding performance through trace-driven simulations, through which the community detection accuracy needs to be inferred. In contrast, in order to evaluate the performance of community detection, we propose to use two methods, which are based on the natural properties of friendship and human mobility prediction model. To the best of our knowledge, no studies have presented analytical methods that evaluate community detection scheme by exploiting natural characteristics of human interaction, as we do in this paper, based on analyzing real-life mobility traces.

2.2. Eigenbehavior-Based Similarity. In order to effectively discover the social groups, social similarities between individuals need to be well-determined. Here, we introduce an existing similarity measurement method called eigenbehavior [21]. This method focuses on measuring the distance between behavioral data.

Eagle and Pentland [21] presented a similarity metric based on the behavioral distance between an individual in society and a user in a predefined community. By measuring and then comparing the distances between that individual and different users in different communities, they inferred to which community that person belonged. In a society of users, there exist multiple communities, such as incoming lab students, senior lab students, and business students. For each of those communities, the authors formed a matrix of m_i by 24 to represent community behavioral data. The number of rows, m_i, corresponds to the number of users in community i, and each row vector corresponds to the behavioral data of an individual user of the community. The 24 columns represent 24 hourly intervals of the day, and each row vector consists of 24 elements. The value of each element represents the average number of users encountered [21] over the experiment period at the corresponding hourly intervals.

In this paper, no given communities are assumed. Therefore, the entire society for all users is considered. The similarities between users were obtained using a principal component analysis- (PCA-) based [23] technique. PCA is widely used to identify patterns in data. Using PCA, a simpler yet meaningful representation of data can be extracted. This transformation is achieved through a linear combination of the principal components (which are obtained from the eigenvectors of the covariance matrix of data) so that most of the variation present in the original data can be preserved. In order to compare the behavioral data between individuals to measure their similarities, PCA technique is employed to characterize the society's behavioral data. More specifically, the social distance between individuals was determined, presuming that each person did not have any prior knowledge about the other's social background. Rather than constructing a different behavioral matrix for each community, only one matrix of M by H was formed to represent all users, where $M = \sum m_i$ equals the number of users and H is the number of hourly intervals in the current scheme. Given the fact that vector Ψ is the average behavior of the society and Γ^j is the behavioral vector of person j, the deviation of an individual's behavior from the mean behavior will be $\Phi_j = \Gamma^j - \Psi$. Then, the covariance matrix is constructed based on the set of Φ_j [21]:

$$C = \frac{1}{M}\sum_{j=1}^{M}\Phi_j\Phi_j^T = AA^T, \tag{1}$$

where $A = [\Phi_1, \Phi_2, \ldots, \Phi_M]$. Here, the set of principal components $u_1, u_2, \ldots, u_H$ (defined as eigenbehaviors) is derived from this behavioral covariance matrix. Based on these eigenbehaviors, an individual's behavior can be reconstructed through the transformation below:

$$\omega_k^j = u_k\left(\Gamma^j - \Psi\right) \tag{2}$$

for $k = 1, 2, \ldots, H$. Generally, for this PCA-based method, a lesser number h, where $h < H$, of eigenbehaviors that correspond to the h largest eigenvalues is sufficient to represent users [23]. After that, the reconstruction weight vector of person j can be formed: $\Omega^j = [\omega_1^j, \omega_2^j, \ldots, \omega_h^j]^T$. As with Eagle and Pentland [21], to determine social distance among users, the Euclidean distance between their reconstruction weight vectors is employed:

$$d(j, m) = \left\|\Omega^j - \Omega^m\right\|, \tag{3}$$

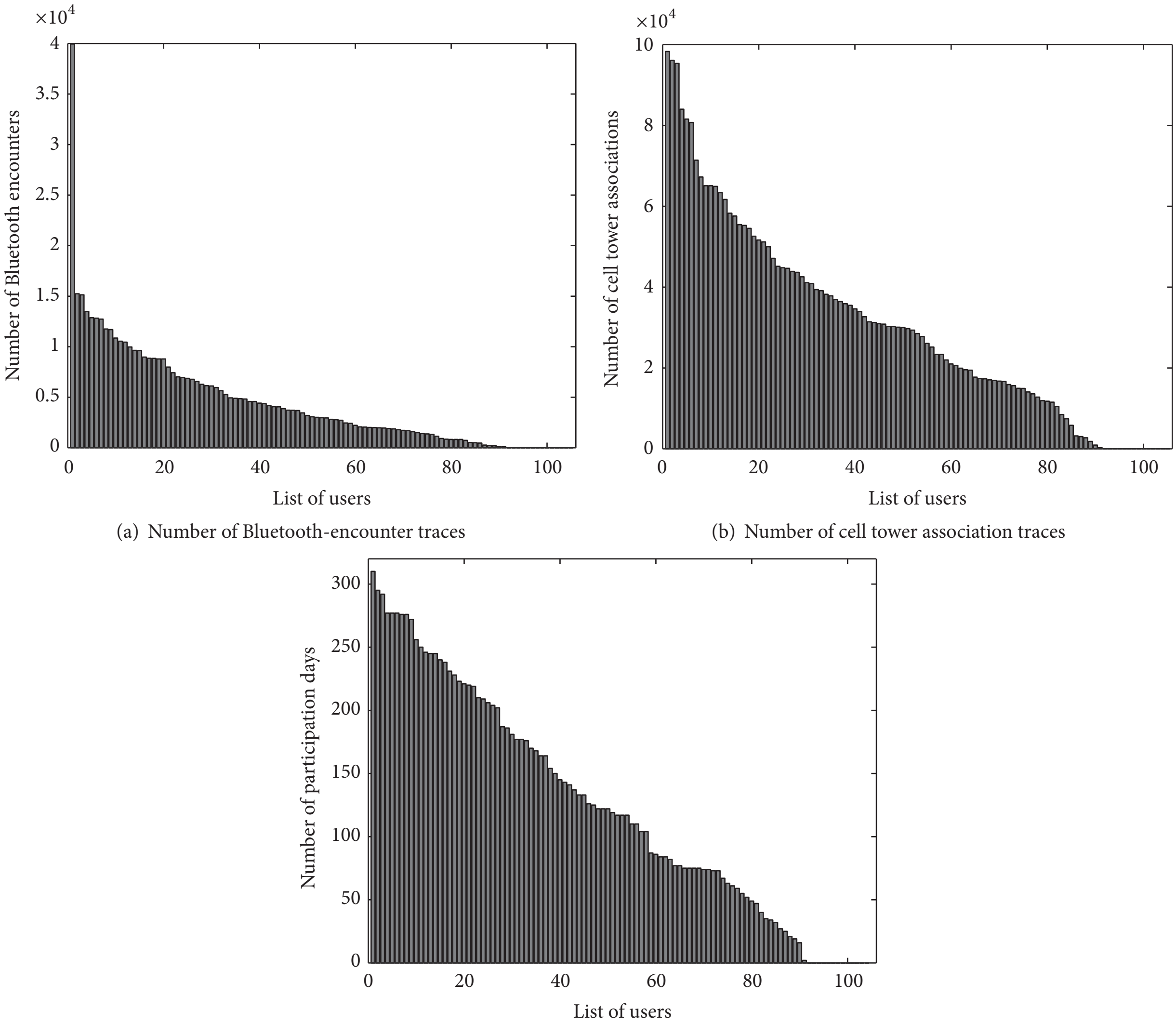
where $d(j, m)$ is the social distance between users j and m, and Ω^j and Ω^m are the reconstruction weight vectors of users j and m, respectively, in the society. The similarity between users is regarded as the inverse of the distance and can be inferred from distance using common transforming techniques, such as Gaussian kernel [24].

3. The Dataset

In this paper, we use the MIT Reality Mining Dataset [25]. This real-life dataset consists of Bluetooth proximity traces, cell tower logs, and communication and mobile application logs that were gathered from more than 90 human-carried devices during part of a school year in 2004. The people who took part in this experiment were from MIT, including media lab members and nearby Sloan Business School students. Since they are in the same academic institute with related positions, there may be many social connections between them, thus implying a social community structure among participants.

Participants in this experiment were given Bluetooth-enabled mobile phones with preinstalled logging software. For the purpose of capturing human interactions, these phones scan the surrounding environment every five minutes and record the list of Bluetooth devices in their proximity with a corresponding timestamp. Since users were likely to have their phones almost all the time during this study, a Bluetooth trace can represent human encounters. Each time a Bluetooth proximity log between mobile devices was recorded, it was assumed to be an encounter between their holders. In addition, because a real-life meeting can last for a longer time than the interval of a Bluetooth scan, it is worth noting that one meeting between users may not be the same as one encounter. In our scope, a meeting of users was identified as a series of consecutive Bluetooth-based encounters between them. Since the Bluetooth scan interval is five minutes, if two persons are in contact for five minutes, it is recorded as two proximity events. Therefore, each proximity event is assumed to be equivalent to 2.5 minutes of contact between users [3].

Besides recording Bluetooth interactions between users, this dataset also contains the cell tower traces that describe the movements of users. In cellular networks, a mobile



(a) Number of Bluetooth-encounter traces

(b) Number of cell tower association traces

(c) Number of participation days

Figure 1: Number of Bluetooth-encounter traces, number of cell tower association traces, and number of participation days, in descending order, for users in MIT Reality Mining Dataset.

phone's service is available when it is located within the coverage area of a cell tower. In general, it will be associated with the nearest cell tower (with the strongest signal) from its current location. Based on natural individual behaviors we observed, a mobile user would move to multiple positions through the day, and his/her connected cell changes accordingly. Every time the device performs a new association with a cell tower, the cell identifier and respective timestamp are logged. Therefore, these traces could indicate symbolic positions, each of which corresponds to the coverage of a cell tower and is labeled with a unique ID. Given that the range of a cell tower is around a few hundred meters in urban areas, using these traces can help to track user movement at a tolerable resolution.

In the MIT Reality Dataset, there are 106 users in total (there are 95 users who have actual data). For a better representation of the social network, we select the set of users who can provide useful data in our analysis. First, as in the eigenbehavior study [21], in this work we only consider data traces by graduate students, who are the majority of participants. Some participants provide no data or a very little amount of data. Thus, we need to select the users who have reasonable amount of data that can represent his/her mobility. Figure 1 shows the statistics obtained for users in MIT dataset. As shown in Figure 1(a), there are 29 users who have less than 1,000 Bluetooth-encounter traces over the entire period (the mean value is 4,446). Similarly, 24 users have less than 10,000 cell tower traces as shown in Figure 1(b) (the mean: 29,781), and 30 users have less than 60 participation days as shown in Figure 1(c) (the mean: 122). Based on the observation from Figure 1, we consider those three thresholds for selecting users and obtained the set of 58 users. Then, we determine the overlapping period in which the selected users participated the experiment. Figure 2

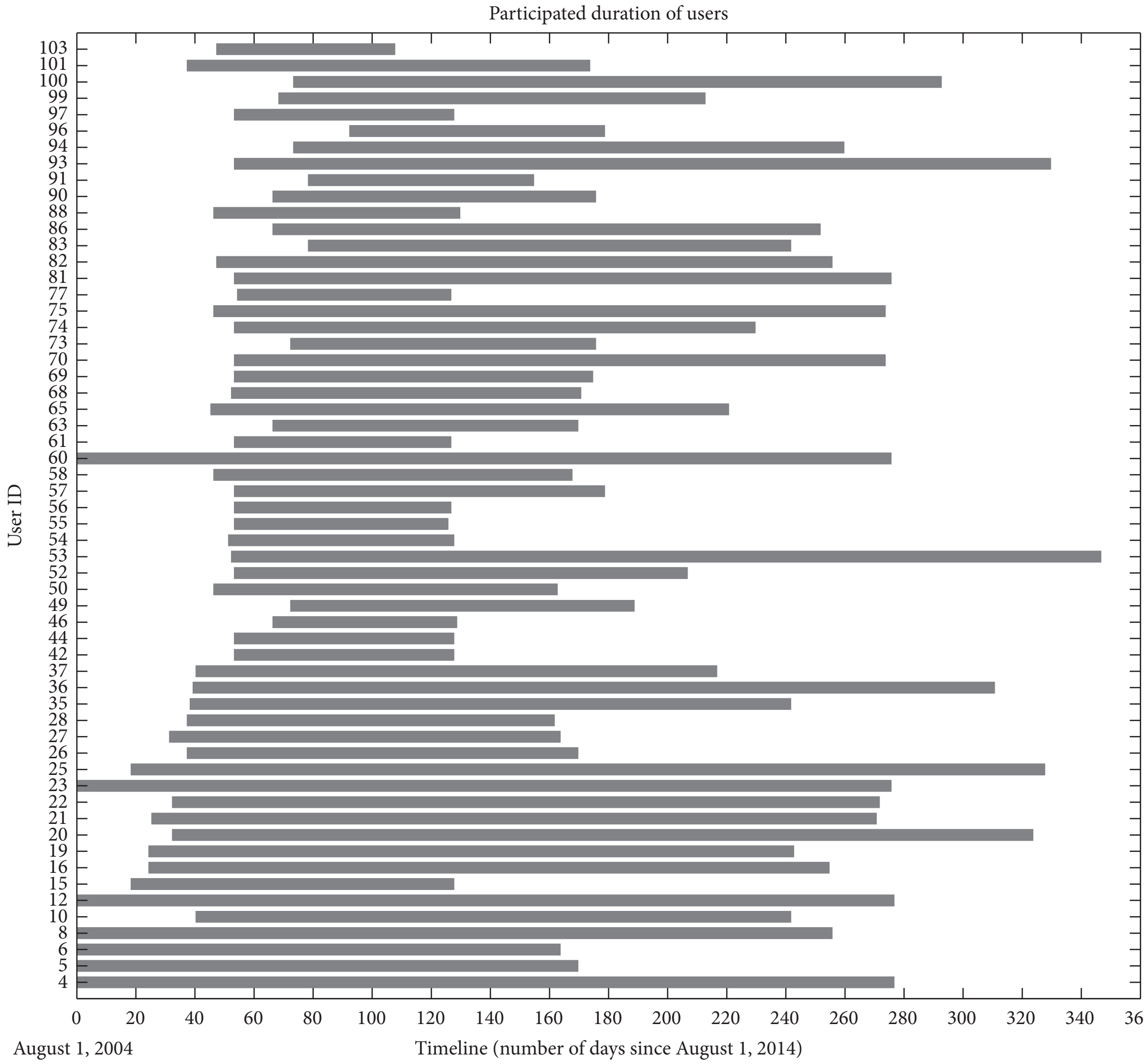


Figure 2: Participated duration of 58 selected users. Each bar represents the participation days between the date a user started and stopped participating in the experiment. August 1, 2014, was the earliest date when user joined MIT experiment.

shows the participated duration of those 58 users. In order to ensure the presence of selected users in the analysis, the period from September 23, 2004, to December 7, 2004, is considered (54 weekdays in total). We also noted that, during the selected period, some users' mobility data are not available most likely due to device issues or occasional inactivity of the user. Therefore, we exclude additional 15 users, who have no Bluetooth encounter recorded for more than 20 days. Finally, we obtained 43 users for analysis.

Figure 3 presents the distribution of the amount of Bluetooth data collected from the chosen users during the overlapping time period. According to our observations, human social patterns are dependent on the time of day. During the work day, especially from 09:00 to 18:00 on a weekday, daily behaviors tend to be more regular than at other times. Moreover, as can be seen in Figure 3, the weekday data account for most of the human interactions. It also indicates that most available interactions happen in normal office hours from 09:00 to 18:00. The fact that participants are from the same educational organization can explain why the recorded interactions are concentrated in those certain periods. Based on the observation, in this work, we focused on studying the data obtained on weekdays, from 09:00 to 18:00.

4. Social Community Detection

In this section, we present new schemes that partition the society into smaller groups, in which mobile users have strong intragroup relations. The scheme includes determining similarities between users, followed by clustering.

The first step is to measure the social similarity between mobile users. Based on the encounter history that reflects the interaction between individuals, the degree of their closeness

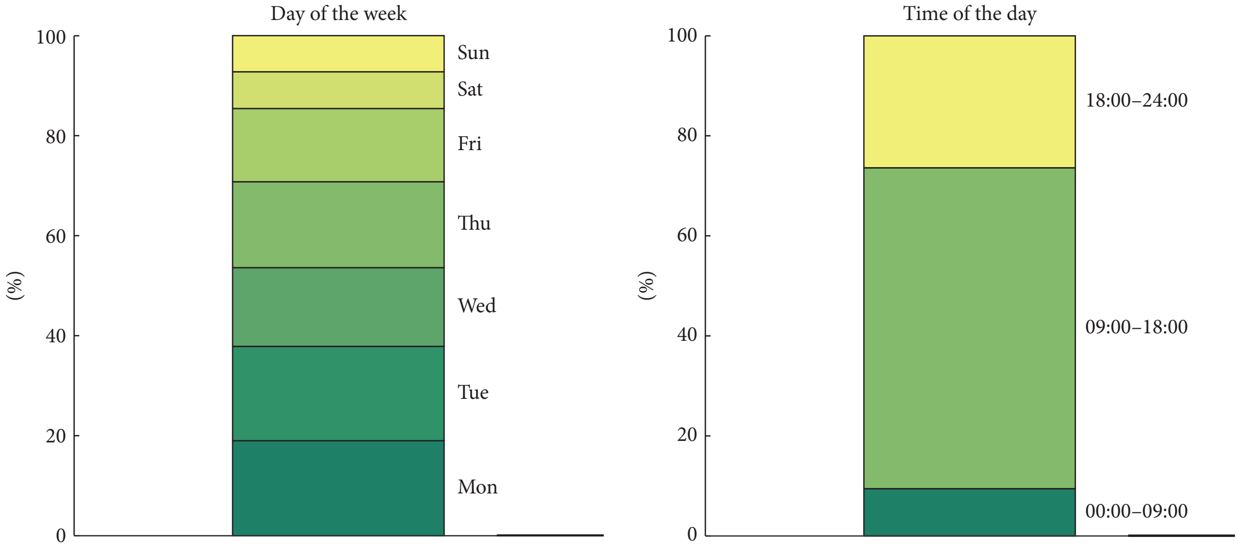


Figure 3: Distribution of the number of Bluetooth traces collected over days of the week and time periods of the day.

is estimated. Since these social similarities will be used as the input of the clustering algorithm, their values need to be measured rationally to ensure the final results of our analysis scheme.

Unlike existing studies that rely on simple context [12, 13, 19, 21] or locally maintained data [14–16, 22], we consider long-term global data of human interactions in order to design similarity metrics that exploit more contextual information for discovering the globally agreed social community structure. Specifically, the information of who meets whom and temporal positions of these contacts is used to estimate social similarities between users. Furthermore, the proposed similarity metrics are based on indirect human interactions as well as direct human interactions.

We first present encounter-rate-based similarity (ERS) and then describe encounters with temporal correlations similarity (ETCS). In ERS, the frequency of encounters is directly used to estimate the strength of social ties between two specific users. In ETCS, both spatial and temporal factors of encounters are considered using the encounters with temporal correlations matrix (ETC matrix) that contains the information of all users in the society. Moreover, in ETCS, the social closeness between two specific users are determined by their interactions with other members in the society. Therefore, ETCS can be considered as an indirect interaction-based similarity metric.

After the estimation of social intimacy, clustering is performed to partition the society of human-carried nodes into social groups, where the members of the same group maintain close relations. The intragroup members are supposed to have intimate interaction properties (e.g., high regularity and long-lasting meeting times), as well as correlated behaviors (e.g., interdependent movement patterns). These characteristics can be employed in the evaluation phase to validate the proposed methods.

4.1. Similarity Measurements

4.1.1. Encounter-Rate-Based Similarity (ERS). In this method, we estimate similarity based on the human-encounter rate. In reality, people who often meet each other tend to have a strong social link; thus, a higher encounter rate suggests that these users have a closer relationship. Therefore, we determine the pairwise similarity between two individuals using the rate of encounters between them:

$$w_{\mathrm{ERS}}(j,m)=\frac{N_{j,m}}{T_{j,m}}, \tag{4}$$

where $w_{\mathrm{ERS}}(j,m)$ is the encounter rate between users j and m, which is the ratio of $N_{j,m}$ (the number of times person j encounters m) to the number of their overlapping days, $T_{j,m}$.

According to our observations, the obtained rate of human encounters is not symmetric. There are many cases in which the values of $w_{\mathrm{ERS}}(j,m)$ and $w_{\mathrm{ERS}}(m,j)$ are not the same. The most probable reasons are asymmetric communications links and different Bluetooth logging times between users. To facilitate the usage of the encounter rate value as a similarity metric, which is the input for the further step of a social structure detection scheme, normalization is performed on those values. We use feature scaling to put the pairwise encounter rate values within the interval $[0\ \ 1]$. Then, the pairwise similarity between persons j and m is represented by the average of two values, $w_{\mathrm{ERS}}(j,m)$ and $w_{\mathrm{ERS}}(m,j)$. As noted, the higher value in the ERS metric indicates a closer pairwise relation between users.

4.1.2. Encounters with Temporal Correlations Similarity (ETCS). Now, we introduce a new similarity measurement method that is based on an encounter matrix, which represents both the temporal and spatial aspects of the pairwise human encounter.

Table 1: Sample M by $(M \times T \times L)$ ETC matrix. In this matrix, as an example, index $n_1 = 1 \times T \times L$ and $n_M = M \times T \times L$. Column $s_u d_v b_w$ corresponds to user u during time interval w of day v in the overlapping period.

	$s_1 d_1 b_1$	$s_1 d_1 b_2$	$\cdots$	$s_1 d_T b_L$	$\cdots$	$s_M d_T b_L$
$s(1)$	$a_{1,1}$	$a_{1,2}$	$\cdots$	a_{1,n_1}	$\cdots$	a_{1,n_M}
$s(2)$	$a_{2,1}$	$a_{2,2}$	$\cdots$	a_{2,n_1}	$\cdots$	a_{2,n_M}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$s(j)$	$a_{j,1}$	$a_{j,2}$	$\cdots$	a_{j,n_1}	$\cdots$	a_{j,n_M}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$s(M)$	$a_{M,1}$	$a_{M,2}$	$\cdots$	a_{M,n_1}	$\cdots$	a_{M,n_M}

The ERS method attempted to estimate the social closeness between users by employing pairwise encounters. However, it only considered the rate of encounters, rather than taking into account the time in which those encounters occurred; thus, the temporal correlation is not utilized. It is potential that embedding both temporal and spatial aspects can be effective in determining the level of human relationships.

We first create the encounters with temporal correlations matrix (ETC matrix) for all users in the society. The overlapping period is segmented into many different days. In order to facilitate identification of temporal factors of human interaction, each daytime period is divided into smaller intervals. We define L as the number of time intervals in a day. More specifically, given the daily period of concern containing most human interactions, the daytime is split into L equal-sized time intervals. As noted in Section 3, we only consider the daytime from 09:00 to 18:00, which is nine hours long. If we choose a time interval of one hour, in each day there are $L = 9$ time intervals (e.g., time interval 1 is 09:00 to 10:00, time interval 2 is 10:00 to 11:00, and so on). Let T be the number of days in the overlapping period, in which all users are active almost all the time. After that, from M users and T days, we construct M by $M \times T \times L$ matrix representing encounters between users during every time interval of the overlapping period. Each of the M rows corresponds to one user. On the other hand, each individual in the society is also associated with a block of $T \times L$ consecutive columns, and the column blocks are formed following the order from user 1 to user M. In every block, one column corresponds to one time interval of one day during the overlapping period.

An element in the matrix displays the number of pairwise encounters between two individuals. The within-block indexes of these columns indicate the temporal positions of the encounters, starting at time interval 1 on day 1, and then follow the order of real time to interval L of day T. As can be seen in the sample matrix in Table 1, the first block of $T \times L$ columns corresponds to user 1. As we can see, it has the start index as 1 and the end index as n_1, which indicate two temporal positions in the overlapping period: time interval 1 of day 1 and time interval L of day T, respectively. For instance, column index 2 is associated with user 1 at time interval 2 on day 1, and, thus, the matrix element $a_{M,2}$ is the number of times user M encountered user 1 during time interval 2 on day 1 of the overlapping period.

From this ETC matrix, we determine the social closeness of users. As can be observed from the sample matrix, a row vector of user j, which is denoted as $s(j)$, represents the encounters between j and every other user in the society (e.g., the row vector of user 2, $s(2) = [a_{2,1}, a_{2,2}, \ldots, a_{2,n_M}]$ represents the encounters of user 2 with every user from 1 to M). Therefore, we note that it can be useful to determine the similarity between users i and j by comparing their representative vectors that depict all interactions between the entire society and each of them.

It should be emphasized that when we determine the intimacy between users i and j by comparing row vectors $s(j)$ and $s(i)$, the direct interactions between users i and j may not be profitable. If they meet each other many times, the vector elements associated with user j (corresponding to column block j in the matrix) in $s(i)$ that represent the number of encounters between i and j will be filled with values other than zero, while the corresponding elements in $s(j)$, showing the times user j encountered himself or herself, will remain all zeros. Therefore, it is clear that the number of direct encounters between two individuals is not helpful in obtaining the similarity between them. Only their interactions with other members in the society will be used to measure their social closeness. Thus, the metric can be regarded as an indirect interaction-based method.

We aim to evaluate the social closeness between individuals by measuring the distance of their social interaction vectors. The constructed ETC matrix is promising in examining the relationships between users. However, in general, human interactions between users are not available in every time interval during a long overlapping period, and, thus, their encounter data would be absent (represented by elements with zeros in the matrix). Therefore, the Euclidean distance-based approaches are not appropriate for determining similarity between users, because it is highly possible that the matrix contains sparse data. Therefore, we measured the social closeness between individuals based on cosine similarity, which is efficient for sparse data because it will ignore coabsences in its computation [26]. From the ETC matrix, we compute the social similarity between individuals as follows:

$$w_{\mathrm{ETCS}}(j, m) = \frac{s(j) \cdot s(m)}{\|s(j)\| \, \|s(m)\|}, \tag{5}$$

where $w_{\text{ETCS}}(j, m)$ is the similarity between users j and m and $s(j)$ and $s(m)$ are the row vectors of j and m, respectively.

4.2. Clustering Methods. After determining the social similarities between individuals, the next step is to uncover the structure of the society. By using clustering algorithms, mobile users are partitioned into social groups. In this paper, we consider spectral clustering and self-organizing map.

4.2.1. Spectral Clustering. Spectral clustering [24] is a simple and efficient algorithm that tends to outperform the traditional clustering approaches. In addition, this graph-based method can transform pairwise similarity or distance into neighborhood connections between network nodes.

First, the adjacency matrix in which elements represent the local neighborhood similarities between individuals is formed. This process maps the similarity or distance into neighborhood relations between network nodes. The proposed similarity metrics between users are transformed into elements of the adjacency matrix.

In case of the eigenbehavior-based method, we use the Gaussian kernel function [24] to obtain elements for adjacency matrix A from the behavioral distances between users:

$$A(i, j) = \exp\left(-\frac{\left\|\Omega^j - \Omega^m\right\|^2}{2\sigma^2}\right) = \exp\left(-\frac{d^2(i, j)}{2\sigma^2}\right). \quad (6)$$

The matrix elements with value 0 indicate that there is no link between users (except $A(i, i) = 0, \forall i$), while a value of 1 demonstrates the highest similarity.

In case of encounter-rate-based and encounters with temporal correlations similarity methods, we directly form the adjacency matrix since those proposed similarities can represent the local neighborhood relationships, and their values are already in a suitable interval.

Based on the adjacency matrix, we construct a Laplacian matrix using the symmetric normalized technique [27]. Then, the set of eigenvectors of the Laplacian matrix is calculated. Before clustering users into k groups, we represent users in a lower-dimensional space, $\mathbb{R}^{M\times k}$, which is formed by the first k eigenvectors (eigenvectors that correspond to the k lowest eigenvalues). The final step is to use K-means algorithm on this data space, in which each row corresponds to a person in the society, and we then obtain the social groups. Because different initializations may give different results, we execute the K-means algorithm multiple times and then choose the result that minimizes the sums of within-cluster point-to-centroid distances. Specifically, given clustering result ϕ, let S_i be the set of data points of cluster i among k clusters, and let μ_i represent its centroid. Then, the result that minimizes loss function $\mathscr{L}(\phi)$ is chosen, where $\mathscr{L}(\phi) = \sum_{i=1}^{k} d_i(\phi)$, and $d_i(\phi) = \sum_{j\in S_i(\phi)} \|\Omega^j - \mu_i(\phi)\|^2$.

4.2.2. Self-Organizing Map. In this paper, we also consider a self-organization map (SOM). The SOM method is able to discover clusters through the unsupervised training process [28].

Technically, a SOM is an artificial neural network. The SOM network is only made up of two layers: the input layer (training data) and the output layer. Hereafter, the term neural map will refer to the output layer of the SOM. The typical topology of the neural map in a SOM is a grid. Each neuron in the output layer is fully connected to each node in the input layer. The number of nodes in the input layer is identical to the dimension of an instance of input data. In a SOM, each neuron has a spatial location and is associated with a weight vector, which has the same dimension as the input vector.

The training process of a SOM is carried out over numerous iterations. First, the weight vector of each neuron is randomly initialized. At each training iteration, a sample is arbitrarily chosen from the input data set. After that, the distance between input sample and neuron (represented by a weight vector) in the map is computed. There are many techniques for distance measurement, and Euclidean distance is the most widely used for this operation. Then, the best matching unit, a neuron where the weight vector is closest to the current input sample, is identified. Next, the winner determines the spatial neighbors on the grid of neurons that will have to adapt their weight vector.

Here, before presenting the adaptation procedure of the SOM, we describe the training parameters. Given the current winning neuron, b, and its neighboring, k, the parameter $h_{bk}(t)$, which is the neighborhood kernel centered on the winning neuron, reflects the influence of the distance between k and b to the training rate of k. At iteration t, it is calculated as follows:

$$h_{bk}(t) = \exp\left(-\frac{\left\|r_b - r_k\right\|^2}{2\sigma^2(t)}\right), \quad (7)$$

where r_b, r_k are coordinates of b and k in the map, respectively, and $\sigma(t)$ decreases with time t. Another parameter is $\alpha(t)$, the learning rate, which also declines with time.

In the adaptation step, each neuron among the winner and its neighbors updates its own weight vector as follows:

$$W_k(t+1) = W_k(t) + \alpha(t)\, h_{bk}(t)\left[X - W_k(t)\right], \quad (8)$$

where X is the input sample in the current iteration.

Based on (8), obviously, the winning neuron will have the largest adaptation rate. Repeatedly, for each input vector, the winning neuron is determined; then it and the nearby neighbors update their weight vectors, according to the procedure described above, until the algorithm converges.

For the purpose of determining the clusters over the input data, a number of neurons in the network can be chosen that is equal to the predetermined number of clusters. This choice is practical in our scheme, since there are a limited number of input samples (43 users). Individually, each neuron of the SOM output layer becomes a cluster center. After the training phase, a sample from the data set will be mapped to the cluster where the center is closest to it.

We select our similarity measurement as the input data for SOM clustering. Since a SOM uses Euclidean distance, the encounters with temporal correlations may not be the most suitable candidate. Therefore, we use the encounter-rate-based similarity as input for this clustering algorithm. Here, each encounter-rate-based vector represents the interactions of an individual with the entire society. More specifically, each individual user is represented by an M-dimensional vector (in our scheme, $M = 43$), in which the vector components are the rates of encounters between that person and other members of the society. Thus, by applying SOM, we aim to reveal the relations between users by comparing their interactions with others, similar to the interaction-based manner of ETCS.

5. Human Mobility Prediction Based on Social Groups

After discovering the community structure, we use the resulting social groups to predict human mobility. In this section, first, we will explain how we process the raw traces to extract meaningful mobility data. After that, the prediction model will be presented, which embeds contextual features from social-group-mates to infer the location of a given user. Human mobility prediction accuracy will be used to validate our proposed social community detection schemes.

5.1. Location Extraction. To indicate the location of a mobile user, the MIT Reality Mining Dataset provides cellular network-based mobility data. The traces record the history of cell tower IDs associated with a phone every time it changes the cell towers. We use symbolic locations that correspond to the associated cell tower at a certain time.

In this paper, we are interested in predicting human movement patterns in the daytime period between 09:00 and 18:00 on weekdays, similar to the social structure detection phase. Based on the raw cell tower log, we can determine the period that a human-carried device is connected to a particular cell tower. However, due to the nature of human movement and signal fading, the time that a phone is connected to the cell tower is volatile. In some cases, a phone stays in a cell for a long time, while there are lots of other cases where a mobile device is only associated with a cell over a very short period and switches to other cells rapidly. Therefore, by using raw information, it is hard to identify accurate positions of users for the purpose of modeling human movements.

In order to use appropriate location data, we split the daytime period into smaller, equal-sized time slots, which are suitable for capturing general mobility. For the time from 09:00 to 18:00, if we set time slot length t_{TS} at 30 minutes, then we will have 18 time slots in each day. During one time slot, the phone may switch to many different cell towers. In this study, the location of a mobile user in a given time slot is considered to be the cell tower that accounts for the largest amount of time in the time slot. We also denote threshold ratio value α for location extraction. In a time slot, the period that a device is associated with the largest amount of time location (i.e., the cell ID) needs to exceed αt_{TS}; otherwise, the location of that mobile user is considered to be undefined. In this paper, α is 0.3. Note that not every individual participated in the experiment at a certain time; that is, some of them may have turned off their phones. It is also possible that mobile devices may receive no cellular signal. The locations of mobile users in these situations are undefined and set to zero.

Table 2: Contextual variables for location prediction.

Variable	Description
Y	Location of a given user
X_1	Day of the week
X_2	Time slot of the day
X_3	Location of social-group-mate 1
X_4	Location of social-group-mate 2
$\cdots$	$\cdots$
X_{n+2}	Location of social-group-mate n

Based on this procedure, we obtained the location for each mobile user in the society's given time index. The temporal position can be depicted by the time slot in the day for the day of the week. The extracted temporal and spatial positions of users were used as contextual data for the human mobility prediction model.

5.2. Prediction Model. Our motivation for social structure detection is to understand the behavior of individuals as well as human interactions. Based on the social groups uncovered through community detection schemes, the human behavior like movement can be better understood. In general, there are many contextual features that affect real-life human movements [29]. In addition to the potential of embedding temporal and spatial factors in mobility prediction [30], the movement of people can be reflected by social relations. We can potentially infer an individual's movement pattern by using contextual information on whom he/she has a strong social link with. At a particular time, if provided the locations of those who are closely related to a user, then, based on that information, we can make predictions about user's whereabouts. For example, on Monday from 10:00 to 10:30, if all social-group-mates of a given user are in the seminar room, then it is highly probable that the user is also there for a Monday meeting.

On the other hand, human mobility prediction can be a useful evaluation tool for social community detection. Different social community detection schemes result in different community structures. If the location of users is predicted most accurately based on the social groups obtained by a community detection algorithm, then that algorithm can be considered the most appropriate one from among the proposed methods. Therefore, movement prediction accuracy can be used as a validation method for social structure detection schemes.

In this paper, we consider using Naïve Bayes for human mobility prediction. Since the objective of predicting human movement is to evaluate the obtained social factors (community structures obtained by using the schemes described in Section 4), we use the maximum likelihood estimation approach due to its simplicity and effectiveness. As shown in Table 2, to predict the location of a given user, we consider the temporal factors, including day of the week and time slot of the day. Regarding the social factors, we utilize location

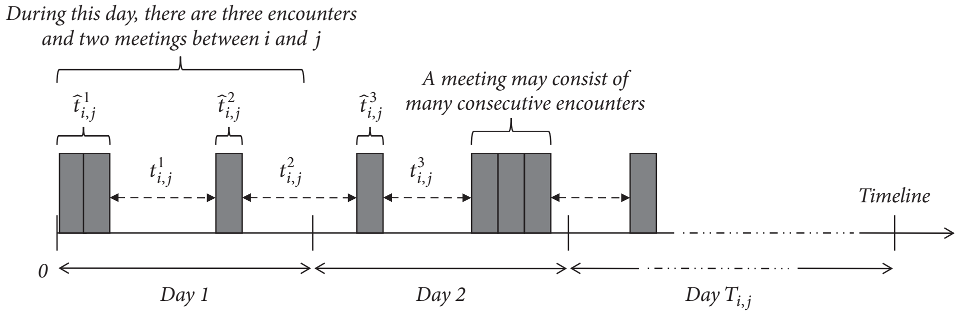


FIGURE 4: Example of the meeting history between person i and person j during the overlapping period of $T_{i,j}$ days. $t^x_{i,j}$ is the time interval between meetings, and $\hat{t}^x_{i,j}$ is the meeting time. Each shaded box represents one encounter.

information from social-group-mates of a given user to predict his/her location. For instance, if the user belongs to a group of five persons, then the locations of his/her four group-mates will be used. In some exceptional cases from our social community detection scheme, where a group consists of only one user, the social factor will be ignored. Given the contextual variables described in Table 2, the predicted location of a user can be determined as follows:

$$\begin{aligned} Y^{\text{predict}} &= \arg\max_{v} P\left(X_1 = x_1 \mid Y = v\right) \\ &\quad \cdot P\left(X_2 = x_2 \mid Y = v\right) \cdots \left(X_{n+2} = x_{n+2} \mid Y = v\right) \\ &\quad \cdot P\left(Y = v\right), \end{aligned} \tag{9}$$

where $x_1, x_2, \ldots, x_{n+2}$ is the input attribute of contextual variable $X_1, X_2, \ldots, X_{n+2}$, and v is an element in the set of possible values of Y (location of the given user).

We randomly select a part of the data as a training set and use the rest as a test set to verify the prediction model. In our model, we chose a training/test data ratio of $8:2$, which means the data from 38 randomly chosen days are utilized in the training step, and data from the remaining 16 days are used to validate the prediction model. Due to the variety of symbolic locations (cellular towers), the zero observation problem can occur where conditional probability $P(X_i = x_i \mid Y = v)$, $\forall v$ ends up with a zero value (i.e., the situation where the location of a social-group-mate never appears in the training set). To avoid this problem, the conditional probability is calculated following the Laplace smoothing technique [31]:

$$P\left(X_i = x_i \mid Y = v\right) = \frac{1 + \operatorname{count}\left(X_i = x_i \mid Y = v\right)}{k + \operatorname{count}\left(Y = v\right)}, \tag{10}$$

where k is the total number of possible values of X_i.

One remaining issue for our prediction model is the existence of undefined locations in the data due to turned-off mobile devices, network connectivity problems, and frequent changes in associated cells. Symbolic location 0 provides no useful information on human mobility or interaction; thus we remove contextual data that contains these meaningless locations. At a particular temporal position, if any spatial position of a user or his/her social-group-mates is undefined, then the data corresponding to this time duration are excluded from our prediction model. Although the amount of validation data may be reduced, we intend to ensure the meaningfulness of all input and output contextual variables.

6. Evaluation Results and Discussion

In this section, we present the performance analysis of the proposed community detection schemes. We first describe the friendship-based and movement prediction-based evaluation methods. Then, we discuss the evaluation results.

6.1. Friendship-Based Evaluation Method. The friendship-based evaluation method is rooted in the natural characteristics of human relationships. We aim to evaluate the social community detection scheme by measuring the degree of intragroup intimacy among society members. According to a study in [15], human relations can be reflected by three basic properties that define friendship; that is, the closeness of two persons is recognized through the frequency, regularity, and duration of their interactions. In the context of our work, human interaction is represented by the encounter. Thus, we estimate these characteristics based on the encounter history between users. We define the metrics for each of the three friendship properties.

Figure 4 illustrates a sample encounter history between two persons. The first friendship property we consider is frequency. We define the frequency metric between two individuals as the ratio of the total number of encounters between them to the total number of days in their overlapping period. The pairwise frequency metric between persons i and j is calculated as follows:

$$\text{Freq}_{i,j} = \frac{N_{i,j}}{T_{i,j}}, \tag{11}$$

where $N_{i,j}$ is the total number of encounters between i and j during their overlapping period of $T_{i,j}$ days.

The second metric is regularity. The interaction between two individuals is regular if they have consistent intermeeting time intervals (i.e., small variance). Another consideration to be taken is the time scale. Given the same variance value, a larger scale of time intervals between two individuals' meetings indicates more regular interactions. Therefore, the regularity between two persons can be measured as the inverse of the index of dispersion [32] of their intermeeting time intervals or the ratio of the mean to the variance of the intermeeting time intervals in their overlapping period. That is

$$\text{Reg}_{i,j} = \frac{E\left(t^{x}_{i,j}\right)}{\text{Var}\left(t^{x}_{i,j}\right)}, \tag{12}$$

where $t^{x}_{i,j}$ is a time interval between two consecutive meetings of persons i and j.

The third characteristic, longevity, is demonstrated by the time length of the interactions between individuals. If, in general, every meeting between two persons lasts for a long time, that suggests they have a close relationship. We define the metric of longevity as the mean time length of meeting events between two persons:

$$\text{Long}_{i,j} = \frac{1}{n_{i,j}} \sum_{x=1}^{n_{i,j}} \hat{t}^{x}_{i,j}, \tag{13}$$

where $n_{i,j}$ is the total number of meetings between persons i and j. Recall that the length of a meeting between persons i and j is calculated as $\hat{t}^{x}_{i,j} = 2.5 \times N^{x}_{i,j}$ minutes, where $N^{x}_{i,j}$ is the number of encounters corresponding to this meeting.

After we obtained those values for frequency, regularity, and longevity, we take the average value to represent the pairwise relationship metric in order to eliminate the asymmetry of the calculated metrics (e.g., the difference between $\text{Long}_{i,j}$ and $\text{Long}_{j,i}$, if it exists). Next, for each of the three metrics, their values are normalized to the interval $[0, 100]$. Given community detection scheme Z that partitioned the society into k social groups, the overall performance of the social structure detection method is calculated as follows:

$$\text{rel}\left(Z_k\right) = \frac{\sum_{n=1}^{k} \sum_{i,j \in G_n, i \neq j} r^{n}\left(i, j\right)}{\sum_{n=1}^{k} R^{n}}, \tag{14}$$

where overall relationship value $\text{rel}(Z_k)$ corresponds to one of the three friendship metrics above; G_n is the set of users in group n, $r^{n}(i, j)$ represents pairwise relationship value between users i and j within group n, and R^{n} is the number of user pairs in group n.

6.2. Movement-Prediction-Based Evaluation Method. In Section 5, we presented a prediction model that uses the results obtained from social community detection schemes. The prediction accuracy of the prediction model embedding social factor is used for an evaluation metric. Here, the social factor is extracted from the obtained social communities. The high mobility prediction accuracy indicates that the social-group-mates have strongly correlated behaviors, which accordingly signifies a better community detection scheme.

Using the community detection scheme, the society is partitioned into a given number of social groups. Suppose that the society is partitioned into k social groups. Then, the average prediction accuracy based on a social community detection scheme Z (e.g., encounters-rate-based similarity with spectral clustering) is determined as follows:

$$\text{acc}\left(Z_k\right) = \frac{\sum_{i=1}^{M} \text{true}^{Z}_{k}\left(i\right)}{\sum_{i=1}^{M} \text{total}^{Z}_{k}\left(i\right)}, \tag{15}$$

where M is the number of individuals ($M = 43$ in this paper), $\text{total}^{Z}_{k}(i)$ is the number of to-be-predicted cases for user i, and $\text{true}^{Z}_{k}(i)$ is the number of correct predictions about user i.

Since the society may be partitioned into various numbers of groups, the overall prediction accuracy based on the social factor of social community detection scheme Z is computed as follows:

$$\text{acc}\left(Z\right) = \frac{\sum_{k=k_{\text{lowest}}}^{k_{\text{largest}}} \sum_{i=1}^{M} \text{true}^{Z}_{k}\left(i\right)}{\sum_{k=k_{\text{lowest}}}^{k_{\text{largest}}} \sum_{i=1}^{M} \text{total}^{Z}_{k}\left(i\right)}, \tag{16}$$

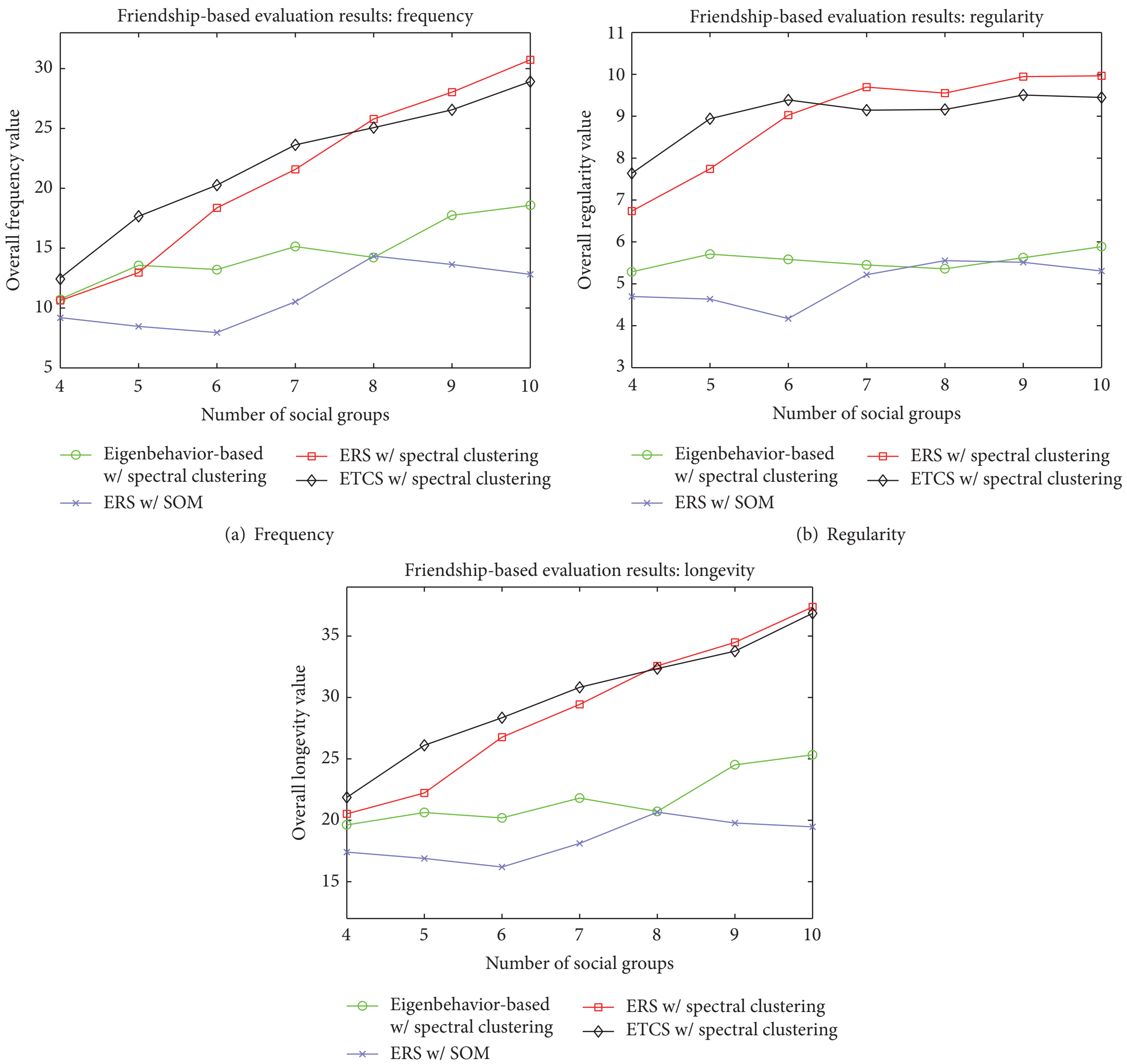
where k_{lowest} and k_{largest}, respectively, correspond to the smallest and largest number of groups the society is partitioned into.

In the prediction model employing Naïve Bayes, we recognize the most likely location to be the output for the predicted location of a given user. In addition, we also consider an extension by taking into account the second most likely location. To be more specific, for every prediction, we consider not only the location with the highest predicted probability, but also the location with the second highest probability. Then, in this extension, if the real to-be-predicted location is identical to the most likely or second most likely locations, we classify it as a correct prediction. Since this modification probably enhances the capability of the current prediction model, it can provide useful validation for our movement-prediction-based social community detection evaluation.

6.3. Results and Discussion. We collected the results using the proposed community detection schemes, ERS (encounter-rate-based similarity) with spectral clustering, ERS with SOM, and ETCS (encounters with temporal correlations similarity) followed by spectral clustering. The results are compared with the eigenbehavior-based approach.

6.3.1. Friendship-Based Evaluation Results. Figure 5 shows the evaluation results of community detection schemes using the friendship-based metrics, where the number of social groups to be found varied between 4 and 10. In each case, the overall values of longevity, frequency, and regularity metrics were obtained.

As shown in Figure 5, ERS and ETCS with spectral clustering show better performances than the eigenbehavior-based method at detecting close relationships at the community level. The eigenbehavior-based metric was outperformed by these two similarity measurements in every friendship property. The eigenbehavior-based similarity measurement only considers the number of users that each individual encountered, while ERS and ETCS embed a more specific context of which users were in contact with him/her. The



(a) Frequency

(b) Regularity

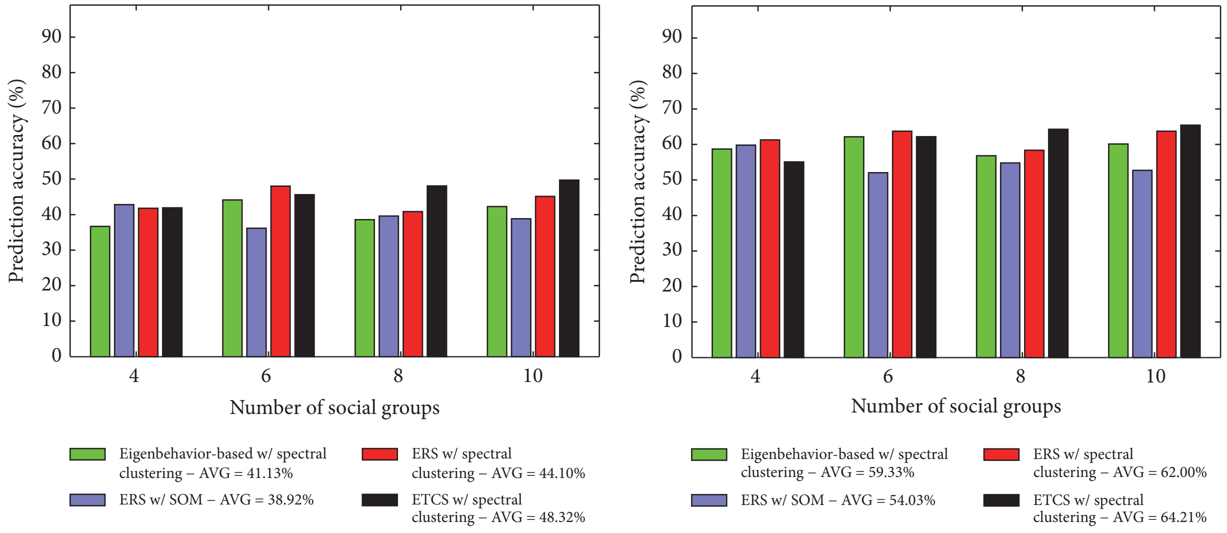
(c) Longevity

Figure 5: Overall friendship-based evaluation results returned from the community detection schemes.

less contextual information is the reason for the lower performance of the eigenbehavior-based similarity method, compared to the others.

As can be seen in Figure 5, ETCS with spectral clustering outperforms other methods when the number of social groups is fewer than eight. For example, when the number of social groups is five, the scheme employing ETCS gains 35% in frequency, gains 15% in regularity, and gains 18% in longevity, compared to the ERS with spectral clustering scheme. This indicates that employing both the temporal and spatial aspects of human interactions can result in better estimations of their relationships. However, when the number of social groups increases, this advantage fades, compared with ERS. Note that if the number of social groups increases, the sizes of the social groups decrease. When analyzing a society of 43 users, if k is larger than eight, the size of many groups is reduced to only two or three members. In these cases, the direct estimation from the human-encounter rate can be more efficient in revealing the relations of group-mates than ETCS. Overall, when the size of social groups is small, it is useful to use the ERS metric. When the sizes of social groups increase, using the ETCS metric is more beneficial.

It is also worth discussing the performance comparison of two clustering algorithms. In general, SOM can accurately obtain the cluster structure of data when a sufficient data amount is available. Recall that the number of clusters we aim to identify can be up to 10, while the input only consists of 43 instances. Given these input data, in our scheme,



(a)

(b)

Figure 6: Movement-prediction-based evaluation results, where (a) shows the prediction accuracy based on the most likely location and (b) displays the accuracy when we include the second most likely location to the prediction output.

SOM was required to train to map each cluster with a few input instances. Due to the small-scale size of the data, the training process of SOM may not be able to return the precise mapping. Based on the observations from Figure 5, spectral clustering shows better performance for our social community detection schemes.

In the friendship-based evaluation results, three different properties of a relationship show a similar trend. The performance of different social community detection schemes is analogous in these three cases. Furthermore, the overall values for the friendship metrics increase when the society is divided into a larger number of groups. This indicates that the three properties of a relationship are highly related. Moreover, if the society is divided into a larger number of social groups, we can obtain higher friendship-based closeness at the community level. In these cases, each social group consists of a few of the socially closest friends; members of the same group are usually in contact with each other, thus having very high degree of social closeness.

6.3.2. Movement Prediction-Based Evaluation Results. Figure 6 shows the prediction accuracy of the Naïve Bayes model based on the proposed social community detection schemes for four different cases when the number of social groups changes to 6, 8, and 10 from 4. For movement prediction-based evaluation, we selected ten different training/test sets from the dataset, then collected the average accuracy over those ten sets.

A similar pattern of evaluation results between the most likely location and second most likely location can be observed in Figure 6. Obviously, when the second most likely location is incorporated into the predicted output, the accuracy is higher.

As shown in Figure 6, the performance of the social community detection schemes changes with different numbers for k. When k is small, the ETCS with spectral clustering scheme does not show remarkable performance, compared to other schemes.

Note that when the society is partitioned into a small number of social groups, k, the group size may be large (i.e., there can be a number of individuals belonging to the same group). At a specific temporal position, some users' locations can be undefined. Recall that when evaluating the movement of an individual, undefined location data about members of his/her social group are removed from our prediction model. Thus, for some cases, the amount of remaining contextual data may be small for the validation of human movement. Since with a larger k the size of social groups becomes smaller, the impact of this issue is reduced, and, thus, the evaluation is more reliable. In these cases, the amount of validation data is satisfactory for assessing the performance of the social factor-based prediction model. Therefore, the results when the society is partitioned into a large number of groups may better illustrate the capability of our model. When the number of social groups is large, as can be observed in Figure 6, the ERS with spectral clustering and the ETCS with spectral clustering schemes display higher movement-prediction-based performance than the other schemes. ERS and ETCS metrics are better measurements of social closeness than eigenbehavior-based similarity in both friendship properties and the mobility-dependency criterion.

As we can also see in Figure 6, similar to the friendship-based evaluation, the scheme using SOM shows a low performance compared to those utilizing spectral.

Figure 7 shows individual mobility prediction accuracy when the society of mobile users is partitioned into 10 social groups. The scheme employing ETCS outperforms the other methods. The use of social factors based on this scheme gives impressive individual prediction accuracy for most of the users in the society.

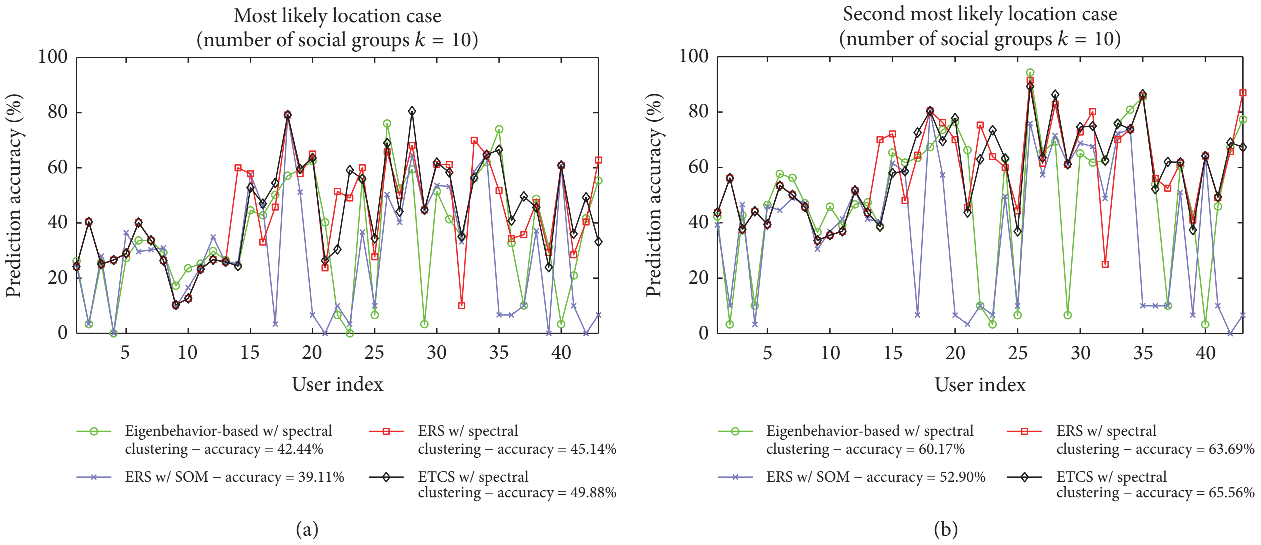


Figure 7: Individual accuracy of the social factor-based mobility prediction model when the number of social groups is 10.

As we can see in Figure 7, there are some cases (when using social groups obtained from ERS with SOM and eigenbehavior-based similarity with spectral clustering) where individual prediction accuracy is extremely low; that is, the social groups obtained from these two schemes are not useful for the prediction model. In addition, the social groups acquired by using these schemes are unbalanced. For example, among the social groups obtained by the ERS with SOM scheme, one group has the number of users at 15 (much larger than groups that only consist of one or two members). Similarly, using the eigenbehavior-based scheme, there is a social group that consists of eight users while another group has only one user; but it is not desirable to divide the society into small partitions.

The results of the movement prediction-based evaluation method indicate that, by embedding the social factor, we can gain considerable prediction accuracy. In addition, the gain is larger if the society is partitioned into a large number of social groups. This is because if the size of a social group can be reduced to a few members, then the group members tend to be highly correlated in their movement decisions. For instance, when the number of clusters is 10, the scheme incorporating the ETCS metric can be employed to achieve accuracy of 50% for the most likely location and more than 65% when including the second most likely location. That is followed by ERS with spectral clustering, which can contribute prediction accuracy of 45% and 63%, respectively, in these two cases.

As can be observed in Figure 7, mobility prediction accuracy varies from user to user, thus implying that different persons may have different levels of behavioral interdependence. Some persons seem to have less social dependence. Using contextual information from the social-group-mates to predict their movement results in limited efficiency. On the other hand, the majority of users tend to have a high level of social dependence in the mobility pattern, where their movements can be more predictable by embedding the social factor. Those results suggest that the social factor is the potential feature in mobility prediction model.

It should be highlighted that there is a different tendency in evaluation results between the movement prediction-based and friendship-based metrics. In contrast to the friendship properties method, even with a large k, using the ETCS metric still maintains higher performance than the ERS metric. The results suggest that, by embedding the temporal factor of human interactions in measuring the social closeness between people, the behavioral dependencies are better reflected in the similarity metric. In general, if two persons have an abundant pairwise encounter history, their relationship can be reflected in large values of friendship property metrics, but it may not always imply high dependency between their behaviors. As an example, we consider a case where three users (A, B, and C) belong to the same organization, and each of them has social relationships with the others. In our example, A and B are teammates, while A and C are personal friends, but they work in different teams. Because there is a personal relationship between A and C, they usually spend time with each other. Thus, compared to A and B, A and C may have higher friendship property metrics values. The fact that A and B are socially close (from the same team) signifies that they have a correlated daily work routine. Therefore, although the two social friends, A and B, may have smaller values in friendship-based metrics than the two personal friends, A and C, the degree of behavioral dependency (e.g., in movement) between A and B is probably higher. The similarity measurement that examines human interactions considering both temporal and spatial aspects will be effective in detecting the members of a social group who are strongly interdependent in movement patterns.

7. Conclusion

In this paper, we investigated social network structure focusing on the community detection issue. Our objective is to determine social groups among individuals in human society. We used a real-life mobility dataset with interactions between people that are represented by Bluetooth-based encounters, as well as human movement data derived from cellular network traces. This contextual information allows us to examine the communities within a society of mobile users. We introduced methods that measure social similarity between individuals by analyzing human contact history. By applying social community detection schemes, we partitioned the society into social groups of users. We also proposed evaluation methods that employ friendship properties and mobility dependence between people.

The performance results demonstrated that ETCS and ERS methods outperform the eigenbehavior-based similarity measurement. The friendship-based evaluation results show that ERS is better than ETCS when the size of social groups is small. On the other hand, the ETCS method gives better performance when the sizes of social groups increase. In the movement prediction-based criterion, the scheme employing ETCS dominates other methods. We also found that the graph-based spectral algorithm is better than SOM in social structure detection.

We explored the potential in employing temporal and spatial aspects of human interactions to the similarity metric, especially in determining behaviorally dependent persons. We found that a social factor-based on which people are socially intimate with one another is a promising contextual feature in modeling human movement patterns. The evaluation results also suggest that partitioning society into a large number of social groups increases the degree of behavioral interdependence and friendly intimacy between individuals at an intragroup level.

Based on the proposed schemes, we can determine the social groups of the users with close relationships and interdependent social characteristics, thus enhancing an understanding of the system of interactions and behaviors in human society. Being able to identify these social structures can facilitate further application of collaborative networks, such as urban data mining and social-based opportunistic data routing.

It is worthwhile noting that in this work a closed community where people have long-term affiliations and similar social backgrounds was considered. However, the type and size of community can affect the results of the proposed community detection schemes. As a future study, we plan to consider various community types and sizes and their effects on the performance of the proposed schemes and extend the schemes accordingly.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1D1A3B03934617).

References

[1] M. Plantié and M. Crampes, "Survey on social community detection," in *Social Media Retrieval*, Computer Communications and Networks, pp. 65–85, Springer London, London, 2013.

[2] Y. Zhu, B. Xu, X. Shi, and Y. Wang, "A survey of social-based routing in delay tolerant networks: positive and negative social effects," *IEEE Communications Surveys and Tutorials*, vol. 15, no. 1, pp. 387–401, 2013.

[3] N. Eagle, A. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 36, pp. 15274–15278, 2009.

[4] S. Fortunato, "Community detection in graphs," *Physics Reports. A Review Section of Physics Letters*, vol. 486, no. 3-5, pp. 75–174, 2010.

[5] S. Fortunato and D. Hric, "Community detection in networks: a user guide," *Physics Reports. A Review Section of Physics Letters*, vol. 659, pp. 1–44, 2016.

[6] E. Estrada, N. Hatano, and M. Benzi, "The physics of communicability in complex networks," *Physics Reports. A Review Section of Physics Letters*, vol. 514, no. 3, pp. 89–119, 2012.

[7] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[8] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 23, pp. 8577–8582, 2006.

[9] M. E. J. Newman, "Detecting community structure in networks," *European Physical Journal B*, vol. 38, no. 2, pp. 321–330, 2004.

[10] V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, Article ID P10008, 2008.

[11] Z. Lu, X. Sun, Y. Wen, G. Cao, and T. L. Porta, "Algorithms and Applications for Community Detection in Weighted Networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 11, pp. 2916–2926, 2015.

[12] E. M. Daly and M. Haahr, "Social network analysis for routing in disconnected delay-tolerant MANETs," in *Proceedings of the 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc'07)*, pp. 32–40, September 2007.

[13] T. Hossmann, F. Legendre, and T. Spyropoulos, "From contacts to graphs: Pitfalls in using complex network analysis for dtn routing," in *Proceedings of the IEEE INFOCOM Workshops 2009*, April 2009.

[14] P. Hui, E. Yoneki, S. Chan, and J. Crowcroft, "Distributed community detection in delay tolerant networks," in *Proceedings of the 2nd ACM/IEEE International Workshop on Mobility in the Evolving Internet Architecture (MobiArch '07)*, 2007.

[15] E. Bulut and B. K. Szymanski, "Exploiting friendship relations for efficient routing in mobile social networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 12, pp. 2254–2265, 2012.

[16] F. Li and J. Wu, "LocalCom: a community-based epidemic forwarding scheme in disruption-tolerant networks," in *Proceedings of 6th Annual IEEE Communications Society Conference*

on Sensor, Mesh and Ad Hoc Communications and Networks, pp. 1–9, Rome, Italy, June 2009.

[17] B. Ball, B. Karrer, and M. E. J. Newman, “Efficient and principled method for detecting communities in networks,” *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 84, no. 3, Article ID 036103, 2011.

[18] N. P. Nguyen, T. N. Dinh, S. Tokala, and M. T. Thai, “Overlapping communities in dynamic networks: their detection and mobile applications,” in *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, pp. 85–95, ACM, September 2011.

[19] S. Pandit, Y. Yang, V. Kawadia, S. Sreenivasan, and N. V. Chawla, “Detecting communities in time-evolving proximity networks,” in *Proceedings of the 2011 IEEE 1st International Network Science Workshop, NSW 2011*, pp. 173–179, usa, June 2011.

[20] D. Boston, S. Mardenfeld, J. Pan, Q. Jones, A. Iamnitchi, and C. Borcea, “Leveraging Bluetooth co-location traces in group discovery algorithms,” *Pervasive and Mobile Computing*, vol. 11, pp. 88–105, 2014.

[21] N. Eagle and A. S. Pentland, “Eigenbehaviors: identifying structure in routine,” *Behavioral Ecology and Sociobiology*, vol. 63, no. 7, pp. 1057–1066, 2009.

[22] M. J. Williams, R. M. Whitaker, and S. M. Allen, “Decentralised detection of periodic encounter communities in opportunistic networks,” *Ad Hoc Networks*, vol. 10, no. 8, pp. 1544–1556, 2012.

[23] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[24] U. von Luxburg, “A tutorial on spectral clustering,” Max Planck Institute for Biological Cybernetics 149, 2006.

[25] N. Eagle and A. Pentland, “Reality mining: sensing complex social systems,” *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006.

[26] J. D. Kelleher, B. M. Namee, and A. DArcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, MIT Press, and Case Studies, 2015.

[27] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in *in Advances in Neural Information Processing Systems*, pp. 849–856, MIT Press, 2001.

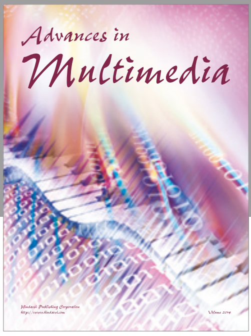
[28] J. Vesanto and E. Alhoniemi, “Clustering of the self-organizing map,” *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586–600, 2000.

[29] T. M. T. Do and D. Gatica-Perez, “Where and what: Using smartphones to predict next locations and applications in daily life,” *Pervasive and Mobile Computing*, vol. 12, pp. 79–91, 2014.

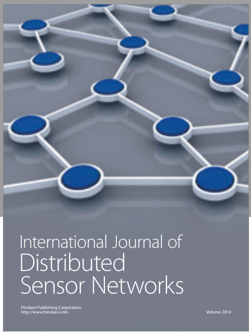
[30] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, “Limits of predictability in human mobility,” *American Association for the Advancement of Science. Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

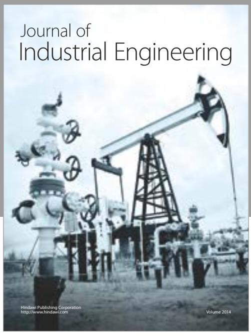
[31] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

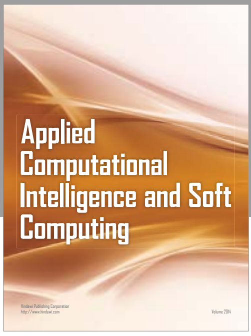
[32] D. R. Cox and P. A. Lewis, *The Statistical Analysis of Series of Events*, John Wiley & Sons, Hoboken, NJ, USA, 1966.

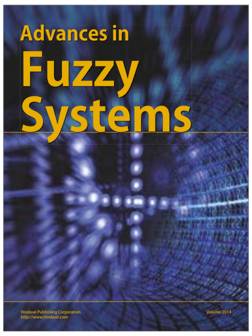

Advances in
Multimedia

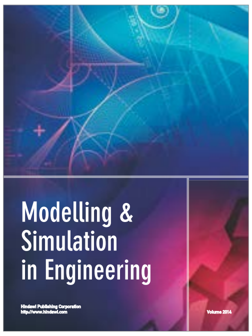
The Scientific
World Journal

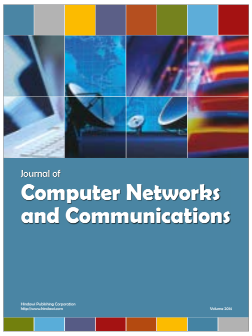

International Journal of
Distributed
Sensor Networks


Journal of
Industrial Engineering

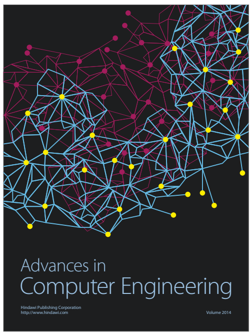

Applied
Computational
Intelligence and Soft
Computing


Advances in
Fuzzy
Systems

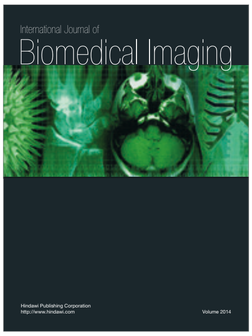

Modelling &
Simulation
in Engineering

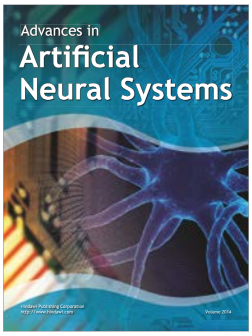

Journal of
Computer Networks
and Communications

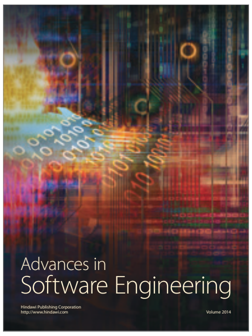
Advances in
Artificial
Intelligence


Advances in
Computer Engineering

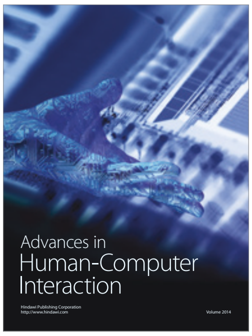
International Journal of
Computer Games
Technology

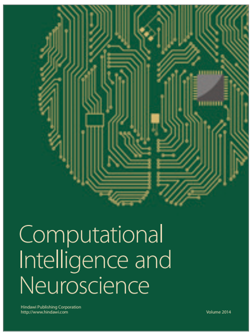

International Journal of
Biomedical Imaging


Advances in
Artificial
Neural Systems


Advances in
Software Engineering

Journal of
Robotics


Advances in
Human-Computer
Interaction


Computational
Intelligence and
Neuroscience

International Journal of
Reconfigurable
Computing

Journal of
Electrical and Computer
Engineering