Doctor of Philosophy

Predicting Human Mobility on Large-scale Mobility Traces

The Graduate School of the University of Ulsan

Department of Electrical and Computer Engineering

DAO THI NGA

Predicting Human Mobility on Large-scale Mobility Traces

Supervisor: Prof. Seok Hoon Yoon

A Dissertation

Submitted to
the Graduate School of the University of Ulsan
In partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Dao Thi Nga

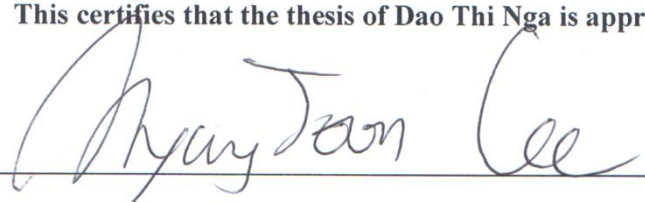
Department of Electrical and Computer Engineering

Ulsan, Korea

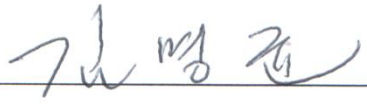
June 2019

Predicting Human Mobility on Large-scale Mobility Traces

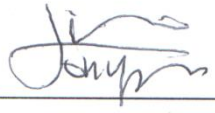
This certifies that the thesis of Dao Thi Nga is approved by:



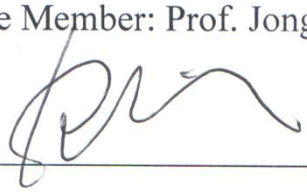
Committee Chair: Prof. Myung Joon Lee



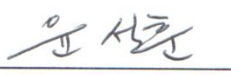
Committee Member: Prof. Myung Kyun Kim



Committee Member: Prof. Jong Myon Kim



Committee Member: Prof. Jang Young Kim



Committee Member: Prof. Seok Hoon Yoon, *Adviser*

Department of Electrical and Computer Engineering

Ulsan, Korea

June 2019

Dedicated to

My beloved parents, sisters, and brother, …

Predicting Human Mobility on Large-scale Mobility Traces

by

Thi-Nga Dao

Submitted to the Department of Electrical and Computer Engineering
on June, 2019, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

Human mobility prediction has attracted a lot of attention because it plays the key point for the success in a variety of applications ranging from location-based recommendation systems to epidemiology. Therefore, the purpose of this thesis is to answer two main questions in human mobility prediction: where a person is most likely to visit and whom a person is most likely to meet. Accordingly, two prediction models which estimate the most probable future locations and encounters are proposed.

In the first model, we aim at estimating the next location of a person-of-interest even when the recent information about the position of that person is unknown. Motivated by the fact that the behavior of an individual is greatly related to other people, a two-phase framework is proposed, which first finds persons who have highly correlated movements with a person of interest, then leverages the position information of selected persons to predict the person-of-interest's location. For the first phase, we propose two methods: community interaction similarity-based (CISB) and behavioral similarity-based (BSB). The CISB method finds persons who have similar encounters with other members in the entire community. In the BSB method, members are selected if they show similar behavioral patterns with a given person, even though there are no direct encounters or evident co-locations between them. For the second phase, a neural network is considered in order to develop the prediction model based on the selected members.

The purpose of the second model is to design a low cost, high accurate human encounter prediction framework that can be applied to large-scale networks. Taking inspiration from the advantages of the distributed system (e.g., low cost and ease of scaling up) and the temporal dependency of human mobility, we propose the distributed human encounter prediction (DHEP) model, which uses the mobility history of only the person of interest to estimate future encounters of that person. The DHEP model based on a recurrent neural

network is constructed, in which recent encounter information is captured and used to make future contact predictions. In addition, for devices with constrained computation capability, we design a feed-forward neural network-based DHEP model which contains a much smaller number of model parameters. Also, an embedding model that learns the low-dimensional representation of a person's location is proposed in order to accelerate the training of the prediction model.

Extensive experiments have been conducted to evaluate the performance of the proposed prediction models with a variety of parameters on different large mobility traces (e.g., MIT, Dartmouth, and UB datasets). The evaluation results show that the designed frameworks achieve higher performance in terms of predictive accuracy than existing studies. Specifically, the human location prediction model under the BSB method outperforms other selection methods on considered datasets. Moreover, the decentralized encounter prediction model with low overhead and high accuracy can protect data privacy and can be applied to large-scale networks.

Thesis Supervisor: Yoon Seok Hoon
Title: Associate Professor

Acknowledgments

First, I would like to express my deep gratitude to my advisor, Prof. Yoon Seokhoon of the department of Computer Engineering, University of Ulsan, for his dedicated guidance and continuous support during my study. Thanks to his critical comments and suggestions on my study, I have learned how to conduct research in an efficient way, gained a lot of valuable knowledge, and also improved my logical thinking. Without his support and guidance, I could not able to fulfill my study in University of Ulsan.

I also would like to thank the committee members for taking time to review my thesis and to give insightful comments in order to further enhance the quality of my thesis. I would like to say thank to my fellow labmates in the Advanced Mobile Networks and Intelligence Systems Laboratory for valuable discussion and fun memories we had together during the past 5 years. I am very grateful to lecturers and colleagues in Le Quy Don technical university in Vietnam for always supporting and encouraging me to conduct research. I also want to thank my special friends for sharing memorable moments and helping me through difficult time during my stay in Ulsan, South Korea.

Last but not least, I am always grateful for the endless love and support from my parents, sisters, and brother throughout the period of my study, my life in general. Since my parents have sacrificed their own life with the hope of better future for the children, I think the thesis book is a very special present that I would like to intend for them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Predicting Human Mobility in Large-scale Networks

The ability of understanding and predicting human mobility plays an important role in a variety of areas, such as location-based recommendation systems, urban planning, traffic forecasting, contagious diseases control, geographic profiling, and among others. For example, in location-based recommendation systems, packets, which contain information of services (e.g., restaurant and ATM) close to the likely visiting location of a person, should be advertised to the person-of-interest in order to enhance user experience. In geographic profiling, for the purpose of ensuring security, law enforcement agencies need to analyze movements of suspects and determine their most likely position. In addition, the accurate prediction of human movement can help to control the outbreak of diseases (e.g., flu and malaria) which are rapidly spread by people's contacts. Specifically, people who tend to meet many other people should be vaccinated first in order to reduce the possibility of the disease spread.

Future encounter estimation can also be used to determine influencers who usually encounter with many other people, allowing those influencers to then be selected to forward packets that contain an advertisement in opportunistic networks. In addition, being able to estimate future contacts allows us to calculate the expected inter-meeting time between people, which is necessary for data forwarding algorithms with constrained deadline in opportunistic networks. For example, a packet routing application may require data to be transmitted to the destination within a given delay bound. Since accurate predicting human movement can be beneficial in a variety of applications, in this thesis we aim at constructing and evaluating human mobility prediction models in large-scale networks.

Due to the daily schedules, members in the community usually expose geographical mobility in a specific area. For example, Figure 1-1 presents geographical movements of 4 people from time t_0 (e.g., early morning) to t_1 (e.g., afternoon). Usually, people move from a place to the next one and then stay there during a period of time for their purpose (e.g, studying and talking with friends). According to previous studies [3–7], even though human mobility have a high degree of freedom and variation, they also show structural patterns due to geographic constraints and social relationship. That means it is possible to discover and predict the most likely trajectories of people in the future.

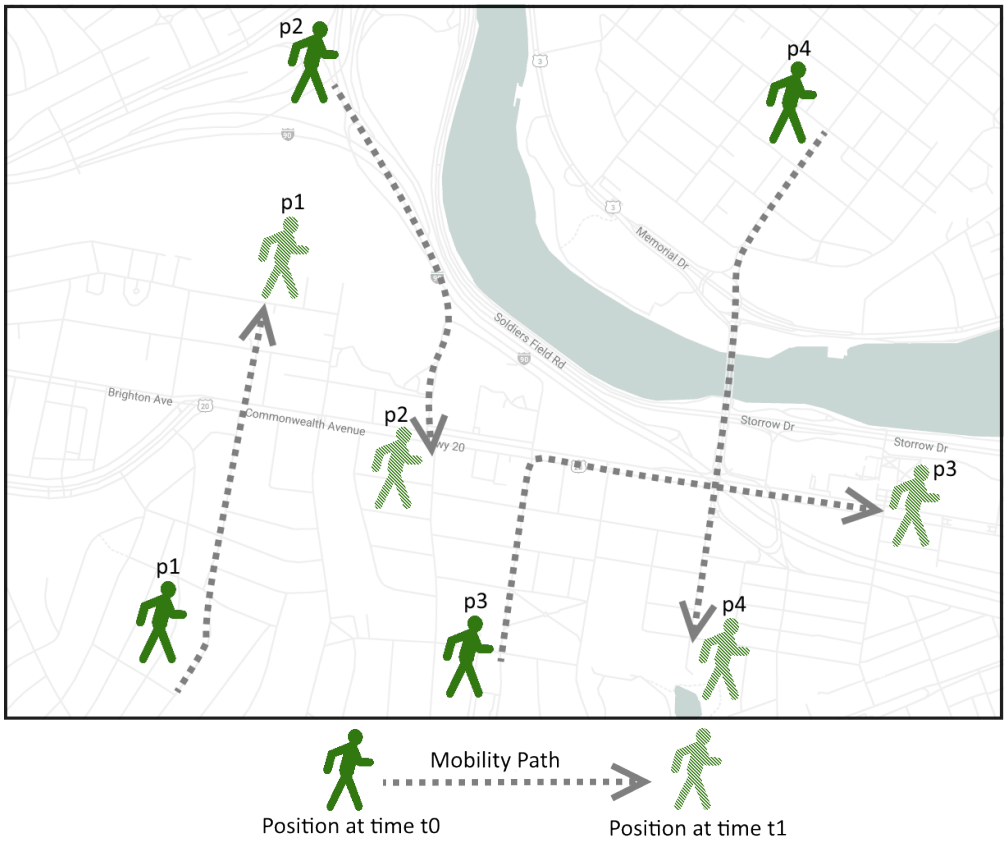


Figure 1-1: Mobility trajectories in a network

However, there are some main challenges in constructing and evaluating human mobility prediction. First, in order to train and examine the proposed prediction models, it is required to have a large human movement traces of many different people during a long period of time (e.g., several months or years). However, mobility data of some people is not available all the time in cases when people are not in the range of communication networks or mobile devices are forced to turn off the global positioning system function to reduce the power consumption. Therefore, when designing the human mobility prediction models, we need to

pay attention to the problem of non-continuous movement data of collected traces. Moreover, the proposed prediction model should be able to address the lack of samples in some real mobility traces.

Secondly, due to the increase in mobile phone users, some prediction models, which require the mobility information of all people in the network, suffer from the scale problem and encounter the privacy concern. Therefore, there is a high demand for a low-cost mobility prediction model which can be used in large-scale networks without worrying about the privacy leakage problem.

Thirdly, neural networks have been widely used to construct different prediction and classification models. In neural network-based models, parameters (e.g., weights and biases) can be adjusted using a gradient descent method to accurately generalize an output value given specific input features. However, if there is an increase in input features, we may face the problem of a long training process. In order to cope with this problem, a common and well-known technique called dimension reduction should be used. Specifically, input feature selection can be applied to find useful input features for prediction models or a low dimensional representation of input features can be used to lower the size of the input features. In the next chapters, we will describe how to accelerate the training process of neural network-based human mobility prediction models.

Moreover, since a person may associate with different base stations during a short period of time, inferring a person's location and encounter from mobility traces should be considered carefully. Usually, the base station to which a person connects most in a time period is used to present the representative location of that person during the considered period. Also, from the raw datasets, it is necessary to extract people whose mobility are most active and an experiment period in which we observe the most mobility samples.

Lastly, existing works usually focused on predicting human movement in the near future (e.g., next several hours) but not the far future (e.g., next several days). Note that some applications (e.g., data forwarding, disease spread) may need information about future human mobility for a long time period, it is necessary to predict human mobility in the far future. Moreover, when predicting human movement some existing works required extra information (e.g., call or message history) of people, which is not easy to obtain due to the increasing privacy concern. In summary, we briefly analyze several challenges that need to be considered when designing a human mobility prediction model. And, we will present how to address the above-mentioned challenges in the next chapters of this thesis.

In human mobility prediction, there are two main questions: where a person is likely to visit and whom the person is likely to meet in the future. Therefore, the purpose of this thesis is to design two models which predict future location and encounter of a person-of-interest. Specifically, we first extract movement features from real mobility datasets and then use these extracted features to construct two movement estimation models. The detail approaches are presented in the next section.

1.2 Proposed Prediction Models

1.2.1 Two-phase Human Location Prediction Framework

In the following subsection, we first introduce the considered problem and then the two-phase location prediction framework is briefly summarized. If someone's recent location is known using a global positioning system (GPS) or other localization techniques (e.g., Wi-Fi, cellular-based systems), it is relatively simple to estimate that person's current or next locations. However, there are cases where someone's recent location sequences are not available. For instance, the GPS function is required to turn off due to lack of battery power in a smartphone even though people want to share their positions. In addition, there are some cases where data is sparse [8, 9](e.g., call detail records, health record of individuals, and credit card spending data). Moreover, in some cases, it is necessary to estimate locations even when someone may not want to share that information (e.g., geographic profiling of criminals). Therefore, we consider a new problem which is estimating someone's current or next location, particularly when recent location sequence information is not available.

Specifically, the mobility information of other Persons with Correlated Movements (PCM) with a given person is used to estimate the person of interest's current or future location, based on the fact that the behavior of one individual is largely related to other members. In other words, the current or next location of a given person is estimated using the known position information of other specific people. Although there were several studies on the prediction of human location using social relationships [6, 10, 11], those works focused on finding direct social ties between people (e.g., friendships) using encounters or calling traces between them. However our work, in addition to friendships, considers behavioral similarity between people, which may not be reflected by direct encounters or co-location events between them.

We propose a two-phase human location prediction model. More specifically, in the

first phase, PCMs with a given person are selected. For this phase, two novel PCM selection methods are proposed: community interaction similarity-based (CISB) and behavioral similarity-based (BSB). In the CISB method, interactions with other community members are considered as well as direct encounters between certain individuals, i.e., the strength of the social relationship between people is measured by not only their direct interactions but also their interactions with other members of society. The motivation is that if two people have a close social tie, then they may have similar patterns of meeting other members of the community. In contrast to the CISB method, which attempts to find PCMs with stronger social ties, the objective of the BSB method is to find PCMs with similar behavioral patterns. As a result, the selected PCMs under the BSB method may not have a strong social relationship with the person of interest. In the second phase, motivated by the fact that the movement of an individual is highly correlated with other people, the current or next location of a person is estimated using the position information of selected PCMs. A PCM-based location prediction (PLP) model is developed based on a neural network (NN) [12]. Chapter 3 will present the two-phase location prediction framework in more details.

1.2.2 Low Cost Decentralized Human Encounter Prediction Model

In this subsection, we summarize limitations of existing works and then briefly introduce our prediction model. Several existing studies, including [13, 14], addressed the problem of predicting people's contacts in the 20-minute [14] or one-hour [13] time slot immediately following. However, they did not consider encounters in the more distant future, rendering them insufficient to address the problem currently under consideration. Moreover, with the rapidly increasing number of mobile device users, there is a need for a scalable encounter prediction model that can be used in large-scale networks. However, in existing studies [15, 16], the mobility traces of all people in the network were collected in order to build the encounter prediction model, which could have caused the scale problem and privacy leakage.

To this end, we propose a distributed human encounter prediction (DHEP) model which requires the mobility traces of only one person to predict the future contacts of that person. Therefore, our model preserves data privacy, since mobility traces are not exchanged throughout the network. In addition, the DHEP model does not suffer from the scale problem and is particularly suitable for OppNets with intermittent connectivity.

Over the past few decades, recurrent neural networks (RNNs) [17] have emerged as a promising model for handling sequential data in various tasks, including natural language

processing, image captioning, and handwriting recognition. In contrast to traditional feed-forward neural networks, an RNN includes states that can capture historical information from an arbitrarily long context window. Therefore, in order to learn temporal encounter dependency, an RNN-based DHEP model, named DHEP/RNN, is constructed. In addition, we propose a feed-forward neural network-based DHEP model (DHEP/FFNN) in which human future contacts are forecasted using only the current encounter data. Compared with the DHEP/RNN model, the DHEP/FFNN model is more suitable for devices with limited computation capability due to the fact that it involves fewer training parameters.

In this study, for comparing with DHEP in terms of performance and applicability, we also develop a centralized human encounter prediction (CHEP) model in which a central entity aggregates the mobility traces of all people in order to train the prediction model. Then, the trained model parameters are distributed to each individual who will make a contact prediction using the received model parameters. Through the use of the global information about human mobility, the CHEP model is expected to exhibit improved performance over the DHEP model. However, the CHEP model suffers from the scale problem and the privacy leakage issue as well as high network overhead, due to the data exchange throughout the network.

The DHEP model leverages the historical encounter information in order to estimate whom a person will meet in the future. The encounter information consists of the encountered people, meeting time, and rendezvous places, which are represented by points of interest (POIs). If there are a large number of POIs and one-hot encoding is used to represent the locations of people, the number of parameters becomes substantial. In order to avoid this problem, we propose an embedding model that outputs the distributed vectors which reflect the people's POIs. Taking inspiration from natural language processing, the embedding model is trained such that nearby physical positions are represented by vectors that are close to each other in the vector space. The detail description of the encounter prediction model will be presented in Chapter 4.

1.3 Evaluation Methods

The purpose of this thesis is to propose human mobility prediction models which can be used in a variety of real applications such as location-based recommendation systems and data forwarding in opportunistic networks. Therefore, we evaluate the proposed prediction

models using the real large mobility traces of communication networks. Specifically, human movement data is extracted from three different datasets including the cellular network traces (i.e., the MIT Reality Mining dataset [18]), the Wi-Fi logs from the Dartmouth college [19], and the Wi-Fi traces from University at Buffalo (UB) [20]. The description of these datasets is presented in next chapters.

Each dataset is divided into training, validation, and test sets. The training data is used to adjust model parameters (i.e., weight and bias values) while we select the hyper-parameters (e.g., the number of hidden layers and the number of hidden units) which produce highest performance on the validation set. Meanwhile, the evaluation results are collected by using only the test set.

In this dissertation, since we construct classification-based prediction models, different performance metrics can be used to evaluate the prediction models as follows.

- The most important metric is *estimation accuracy* which is defined as the percentage of correct prediction samples. More generally, we consider top-k accuracy, in which the model outputs a list of k labels (e.g., locations) with the highest probability. If the target label (e.g., future location) belongs to the list of k element(s), the prediction is considered accurate.

- Since some datasets may expose bias in favor of negative samples, we record the confusion matrix on the test set. Then, other metrics, which are extracted from the confusion matrix, could be also considered including *sensitivity, precision, F1 score*.

- In order to observe the performance of prediction models under different decision threshold values, *receiver operating characteristic* (ROC) and *area under curve* (AUC) are collected to evaluate the prediction models.

- Because we aim at designing prediction models for large-scale networks, *time and space complexity* of prediction models should be also computed. Specifically, the number of training parameters of the prediction models are estimated and compared with other methods.

1.4 Dissertation Organization

The rest of this thesis is organized as follows. Chapter 2 summarizes literature related to human mobility prediction including human mobility analysis, mobility models, mobility

estimation schemes, and social circle-based mining. Then, two models that predict next visiting locations and future contacts of the person-of-interest are described and evaluated in Chapters 3 and 4, respectively. Finally, in Chapter 5 the concluding remarks are drawn to summarize our work and discuss future directions of our studies.

Chapter 2

Related Works of Mobility Prediction

2.1 Human Mobility Analysis

A number of studies aimed to reveal human movement characteristics [3–5]. For example, Karamshuk *et al.* [5] classified the properties of human mobility into three groups consisting of spatial, temporal, and connectivity. With regard to the spatial characteristic, they focused on geographic movement, i.e., how far a user moves and where a user goes. Flight was defined as an Euclidean distance between two consecutive spots visited by the same individual. Temporal features were also considered, e.g., pause-time indicates the time period a user stays at a specific location. Meanwhile, the connectivity property reflects the contact or encounter between two users, e.g., inter-contact time was defined as the elapsed time between two adjacent contacts for a pair of users.

González *et al.* [4] analyzed the 6-month trajectory of 100,000 anonymized mobile phone users and found that human mobility follows simple reproducible patterns. Specifically, the distribution of flight can be approximated by a truncated power-law. Moreover, the authors measured the radius of gyration, which is the characteristic distance travelled by a person during a time period, and concluded that the radius of gyration also follows a truncated power-law distribution.

In addition, there were a few studies that considered the effect of social relationship on the user movement [6], [7]. Cho *et al.* [6] found that human movement is a combination of periodic mobility, which is geographically limited, and movement which is related to social relationship. Moreover, people are likely to visit a distant place where a friend stays nearby, based on the datasets of online location-based social networks and cell phone location trace. Meanwhile, the short travel is less affected by social ties and more related to periodic

movement. Crandall *et al.* investigated the effect of social ties on the co-occurrence events in time and space. More specifically, the authors tried to answer the following question: what is the probability that two people have a social tie (i.e., they know each other), given that they encounter k times during a period of time. By analyzing a dataset of 38 million geo-tagged photos from Flickr, the authors found that even a very small number of co-occurrences can result in a high likelihood of a social tie.

Those studies differ from ours in that they mainly focus on capturing human mobility characteristics. Whereas we aim to design human movement prediction models to estimate a person's future location and encounter. Note that embedding the properties of human mobility will be helpful for accurately predicting human movement. Therefore, in this work, a mobility prediction model is proposed considering human movement characteristics such as encounter frequency and social correlation.

2.2 Human Mobility Model

2.2.1 Synthetic Mobility Models

This subsection presents several simulation-based human mobility models which can be used to reproduce movement patterns of individuals. For example, a classic mobility model (called random walk) is based on a random selection of direction and speed for each individual [21–25]. A person starts from a current location and moves to a new one by choosing speed and direction randomly from the given intervals. Right after reaching each position, the person will calculate the new direction and speed. When an individual reaches an area boundary, he/she will bounce off the border with an opposite angle with the incoming direction.

The probabilistic random walk mobility model considers the probability of moving to next positions of a person. Specifically, assume that there are n_l points-of-interests (POIs). The mobility model computes a square matrix which has n_l rows and n_l columns. The element at i^{th} row and j^{th} column represents the probability that a person, who currently stays at location i, will move to position j in the future. Another extension of the random walk mobility model is called the random waypoint model, which assumes that people stay at a position for a pause time before moving to the next spot. This synthetic model is usually used when conducting movement simulations in ad-hoc networks. Unlike the random walk model, the random direction model assumes that a person will select new direction between

0^o and 180^o degree whenever he/she arrives at the area border.

Random walk is a memoryless model since no information about previous movement is stored for a future mobility decision. Therefore, there can be sudden changes in direction and speed and some characteristics of realistic mobility may not be captured (e.g., temporal and spatial dependency), thus rendering the random walk model unrealistic. However, due to its simplicity of implementation and analysis, the random walk model has been widely used.

The Gauss-Markov mobility model, which is inspired by a Gaussian distribution, removes sudden changes in direction and speed in human mobility. Particularly, this model takes into account the historical mobility of people (e.g., speed and direction) in order to determine future movement. According to [21], the mean speed and direction are random variables following the Gaussian distribution.

Some mobility models [26, 27] tried to reflect human mobility features. For example, the city section model [26] provides realistic movements for people who are located in a certain city area. The area includes a number of streets and there are intersections between them. Each street is associated with a pre-determined speed limit. A person starts at a street intersection and determines the next visiting intersection. The shortest path from the starting spot to the destination is calculated and the person travels on the determined shortest path. After arriving at the spot, this person will stay there for a pause time. Because of considering the real map of a specific city, the city selection model is widely applied in vehicular ad-hoc networks.

If two individuals have a social interactions (e.g., friends, family, colleagues), their movements are usually correlated [28–30]. For example, two friends may have an appointment of visiting a place in which they are both interested. In order to better understand the relationship between movement and social tie, [7,31,32] studied the impact of social interactions on human mobility. For example, two people tend to have stronger relationship when they concurrently visited more geographic locations during a given time period. The reason for this finding is that friends are more likely to spend time together in the same place or to live close to each other. Therefore, [33] proposed a model which predicts location of a person using the information of his/her near contacts.

There are several human mobility models [34–36] which consider only the effects of social relationship between people on human movements. However, these models are not able to capture realistic human movements because next destinations are determined using only

social ties instead of considering human movement characteristics (e.g., flights, inter-contact time, and pause-time) which were explored in existing studies [4, 37–39].

Some mobility models motivated by the generated social networks were studied [40, 41]. For example, in [41] people are initially located randomly according to population density in the area. Then, each person decides to visit a social contact with probability p_v, or follows the movement with a Levy-like flight with probability $1 - p_v$. The destination is one of the physical locations in order to prevent people from moving to unphysical spots. Meanwhile, the study in [40] assumed that each step a person goes back to previously visited location with probability $1 - p$ or discovers a new place with probability p. The new location is selected from the list of his/her contact's previously visited locations.

Duong *et al.* [42] proposed a social relationship-aware mobility model (SRMM) which not only captures main characteristics of human movements (e.g., flight, inter-contact time, and pause-time follow the truncated power-law distribution) but also considers the impact of social relationship on human mobility. Specifically, a clustering algorithm is proposed to partition people into social groups where each group represents a certain community in the real world (e.g., family, class, baseball club). Spots (such as classroom, laboratory, clothing stores, and cafeteria) are clustered into a number of places (e.g., a university, mall). Then, the SRMM model allows people in a social group to visit places with the same probability and to decide some frequently visiting spots in a place for movement.

2.2.2 Real Mobility Traces

Even though synthetic mobility models tried to capture main features of human movements, the realistic movement traces should be collected and used to evaluate mobility prediction models and data routing algorithms. By using real large-scale mobility data, evaluation results have become more reliable and convincing. Therefore, in this subsection, a number of real large-scale movement data, which is extracted from traces of communication networks, will be briefly introduced and analyzed.

Thanks to the development of the personal devices' communication technologies (e.g., cellular network, Wi-Fi, Bluetooth, and LTE direct) during the last few decades, human trajectories can be inferred using traces of mobile networks. For example, call detail records [1] from mobile phones provide the time of incoming and outgoing call/message as well as cell towers with which a user was associated. Since people are likely to carry mobile devices most of the time, the geographic coordinates of the cell tower can be used as the representative

location of the user. Figure 2-1 shows the recorded movements of 2 mobile device users. There are a number of available CDR datasets as listed in Table 2.1. For example, the Nodobo dataset [43] provided a mobile phone traces of 27 high school students over a 5-month period in 2011. The data contains 13,035 call records, 83,542 message records, and 5.2 million proximity samples. Moreover, authors in [44] collected mobile phone activity data over 2 months (from November 1st 2013 to January 1st, 2014) in Milan and Trentino, Italy. The dataset provides information of the message, call, Internet activities of people in Milan and Trentino. Note that the population in Milan and Trentino is 1.3 and 0.5 million, respectively. The area of Milan is divided into a grid of the 1,000 cells of 235x235 m^2 and an cell ID is used to approximate a user's location.

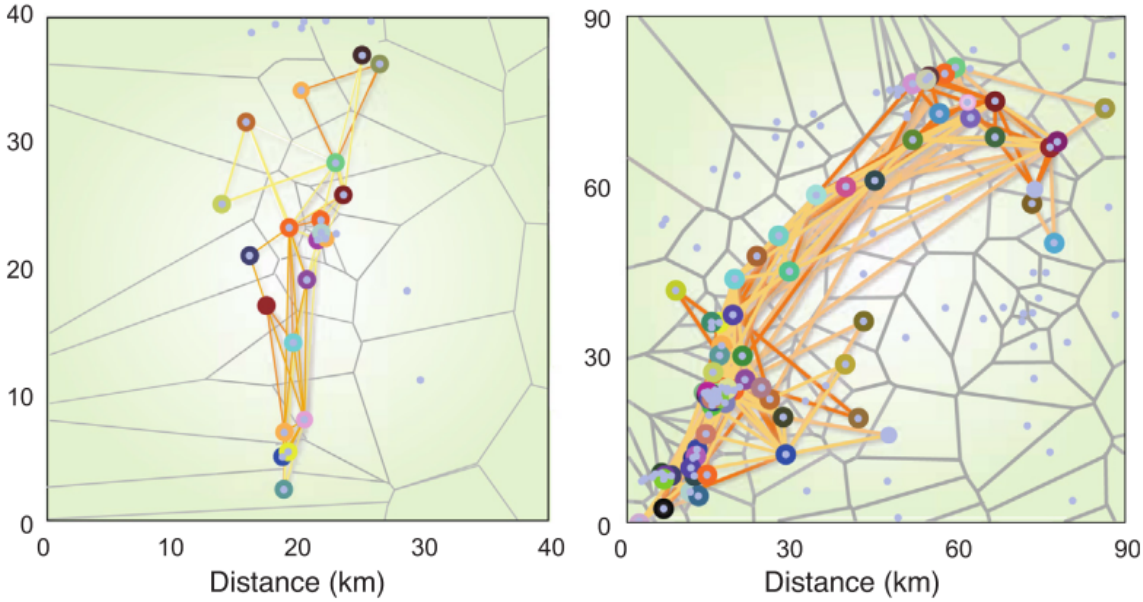


Figure 2-1: Trajectories of two anonymized mobile phone users travelling in the vicinity of 22 and 76 different cell phone towers during a 3-month period. Each dot represents a mobile phone tower, and each time a user makes a call, the closest tower that routes the call is recorded. The gray lines correspond to the Voronoi lattice, approximating each tower area. The colored lines represent the recorded movement of the user. *Source*: Figure from [1].

Since the extract position of a user may be 100 meter away from the cell tower location, inferring human mobility from CDR datasets has low spatial resolution. In contrast, global positioning system (GPS) data can provide more fine-grained positions of people. For instance, [2] studied the human mobility patterns extracted from a GPS data which stores information about the trips of 46,000 vehicles in May 2011 in Italy. Figure 2-2 visualizes human trajectories started from 2 cities in Italy. Moreover, another GPS dataset [45] collected 2.5×10^6 different trajectories of 35,000 vehicles moving in a circular area with the

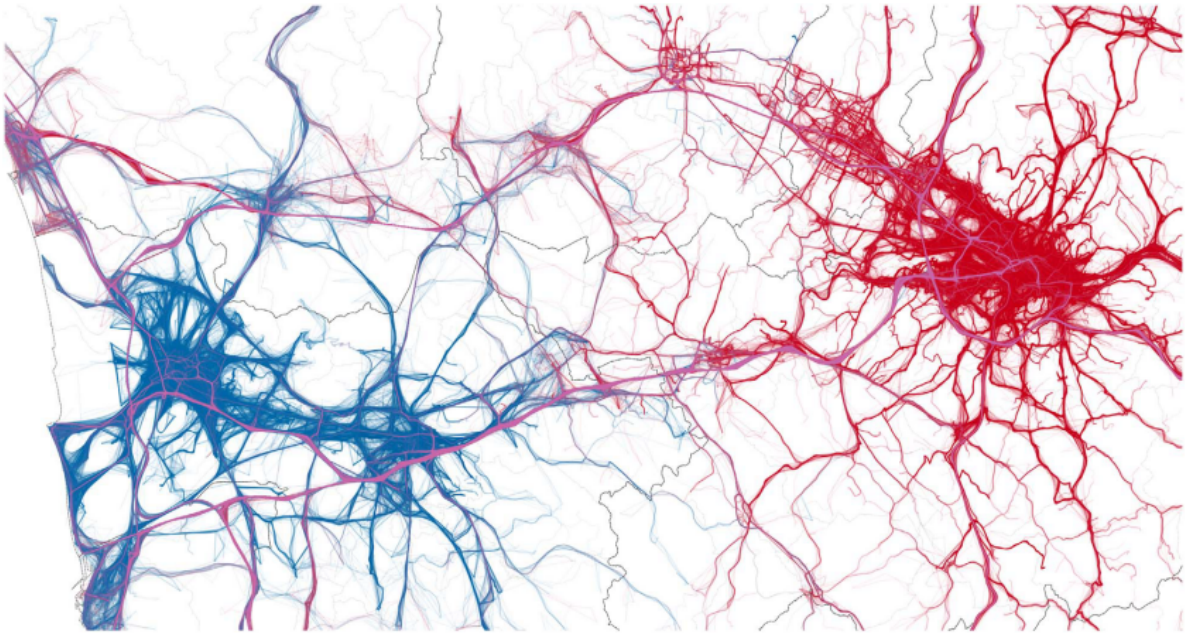


Figure 2-2: A visualization of the complexity of the explored mobility area. A fragment of the GPS trajectories display trips originating in the metropolitan area of Pisa city (in blue) and Florence city (in red). *Source*: Figure from [2].

radius of around 30 km.

Now, several datasets of wireless local area networks (WLANs) will be presented. The WLANs traces [46] collected the Wi-Fi logs of 275 freshmen students for 11 weeks from Sep 22 to Dec, 8, 2002. Students are assumed to have a registered wireless card MAC address and there is one-to-one mapping between users and wireless cards. Each record contains the following information: detected access point signal strength, AP MAC address, and current AP association. Another Wi-Fi dataset in [47] recorded human mobility data from a lot of APs located mostly in cafes, restaurants, bars, and libraries in Canada. The experiment collection lasts 1095 days with the participation of 140 people.

In order to generate innovations around smart phone-based research, the mobile data challenge (MDC) dataset of nearly 200 volunteers was collected by Nokia started in October 2009 in the Lake Geneva region, Switzerland. The MDC data provides a variety of data types (GPS, WLAN, Bluetooth, and phone call/sms). This kind of mobility data can be used to evaluate location/encounter prediction models and to understand the effect of the demographic attribute on human mobility. The UIM traces [48] provided both the location and contact information of 28 people on University of Illinois campus. Each mobile device is able to detect nearby devices by using Bluetooth and Wi-Fi communication technologies.

In order to have the better view of real mobility datasets, Table 2.1 summarizes main features of real large-scale mobility traces including data type, the number of people, and the experiment collection period.

Table 2.1: Summary of real large-scale mobility traces

Real movement traces	**Data type**	**Number of participants**	**Duration**
MIT reality mining dataset [18]	CDR	106	75 days
Nodobo dataset [43]	CDR	27	5 months
Mobile phone activity dataset [44]	CDR	1.8 million	2 months
GPS data [2]	GPS	46,000	1 month
GPS data [45]	GPS	35,000	1 month
WLAN data [46]	Wi-Fi	275	11 weeks
Dartmouth dataset [19]	Wi-Fi	17,414	4 months
WLAN data [47]	Wi-Fi	140	1095 days
MDC dataset [49]	Bluetooth	185	1 year
UIM trace [48]	Bluetooth/Wi-Fi	28	19 days

2.3 Human Location Estimation

Several studies tried to predict where a person stays given the prior information on historical locations of that person [8, 50–58]. Most of those studies are based on the Markov model. For instance, in order to predict the person's position in upcoming time slots, Pang *et al.* [53] proposed a modified Markov model considering spatio-temporal information (i.e., sojourn time and location transition preference). The authors claimed that the modified Markov model achieves higher prediction accuracy than the original Markov model. Ghosh *et al.* [50, 51] found that human mobility exhibits the partially orbital movement pattern, i.e., people routinely stay at a few certain spot(s) for a considerable amount of time. This partially deterministic movement pattern can be used for human location prediction without the need for constant tracking.

Al-Molegi *et al.* [59] constructed a Markov Chain-based model to predict the most likely future regions-of-interest (ROIs) of a person given the current ROI and time information of that person. Specifically, a state in Markov Chain contains the person's ROI and time of day while state transition corresponds to the movement from one ROI to the next. The location, to which the highest transition probability is produced, is predicted as the next ROI of that person.

Alhasoun *et al.* [8] constructed a prediction model based on dynamic Bayesian networks with the assumption of knowing the last visited location of the person of interest and histor-

ical position data of 'strangers'. Strangers were defined as members who do not necessarily have a social link to the person of interest. Authors proposed three methods to determine strangers: temporal closeness, spatial closeness, and spatiotemporal closeness. In case of the temporal closeness approach, the members who have the most similar pattern of communication (e.g., call, sms, and data) to the person of interest were chosen. Meanwhile, the spatial closeness method selected the person with the most similar distribution of visited locations to the person of interest. The spatiotemporal closeness considered the chi squared test value to compute the closeness between two people p and q. Specifically, the contingency table was first constructed where the element at row i and column j represents the number of times persons p and q concurrently stay at location i and j, respectively. Then, the chi squared test value was used to measure the degree of association between two persons on the contingency table.

In addition, Zeng *et al.* [58] first determined the missing data of human trajectories by using the Gibbs sampling algorithm. Then, a high-order Markov chain model was constructed to predict the most likely location of the person of interest. Meanwhile, Noulas *et al.* [57] considered two location prediction models based on linear regression and M5 model trees to address the problem of predicting the person's next location. Mobility features such as historical visits and temporal information were fed into the prediction model. The authors concluded that combining different mobility features achieved noticeably higher performance than using a single feature approach. Unlike models in [8, 53–58], we design a prediction model which does not require a prior location history of the person of interest.

In addition to historical visit information, a number of factors can be used to predict human mobility, such as social friendships, location preferences, and temporal information. Among these factors, the strong relationship between person movement and friends was revealed in some studies [6, 33, 60]. Consequently, a number of mobility prediction models were designed with the support of friendship information [6, 10, 11, 56, 61, 62].

For instance, Cho *et al.* [6] decomposed human mobility into two parts: periodic and socially correlated movements. The authors demonstrated that short-distance travel is usually affected by periodic mobility, whereas friendship tends to influence long-range movement. They first developed a human mobility prediction model assuming that people's periodic travels follow a mixed Gaussian distribution of home and workplace. Then, with consideration for social friendships, the probability of being in a location is estimated as a function of the time period during which a friend stays in that location and the distance from the

person to his/her friend. Finally, a mobility probability distribution combining periodic and socially correlated movement is formulated using Bayes' theorem.

Even though certain studies [6, 10] did not require a location history for people, their models only work with datasets of geographic locations, e.g., GPS traces. Therefore, to address the limitation of [6, 10], our work attempts to design a more flexible and applicable model that does not require a dataset of physical locations. Moreover, our model considers behavioral similarity between people as well as friendship. Note that behavioral similarity may not be revealed based on direct interactions or encounters between people.

In regard to datasets of symbolic locations, among the models considering social correlation features to predict person movements, a few approaches worked with datasets of non-geographic locations [11, 56]. In [56], a location prediction model was considered based on two factors: periodic movements and social relationships. Specifically, a Markov-based model was constructed to capture periodic movements while colocation frequency was used to measure the closeness between people. In order to reflect the impacts of both factors on human mobility, the location prediction model was built where a different weight was assigned to each factor. Meanwhile, Zhang *et al.* [11] proposed an algorithm called NextCell to predict the future locations of people. A boosting technique was used to combine two predictors that are based on periodic behaviors and social interplay. The periodic behavior predictor considers a probability distribution over locations. Meanwhile, in the social interplay predictor, the probability that two people co-locate at a given time was estimated as a function of phone call features.

Although the above studies [11,56] consider the datasets of symbolic coordinations, those models do not take into account the impact of time features on human mobility which are believed to be an important factor for human mobility prediction [6, 53, 57]. Meanwhile, our work aims at considering both temporal and spatial information of social friends to design the movement prediction model. Moreover, for predicting the current or future location of a person, the proposed framework also takes into account other human mobility characteristics, e.g., encounter frequency, community interactions, and behavioral similarity.

2.4 Human Encounter Estimation

Note that since the purpose of location prediction is to estimate where a person will be in the future, encounter information between two people can be inferred if their future trajec-

tories are provided. However, in order to predict encounters between a person and other individuals, location traces of all people in the network are required for constructing location prediction models. This requirement becomes definitely costly if the network consists of a large number of people. Therefore, instead of indirectly inferring the future encounters of a person through location prediction, in this work, we propose a model which directly predicts the future contacts of that person.

The rest of this section describes recent studies related to future contact prediction. Several studies [63–66] have leveraged the temporal context (e.g., the day of a week and time of a day) to estimate whom a person of interest is likely to meet in the future. In [63], the Naive Bayes (NB) predictor with a modification was used to construct a link prediction model. In the traditional NB model, input features are assumed to be independent given the output value, e.g., Jyotish's model [64] applied the Bayesian classifier to determine whom a person is likely to meet given the type of day (weekday or weekend) and the time slot index. On the other hand, the modified NB, which is called Jahromi's model hereafter, relaxes the assumption of the traditional Naive Bayes by assigning a different weight to each temporal feature, according to [67]. Note that the studies in [63–66] only took the temporal context into account. This consideration may be insufficient, since, according to several studies [3–5], human mobility highly depends on a number of other features such as location preferences and past locations. Therefore, in order to address the limitations of [63–66], we consider a future encounter prediction model that embeds not only temporal information but also the historical contacts and spatial information of the person of interest.

Other studies, such as [13, 14], aimed at predicting people's encounters given historical contact information. Nguyen *et al.* [13] proposed a prediction model that considered two features to forecast the next encounters of a person at the very next time, $t+1$. One feature was the encounter at current time t on the same day, while the other was the encounter distribution of that person at time $t+1$ on previous days. These features were then fed to a boosting algorithm which adjusted the features' weights in order to minimize the error function.

While many studies only consider the direct encounter between people, Tatar *et al.* [14] stated that the direct encounter may provide a limited view about transition possibilities. Therefore, given the historical encounter information of all people in the network, their prediction model outputs contacts who are k nodes away from the person-of-interest. Through simulations, it is concluded that predicting that nodes stay at a distance of at most two hops

from the person-of-interest can produce twice the prediction performance of a direct contact prediction. Moreover, indirect encounter estimation can bring us higher transmission opportunities in applications such as data forwarding in opportunistic networks.

Unlike the studies in [13, 14], which only considered the prediction of the very next contacts of people, our work investigates the predictability of human encounters after k time slots ($k \geq 1$).

In [15, 16] human encounters were viewed in the form of a time-varying network graph. People were represented by a set of nodes, and an encounter between two persons are denoted by a bidirectional edge between a pair of corresponding nodes. The prediction of future human encounter was made based on historical graphs which included all nodes' contacts during a certain period in the past. Specifically, the study [15] extracted several features, which include the number of common neighbors, time spent with common neighbors, overlapping time between two nodes u and v. Then, these extracted features are used as the input of the Naive Bayes-based classifier which predicts the future contact between two nodes u and v.

Jahanbakhsh *et al.* [68] studied the problem of estimating the missing part of a contact graph by computing a similarity score between neighbor sets of 2 nodes. They assumed that only a subset of nodes are equipped with a sensor device and able to sense nearby contacts. The experimental results collected on real life mobility traces indicate that combining social information with time-spatial information achieves higher performance than using each of them separately. The evaluation results also show that reconstructing the missing parts of contact graphs enables researchers to expand the existing collected traces to include external people as well.

While the studies in [15,16] required the past encounters of all nodes to predict the future contacts of a person, our work only needs the historical encounter data of that person.

2.5 Social Circle-based Mining

This section aims at summarizing recent studies based on a social circle including social community detection and social circle-based applications.

A number of studies were conducted to detect social circles (i.e., social communities) using graph clustering [69, 70] where a network was divided into disjoint social circles by using clustering techniques. There were several studies on community detection based on

the contact history of members in the network, e.g., encounter frequency and duration [71] and the total number of past encounters of a person [72]. Eagle and Pentland represented the behavior of individuals from a set of primary vectors called eigenbehaviors [72]. Then, community affiliation can be inferred by computing the social behavior distances (e.g., the total number of past Bluetooth encounters) between a person and other members of a social circle.

Those community detection schemes differ from ours in that they were mainly focused on partitioning people into groups, whereas in our work, people can have overlapping PCMs. Moreover, in order to select PCMs for a given person, community interactions and behavioral similarities are considered, in addition to encounter frequency.

A number of recent studies have shown that the social circle information can play an important role in various areas because human activities are certainly shaped by the social relationship, rather than are randomly and independently determined.

For example, there were a number of social circle-based recommendation systems [73–75]. The problem of suggesting a sequence of points of interests (POIs) for travelers was considered in [74]. First, socially close friends who have similar travel records with a person of interest were discovered, and then the system recommends a list of POIs where the close friends visited in the past. While social circle-based recommendation systems mainly focus on estimating rating and preference of the person of interest by using top-ranked items from his/her close friends, our work proposes the human location prediction framework with the support of the recent spatial data of PCMs.

Another social circle-based application is social messages geolocation which infers the location associated with a specific post from a person of interest in the social network [76–78]. In [76], a modified majority voting scheme was used which returns the most popular location among positions of social friends of the person of interest. Unlike social posts geolocation which does not take into account human mobility prediction, our work first selects persons with correlation movements (PCMs) of the person of interest based on the social relationship and interaction between people in the social circle. Then, the selected PCMs' position information is fed into the prediction model to estimate where the person of interest stays at the current or future time.

Chapter 3

Two-phase Human Location Prediction Framework

3.1 Preliminaries

3.1.1 Dataset

In this work, we consider two different datasets. The first one is called the MIT Reality Mining dataset [18], which was collected during a period of nine months with the attendance of 106 subjects, including students and faculty members of MIT. Since subjects in the MIT dataset are involved in the same university, the social relationships between people exist with a high probability.

The MIT dataset provides cell tower logs including the tower transition events and a set of base stations seen by the participants. In cellular networks, a mobile phone can be within the ranges of several cellular towers. However, the phone is only associated with the tower with the strongest signal. The events of tower transitions are recorded with cell tower ID and a timestamp. Due to the fact that small and short-range cells that cover distances of few hundred meters are more popular in metropolitan areas [79], the cell tower logs can be used to represent human locations [11, 56, 80–83]. Hence, this work uses the dataset of cellular traces in order to evaluate the movement prediction model. The subjects in the dataset participated in the experiment in different time periods, and some subjects have no data or very little data [70]. Therefore, by considering overlapping periods and available data, 43 people with sufficient mobility data were selected. $\mathbb{M}$ and m are defined as the set of chosen people and the cardinality of $\mathbb{M}$, respectively.

The second one, called Dartmouth dataset [19], provides the mobility traces extracted from logs of APs in the Dartmouth university campus. A log message including the timestamp, user ID, and AP ID was recorded when a mobile device connects or disconnects to the AP. Because of the short range of the Wi-Fi technology, human mobility can be represented as a sequence of connected APs [84–86]. For the Dartmouth dataset, a 4-month period from January 3th to April 30th, 2004 was considered since the during this period the academic campus was relatively consistent [87, 88]. Similar to the MIT dataset, persons whose mobility data was provided less than 75% over the experiment period was filtered out. Then, the Dartmouth dataset includes 162 mobile users.

3.1.2 Location Extraction

In this subsection, the way to extract locations for people is described. In this work, location information based on time slots is used. Note that a mobile device may be connected to several base stations (e.g., cell towers or access points) during a time slot. Therefore, in such cases, the way to determine the locations of people in a time slot is needed.

Let λ denote a threshold for location extraction ($0 \leq \lambda \leq 1$). By using λ, the representative base station with which the phone is associated the most is determined, since a mobile device may connect to several base stations during a time slot. Specifically, if the ratio of the time a person spends at a base station to the time slot length exceeds λ, this base station is regarded as the representative human location in the specific time slot. In this work, λ and time slot length are set to 0.5 and 30 minutes, respectively. In cases where no base station satisfies the conditions for location extraction, or where a mobile phone does not receive any signal during a time slot, the person's location in that time slot is marked as undefined.

In case of the MIT dataset, positions at which people rarely stay should be pruned. The locations are first arranged in descending order of occurrence frequency in the dataset. Then, a location set that contributes 98% of the cellular traces is selected for use in this work. Let $\mathbb{L}$ denote the set of all symbolic locations in the dataset after the pruning process. In the MIT dataset, the total number of positions remaining after the above pre-processing is 482, i.e., $|\mathbb{L}| = 482$. Meanwhile, in the Dartmouth dataset, there are total 399 APs. Table 3.1 summarizes the main characteristics of both datasets.

Table 3.1: Summary of two datasets after location and user extraction

Dataset	No. of people	No. of locations	Duration (Days)	Period of a day
MIT	43	482	75	8 am to 12 pm
Dartmouth	162	399	118	8 am to 12 pm

3.2 The Proposed Mobility Prediction Model

3.2.1 Problem Definition

Recall that $\mathbb{M}$ and $\mathbb{L}$ are the set of m members in the entire community and the set of locations, respectively. We take into account the problem of predicting the position of person p ($1 \leq p \leq m$) at current or future time t with the precondition that the recent locations of person p are unknown. Assume that the historical positions of $(m-1)$ other members can be observed. Therefore, we can make use of the historical information of the rest of people for predicting where person p stays at the current or future time.

Let us define $\boldsymbol{y_t}$ as the location of person p at time t and R as the spatio-temporal information of the rest of members during the historical period. Specifically, $R = \{R_j^i | R_j^i \in \mathbb{L}, i \in \mathbb{M}, i \neq p, 1 \leq j \leq t\}$. The problem of predicting human location is formulated as follows. The objective is to predict where person p is the most likely to visit at time t given the recent information R, or to maximize the following probability:

$$P(\boldsymbol{y_t}|R) \tag{3.1}$$

3.2.2 Two-phase Mobility Prediction Framework

The proposed human mobility prediction framework consists of two phases, i.e., detection of persons with correlated movements (PCMs) and PCM-based location prediction, as shown in Fig. 3-1. Let r denote the number of selected PCMs. In the first phase, r from among $(m-1)$ remaining members are extracted as PCMs of person p. Then, given the location information of the chosen members, the current or future position of person p is predicted.

In the first phase, we measure independently the social correlation between person p and each of the $(m-1)$ other members. In order to choose r useful PCMs for person p, these $(m-1)$ people are arranged according to their correlation scores, and r members with the highest scores are selected. The training set is used in the first phase where the input is vector $\boldsymbol{x}$ that represents positional information of the $(m-1)$ remaining members. Table 3.2 shows an example of vector $\boldsymbol{x}$ when the location of an arbitrary person, p, is estimated,

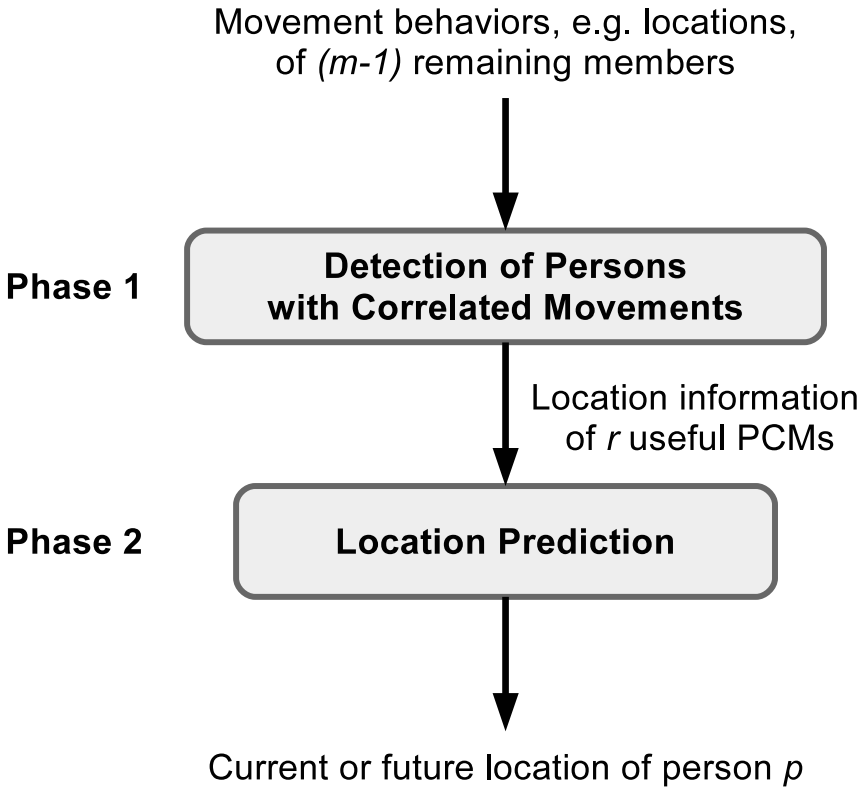


Figure 3-1: The proposed framework for predicting the current or future location of person p

where $\boldsymbol{x}$ consists of spatial and temporal information. More specifically, $x_1^{\text{loc}}, x_2^{\text{loc}}, ..., x_{p-1}^{\text{loc}}$, x_{p+1}^{loc},..., x_m^{loc} denotes the positions of the $(m-1)$ remaining members, while x^{day} and x^{time} account for the day and time slot indices, respectively. The first phase returns r PCMs of person p as the output.

Then, the location prediction phase estimates the most likely current or future position of person p using the spatio-temporal information of the r selected PCMs. Each sample of the training set is given by $(\boldsymbol{z}_{i_1}^{i_2}, \boldsymbol{y}_t)$, where label $\boldsymbol{y}_t$ is location information of person p at time t, and feature vector $\boldsymbol{z}_{i_1}^{i_2}$ is spatio-temporal information of PCMs from time slots i_1 to i_2 where $1 \leq i_1 \leq i_2 \leq t$. In the proposed framework, a conditional probability distribution is estimated with the objective of maximizing probability $P(\boldsymbol{y}_t|\boldsymbol{z}_{i_1}^{i_2})$. More specifically, given the location information of PCMs at time slots between i_1 and i_2, the position of person p at time t needs to be predicted.

In this work, $(t-1) \leq i_1 \leq i_2 \leq t$ is considered, i.e., $(i_1, i_2) = \{(t-1, t-1), (t-1, t), (t, t)\}$. In cases where $i_1 = i_2 = (t-1)$, the location of PCMs in the previous time slot $(t-1)$ is used to predict the next location of person p (i.e., estimate person p's future location at time t). Meanwhile, $(i_1, i_2) = (t-1, t)$ indicates that the model estimates the current position of the

person using previous and current location information of PCMs. Whereas, if $i_1 = i_2 = t$, the current position of the person is estimated given the information of PCMs at time t. Hereafter, the model is regarded as estimating the location of a person at time slot t. In the second phase, the PCM-based location prediction model is used to label an unknown sample as one out of $|\mathbb{L}|$ possible locations.

Table 3.2: Elements of the input vector in the first phase of the model for predicting the location of person p

Feature	**Description**
$\{x_i^{\text{loc}}, 1 \leq i \leq m, i \neq p\}$	location of $(m-1)$ remaining members
x^{day}	index of day in a week
x^{time}	index of time slot in a day

There are several benefits from the proposed framework. Our model does not require a recent location sequence of person p when predicting the current or future location of person p, which is beneficial in cases where the location information of that person is not available. Note, however, that the location information of person p is necessary for training parameters of the prediction model.

The proposed framework first selects r PCMs based on human mobility characteristics, e.g., encounter frequency and behavioral patterns. Then, the location information of only these chosen PCMs is used to predict a person's location at time t. As a result, by reducing redundant input features of the second phase, the overfitting problem can be mitigated. Moreover, the proposed framework allows for low time and space complexity.

3.3 Detection of Persons with Correlated Movements

In this section, two methods for selecting PCMs are proposed: community interaction similarity-based (CISB) and behavioral similarity-based (BSB) methods. Each uses a different measurement score to estimate the closeness between people's movement patterns. Then, the r PCMs with the highest scores are selected.

3.3.1 Community Interaction Similarity-Based Method

Recall that the encounter frequency-based (EFB) method only uses direct encounters between certain individuals to estimate their friendship. In contrast, to measure correlation scores between persons p and q, CISB considers interactions between these persons and other

community members as well as direct encounters between them. The CISB method is inspired by the fact that interactions and relationships with other members in the community have a great impact on a person's behaviors.

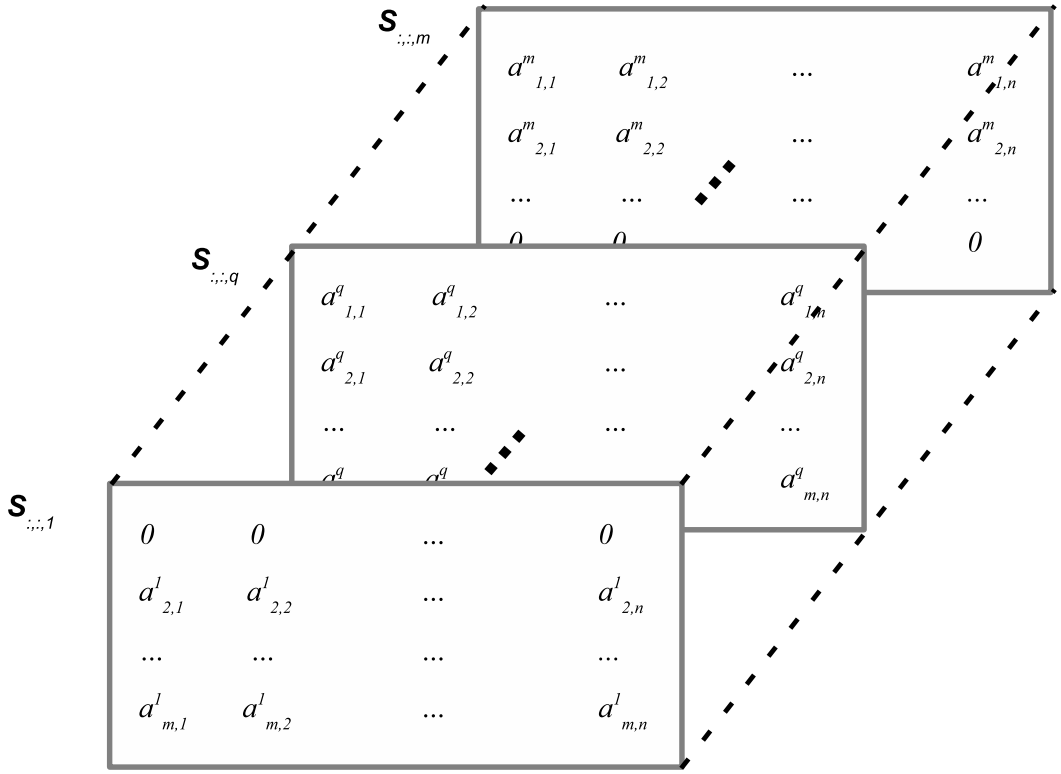


Figure 3-2: A three-dimensional CIS tensor, denoted by $\mathbf{S}$

The CISB method utilizes the community interaction-based similarity (CIS) tensor shown in Fig. 3-2, which represents the spatio-temporal encounters between people. The 3D CIS tensor, $\mathbf{S}$, consists of m layers each of which is a matrix, e.g., $\mathbf{S}_{:,:,q}$ reflects the interactions between person q and the rest of the $(m-1)$ people. More specifically, the p^{th} row in the matrix, i.e., $\mathbf{S}_{p,:,q}$, indicates the temporal encounters of persons p and q. Let n_{day} and n_{time} denote the total number of days in the overlapping period and the number of time slots per day, respectively. $n = n_{\text{day}} \times n_{\text{time}}$ indicates the total number of samples per person during the whole period. $a^q_{p,t} := \mathsf{S}_{p,t,q}$ accounts for the encounter between persons p and q in time slot t.

Vector $\mathbf{S}_{p,:,q}$, representing the encounters between persons p and q, is defined as follows:

$$\mathsf{S}_{p,t,q} = \begin{cases} 1, & \text{if persons } p \text{ and } q \text{ are in the same cell} \\ & \text{during time slot } t \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

Person p is not considered able to encounter himself or herself, because tensor S represents the interactions between a person and the entire community, i.e., $\forall t, p : \mathsf{S}_{p,t,p} = 0$. The CIS tensor owns a symmetric property, i.e., $\forall t : \mathsf{S}_{p,t,q} = \mathsf{S}_{q,t,p}$.

From the CIS tensor, we reshape every 2D matrix, $\mathbf{S}_{:,:,q}$, to a one-dimensional vector with $(m \times n)$ elements, which is denoted by $\boldsymbol{s}^{(q)}$. In order to evaluate the closeness between persons p and q, we consider the meet/min correlation coefficient [89] for measuring the similarity between $\boldsymbol{s}^{(p)}$ and $\boldsymbol{s}^{(q)}$, which can be used if the vectors include a lot of 1s. Thus, score $d_{p,q}$ between nodes p and q is calculated as follows:

$$d_{p,q} = \frac{\boldsymbol{s}^{(p)\top}\boldsymbol{s}^{(q)}}{\min\left(|\boldsymbol{s}^{(p)}|, |\boldsymbol{s}^{(q)}|\right)} \tag{3.3}$$

where $|\boldsymbol{s}^{(q)}|$ is the L^1-norm of vector $\boldsymbol{s}^{(q)}$. After obtaining scores between person p and $(m-1)$ others, r people with the highest scores are chosen as PCMs.

Because of the characteristics of CIS, i.e., $\forall t, p : \mathsf{S}_{p,t,p} = 0$, and the dot product in the numerator of Eq. (3.3), when measuring similarity $d_{p,q}$, the interactions between persons p and q vanished unintentionally. For example, in calculating $d_{1,2}$, two vectors $\boldsymbol{s}^{(1)}$ and $\boldsymbol{s}^{(2)}$ are as follows:

$\boldsymbol{s}^{(1)} = [0 \;\; 0 \;\; ... \;\; 0 \;\; a^1_{2,1}\, a^1_{2,2} ... a^1_{2,n} \; ... \;\; a^1_{m,1} \;\; a^1_{m,2} \;\; ... \;\; a^1_{m,n}]$

$\boldsymbol{s}^{(2)} = [a^2_{1,1}\, a^2_{1,2} ... a^2_{1,n} 0 \;\; 0 \;\; ... \;\; 0 \;\; ... \;\; a^2_{m,1} \;\; a^2_{m,2} \;\; ... \;\; a^2_{m,n}]$

Since the first n elements of $\boldsymbol{s}^{(1)}$ and the $(n+1)^{\text{th}}$ to $(2n)^{\text{th}}$ elements of $\boldsymbol{s}^{(2)}$ are zeros, similarity score $d_{1,2}$ that is estimated using Eq. (3.3) does not consider the encounters of persons 1 and 2.

To avoid this vanishing problem, when calculating $d_{p,q}$, the CIS tensor is temporarily modified as follows: $\mathsf{S}_{p,t,p} = \mathsf{S}_{p,t,q}$ and $\mathsf{S}_{q,t,q} = \mathsf{S}_{q,t,p}$. For example, the modified vectors $\boldsymbol{s'}^{(1)}$ and $\boldsymbol{s'}^{(2)}$ corresponding to $\boldsymbol{s}^{(1)}$ and $\boldsymbol{s}^{(2)}$, respectively, can be obtained as follows:

$\boldsymbol{s'}^{(1)} = [a^2_{1,1}\, a^2_{1,2} ... a^2_{1,n}\, a^1_{2,1}\, a^1_{2,2} ... a^1_{2,n} \;\; ... \;\; a^1_{m,1}\, a^1_{m,2} ... a^1_{m,n}]$

$\boldsymbol{s'}^{(2)} = [a^2_{1,1}\, a^2_{1,2} ... a^2_{1,n}\, a^1_{2,1}\, a^1_{2,2} ... a^1_{2,n} \;\; ... \;\; a^2_{m,1}\, a^2_{m,2} ... a^2_{m,n}]$

By using the modified CIS tensor, the CISB method takes into account the interactions with the entire community members.

Table 3.3: Behaviors of Persons p, q, and u

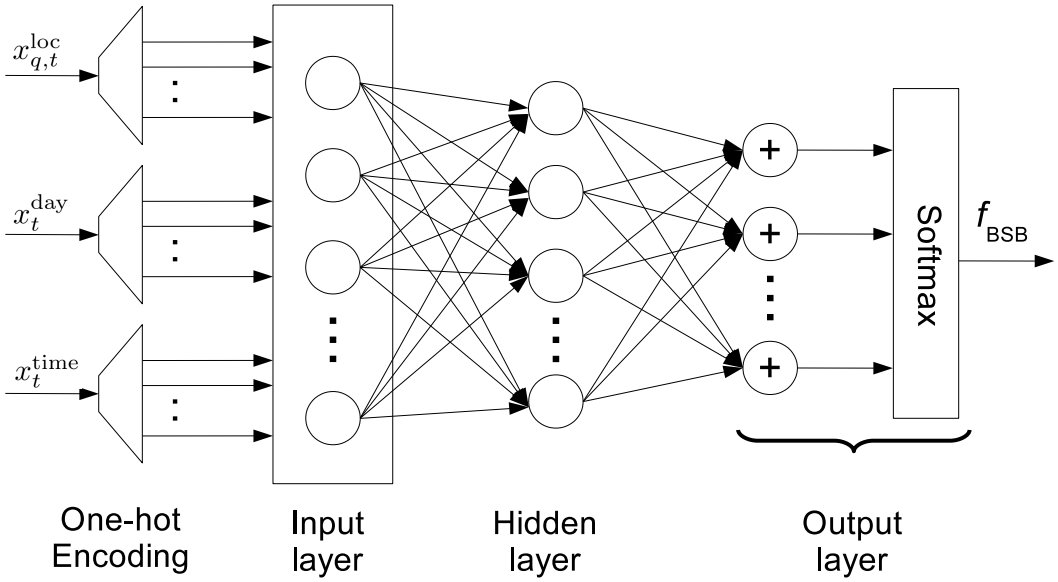
	day 1				**day 2**			
Time	1 pm	3 pm	5 pm	7 pm	1 pm	3 pm	5 pm	7 pm
p	math class	gym	restaurant B	tea house	lab D	gym	restaurant B	home
q	office A	café	restaurant C	piano class	office A	café	restaurant C	piano class
u	math class	market	park	home	lab D	lab D	library	concert hall

3.3.2 Behavioral Similarity-Based Method

Persons p and q are said to have similar behavior if they experience a correlated visit behavior pattern, and hence, the location of a person can be estimated by using the other's positional information. The places two people with similar behaviors visit at the same time are denoted as correlated locations. Note that correlated locations may be different. As seen in Table II, persons p and q show a lot of similar behaviors even though they do not encounter each other. For example, person p exercises at a gym whenever person q visits a café. It is beneficial to measure the behavior similarity between persons because the location of a person can be inferred by using information of another person with a correlated behavioral pattern. For instance, the location of p at 3 p.m. on day 2 can be predicted using person q's location (café). Based on this observation, in the BSB method, persons with the highest behavioral correlation are chosen as PCMs.

The key difference between the BSB method and friendship-based methods (i.e., CISB and EFB) is as follows. When selecting a PCM, the BSB method does not consider real friendships between people while friendship-based methods extract co-location information to measure the friendships between them. Specifically, in the EFB method, members with the highest number of encounters with a given person are chosen as PCMs. As shown in Table II, persons p and u encounter each other several times (in math class and the laboratory), which indicates a close friendship between them. Therefore, in friendship-based methods, person u is likely to be chosen as a PCM of person p. Meanwhile, in case of two people with similar behaviors but different correlated locations, a person is not likely to be selected as a PCM of the other person. For example, persons p and q have the same behavior of having dinner at 5 p.m. at restaurant B and restaurant C, respectively. Since there is no encounter between people p and q, in friendship-based methods person q does not obtain a high enough score to be selected as a person p's PCM. On the other hand, the BSB method considers this correlated behavior for choosing PCMs.

In the BSB method, a feed-forward neural network (NN), shown in Fig. 3-3, is constructed to evaluate the behavioral similarity between persons p and q. $x_{q,t}^{\text{loc}}$ is defined as the location of person q at time t, while x_t^{day} and x_t^{time} are indices of the day and the time slot at time t, respectively. The objective is to learn the underlying conditional probability distribution $f_{\text{BSB}}(x_{q,t}^{\text{loc}}, x_t^{\text{day}}, x_t^{\text{time}}) = P(x_{p,t}^{\text{loc}} | x_{q,t}^{\text{loc}}, x_t^{\text{day}}, x_t^{\text{time}})$, i.e., the probability mass distribution that person p stays in a specific location during time slot t given the spatio-temporal information of person q.



Output of NN: $f_{\text{BSB}} = P(x_{p,t}^{\text{loc}} | x_{q,t}^{\text{loc}}, x_t^{\text{day}}, x_t^{\text{time}})$

Figure 3-3: Architecture of the neural network for measuring behavioral similarity between persons p and q in the BSB method

Here, the training process is described. First, symbolic location $x_{q,t}^{\text{loc}}$ is embedded into an $|\mathbb{L}|$-dimensional indicator vector using one-hot encoding, because there is no ordinal relationship between locations. Similarly, time slot and day indices (i.e., x_t^{day} and x_t^{time}, respectively) are also converted to indicator vectors. Then, these vectors are used as input units of the neural network including one hidden layer with 150 logistic sigmoid activation nodes. These input features propagate progressively throughout the network until reaching the softmax output layer, which obtains a conditional probability distribution over locations. Θ is defined as the parameters of the NN consisting of weights, denoted by $\boldsymbol{W}$, and biases that are randomly initialized. Let η denote the number of samples in the training dataset. After retrieving the conditional probability at the output layer, regularized cost function $J(\Theta)$ using cross-entropy error [90] is calculated as follows:

$$J(\Theta) = -\frac{1}{\eta} \sum_{j=1}^{\eta} \boldsymbol{t}^{(j)\top} \log f_{\text{BSB}}^{(j)}(x_{q,t}^{\text{loc}}, x_t^{\text{day}}, x_t^{\text{time}}) + \Omega(\boldsymbol{W}) \tag{3.4}$$

where $f_{\text{BSB}}^{(j)}(x_{q,t}^{\text{loc}}, x_t^{\text{day}}, x_t^{\text{time}})$ and $\boldsymbol{t}^{(j)}$ are the estimated output and target vectors of the j^{th} instance, respectively. Note that log(.) is an element-wise operation. $\Omega(\boldsymbol{W})$ is the L^2-based regularization term. The neural network model is trained to minimize cost function $J(\Theta)$ using a back-propagation algorithm combined with a gradient descent method.

In order to select r PCMs of person p, $(m-1)$ NNs are trained separately, where each NN measures the similar behavior pattern between individuals p and q $(p, q \in \mathbb{M}, p \neq q)$. The prediction accuracy, $\xi_{p,q}$, of the q^{th} NN, which indicates the behavioral similarity of persons p and q, is defined as the ratio of the number of correctly predicted instances to the number of test samples. In the BSB method, r PCMs among the $(m-1)$ remaining members with the highest $\xi_{p,q}$ are selected.

It is worthwhile to note that there are alternative ways of selecting a set of PCMs. For example, forward search (FS) [91] begins with an empty set and then progressively incorporates each feature (i.e., PCM) into the set. However, FS has an inherent weakness, i.e., high time and space complexity when a large number of features need to be selected.

Another simple approach that can be considered is that positional information of all $(m-1)$ remaining people is used as input features for a prediction model that labels the location of a person at time t with one of the $|\mathbb{L}|$ positions. In our work, this simple method is called ignored feature selection (IFS) since there is no feature selection process. Note that IFS may need to train a model with tens or hundreds of thousands of features, which may result in a long training time, and might suffer from the overfitting problem.

In order to gain insight into computational costs, the number of training parameters for three methods including IFS, FS, and BSB is compared. For IFS, there is no PCM selection phase, i.e., the location of person p is predicted given the positional information of the $(m-1)$ other people. With FS, to choose the first PCM, $(m-1)$ NNs are trained independently. To select the second PCM, a combination of the first PCM and one of the $(m-2)$ remaining members is examined; $(m-2)$ NNs need to be learned, each of which estimates person p's location at time slot t given the contextual information of two examined people. Similarly, to select the r^{th} PCM, $(m-r)$ NNs are used to evaluate the subset of the $(r-1)$ already selected PCMs and one of the $(m-r)$ remaining members. In the FS method, a large r value leads to a significant increase in the number of training parameters.

In the BSB and FS methods, it is assumed that three selected PCMs are used for the location prediction phase, where a neural network with three hidden layers of 500, 150, and 50 neurons is considered. For a fair comparison, the number of hidden neurons for IFS is proportional to that of the BSB/FS methods in the mobility prediction phase with the ratio $(m-1) : r$ because $(m-1)$ and r are the numbers of people whose location information is given to the IFS and BSB/FS methods, respectively.

Table 3.4 compares the number of training parameters of the three methods, where

Table 3.4: Comparison of the number of parameters among behavioral similarity-based (BSB), ignored feature selection (IFS), and forward search (FS) methods

m	l_{BSB} (millions)	l_{IFS} (millions)	l_{FS} (millions)
43	7.00	59.02	158.30
100	15.36	142.09	878.17
500	74.06	725.03	2,229.0
1000	147.42	1,453.7	8,933.1

l_{BSB}, l_{IFS}, and l_{FS} denote the number of parameters, i.e., weights and biases, in behavioral similarity-based, ignored feature selection, and forward search methods, respectively. When the total number of people, m, increases, the three considered methods need to train more parameters. However, the BSB method always learns the least number of parameters. Moreover, the increase rate is quite different in each method, e.g., when m varies from 43 to 1000 people, l_{BSB} rises $\frac{147.42}{7.0} \approx 21$ times, compared to $\frac{1,453.7}{59.02} \approx 24.6$ times with l_{IFS} and $\frac{8,933.1}{158.3} \approx 54.6$ times with l_{FS}. In summary, because only location information of PCMs is used to predict human mobility of a given person, the proposed BSB method incurs a much lower computation cost in terms of time and space complexity than IFS and FS, which makes the BSB method more practical, particularly when there is a large number of people.

3.4 The PCM-based Location Prediction (PLP) Model

In this section, we propose the PCM-based location prediction (PLP) model based on the neural network to predict human mobility given the movement patterns of r selected PCMs. Figure 3-4 illustrates the NN-based prediction model where the location of a person at time t is estimated given the information of r PCMs, i.e., $\boldsymbol{z}_{i_1}^{i_2} = \{z_{l,\tau}^{\text{loc}}, z_t^{\text{day}}, z_t^{\text{time}}\}$. Spatial information $z_{l,\tau}^{\text{loc}}$, where $1 \leq l \leq r$ and $i_1 \leq \tau \leq i_2$, indicates the symbolic location of the r PCMs from time i_1 to i_2. Meanwhile, temporal information z_t^{day} and z_t^{time} are the indices of the day of the week and the time slot in the day at time t, respectively. Let $y_{\text{NN}}^{\text{pred}} \triangleq f_{\text{NN}}(z_{l,\tau}^{\text{loc}}, z_t^{\text{day}}, z_t^{\text{time}})$ denote the output of the NN-based predictor, i.e., the most likely location of the person during time slot t. Note that to predict the location at time t of m persons in the community, m corresponding NN-based prediction models need to be trained separately. The remainder of this subsection describes an NN-based predictor.

The function $f_{\text{NN}}(z_{l,\tau}^{\text{loc}}, z_t^{\text{day}}, z_t^{\text{time}})$ can be decomposed into three parts, as follows.

1. The contextual information of r PCMs, i.e., $z_{1,i_1}^{\text{loc}}, z_{1,i_2}^{\text{loc}}, ..., z_{r,i_2}^{\text{loc}}, z_t^{\text{day}}, z_t^{\text{time}}$, is mapped

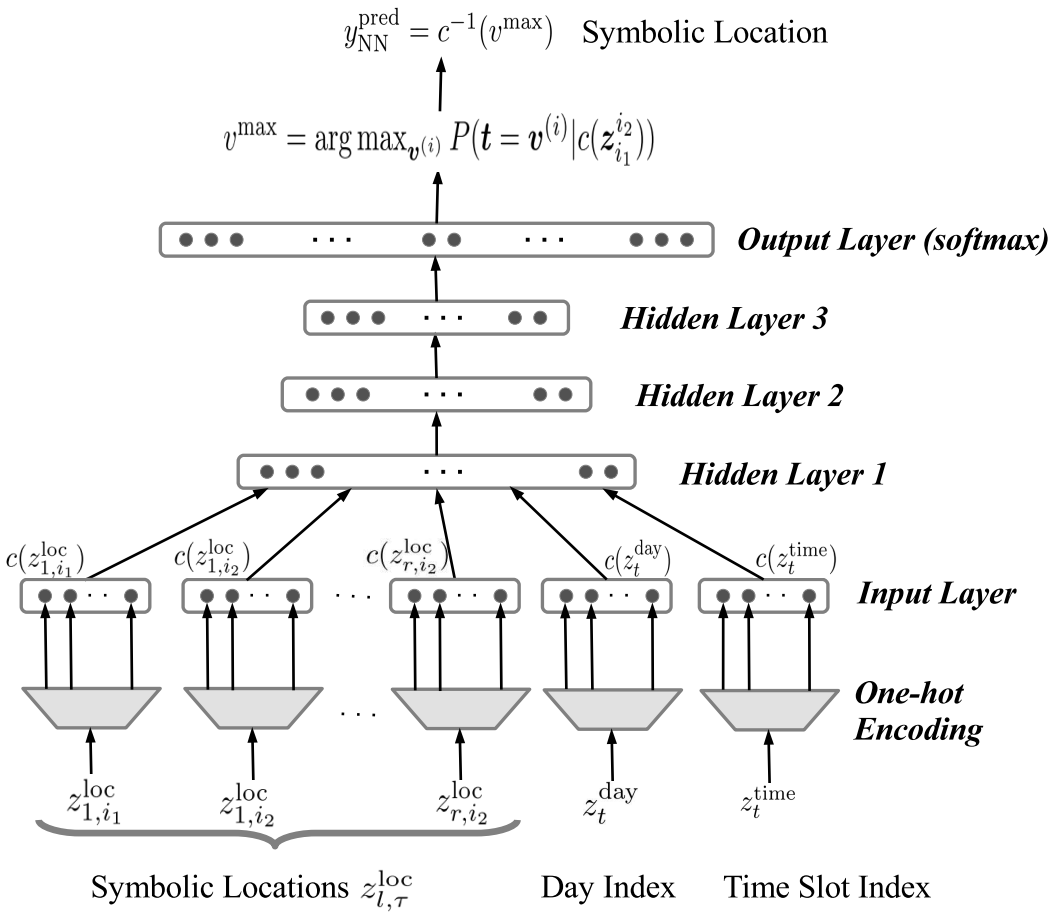


Figure 3-4: The PCM-based location prediction model

onto real vectors using the one-hot encoding function $c(.)$, where output is the indicator vector in which only one element is set to 1 while others are 0. Since $|\mathbb{L}|$ is the size of the extracted location set, $c(z_{1,i_1}^{\mathrm{loc}}), c(z_{1,i_2}^{\mathrm{loc}}), ..., c(z_{r,i_2}^{\mathrm{loc}})$ are vectors in $\mathbb{B}^{|\mathbb{L}|}$, where $\mathbb{B}$ is a set of binary numbers; $c(z_t^{\mathrm{day}})$ is a day-index vector in $\mathbb{B}^7$. Meanwhile, the cellular trace between 8 a.m. and 12 p.m. is extracted and each time slot lasts 30 minutes; therefore, z_t^{time} is mapped onto a 32-dimensional binary vector. In addition, grouped time slots are considered. In this case, $c(z_t^{\mathrm{time}})$ is a three-dimensional vector where each dimension corresponds to one of the three parts of a day, i.e., morning, afternoon, and evening.

2. The NN classifier maps input sequence $c(\boldsymbol{z}_{i_1}^{i_2}) = \{c(z_{l,\tau}^{\mathrm{loc}}), c(z_t^{\mathrm{day}}), c(z_t^{\mathrm{time}})\}$ to a conditional probability distribution over locations. Let $\boldsymbol{v}^{(i)}$ denote an indicator vector in which only the i^{th} element is 1 and others are 0 ($1 \leq i \leq |\mathbb{L}|$). Vector $\boldsymbol{t}$ is defined as

the target vector of the NN classifier. The output of the NN is denoted by vector $\hat{\boldsymbol{y}}_{\text{NN}}$ where the i^{th} element, $\hat{y}_i$, estimates posterior probability $\hat{y}_i = P(\boldsymbol{t} = \boldsymbol{v}^{(i)}|c(\boldsymbol{z}_{i_1}^{i_2}))$ that the person stays in location i given the information of r PCMs.

3. The vector $v^{\max} = \arg\max_{\boldsymbol{v}^{(i)}} P(\boldsymbol{t} = \boldsymbol{v}^{(i)}|c(\boldsymbol{z}_{i_1}^{i_2}))$ that indicates the most likely location of the person is mapped onto symbolic position $y_{\text{NN}}^{\text{pred}}$ by the inverse function, i.e., $y_{\text{NN}}^{\text{pred}} = c^{-1}(v^{\max})$.

Now, the neural network training process that includes feed-forward propagation, a cost function calculation, and a parameter update is presented. First, feed-forward propagation is described. In this work, μ is defined as the number of hidden layers, and a neural network is considered with logistic sigmoid activation units denoted by $\sigma(.)$. Hereafter, the input layer is referred to as layer 0, the j^{th} hidden layer as layer j $(1 \leq j \leq \mu)$, and the output layer as layer $(\mu + 1)$. Let $\Theta = \{\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, ..., \boldsymbol{W}^{(\mu+1)}, \boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, ..., \boldsymbol{b}^{(\mu+1)}\}$ denote all learning parameters consisting of weights and biases of the NN classifier. More specifically, $\boldsymbol{W}^{(j)}$ and $\boldsymbol{b}^{(j)}$ are the matrix of weights and the vector of biases, respectively, that indicate connections from layers $(j-1)$ to j; $\boldsymbol{h}^{(j)}$ is defined as the state vector at layer j, and $\boldsymbol{h}^{(j)}$ is given by:

$$\boldsymbol{h}^{(j)} = \sigma\big(\boldsymbol{W}^{(j)\top}\boldsymbol{h}^{(j-1)} + \boldsymbol{b}^{(j)}\big), \quad 1 \leq j \leq \mu \tag{3.5}$$

Note that $\boldsymbol{h}^{(0)}$ is a vector of input features at layer 0. For the output layer, the unnormalized output vector $\boldsymbol{h}^{(\mu+1)}$ is calculated by $\boldsymbol{h}^{(\mu+1)} = \boldsymbol{W}^{(\mu+1)\top}\boldsymbol{h}^{(\mu)} + \boldsymbol{b}^{(\mu+1)}$.

Then, the output of a softmax function represents the categorical distribution over locations where the i^{th} element of the output vector reflecting the probability that the person stays at time t in location i is calculated as follows:

$$\hat{y}_i = P(\boldsymbol{t} = \boldsymbol{v}^{(i)}|c(\boldsymbol{z}_{i_1}^{i_2})) = \frac{e^{h_i^{(\mu+1)}}}{\sum_{j=1}^{|\mathbb{L}|} e^{h_j^{(\mu+1)}}} \tag{3.6}$$

where $h_i^{(\mu+1)}$ is the i^{th} value of unnormalized output vector $\boldsymbol{h}^{(\mu+1)}$.

Secondly, the calculation of the cost function is briefly described below. Assume that there are η training instances. Cost function $J(\Theta)$ using the cross entropy metric and regularization method can be expressed as follows:

$$J(\Theta) = -\frac{1}{\eta}\sum_{j=1}^{\eta} \boldsymbol{t}^{(j)\top} \log \hat{\boldsymbol{y}}_{\mathrm{NN}}^{(j)} + \Omega(\boldsymbol{W}) \tag{3.7}$$

where $\hat{\boldsymbol{y}}_{\mathrm{NN}}^{(j)}$ and $\boldsymbol{t}^{(j)}$ are, respectively, the estimated output and target vectors of the j^{th} instance, and $\boldsymbol{W}$ denotes the weight matrices. Note that $\log \hat{\boldsymbol{y}}_{\mathrm{NN}}^{(j)}$ is an element-wise operation. Regularization $\Omega(\boldsymbol{W})$ is implemented using the L^2 parameter norm penalty.

Finally, parameters including weights and biases are obtained by applying a back propagation algorithm, which minimizes the cost value using the gradient descent method with momentum. Let α and γ denote the learning rate of gradient descent and the momentum coefficient, respectively, $(0 \leq \gamma \leq 1)$. Let $\tilde{\theta}$ denote a training parameter in the set Θ, i.e., $\tilde{\theta} \in \Theta$. Then, each $\tilde{\theta}$ is updated in the j^{th} epoch as follows:

$$\omega = \gamma\omega + \alpha\frac{\partial J(\Theta)}{\partial \tilde{\theta}} \tag{3.8}$$

$$\tilde{\theta}^{(j)} = \tilde{\theta}^{(j)} - \omega \tag{3.9}$$

where partial derivative $\dfrac{\partial J(\Theta)}{\partial \tilde{\theta}}$ is the gradient of the cost function with respect to $\tilde{\theta}$, and ω is the current velocity vector with the same dimensions as parameter $\tilde{\theta}$.

Table 3.5: Setup for the location prediction framework

Parameters	**Values**
PCM selection method	{SC, STC, EFB, CISB, BSB }
Prediction model	{PLP}
Temporal location type	{pre-loc, cur-loc, **pre-cur-loc**}
Time slot feature	{un-grouped, **grouped** }
Number of selected PCMs (r)	{1, 2, **3**, 4, 5}
Top-k accuracy	{**top-1**, top-2, top-3, top-4}

3.5 Evaluation Results and Discussion

In this section, the performance of the human mobility prediction framework is examined under different PCMs detection methods and is compared with the baseline prediction model, most frequent location (MFL). The whole dataset is randomly partitioned into training, validation, and test sets at the ratio of 5:2:3. Recall that in case of the Dartmouth dataset the 118-day period is selected and for each day human mobility from 8h to 24h is considered.

Since each time slot lasts 30 minutes, there are total 32 time slots per day. Therefore, the number of samples in the Dartmouth dataset is $118 \times 32 = 3,776$. The training, validation, and test sets consist of 1,886; 755; and 1,133 instances, respectively. The training set is used to fit the model while the purpose of validation set is to determine the appropriate hyper-parameters for the neural network, e.g., the number of hidden layers, hidden units, activation units, and learning rate. Note that only the training set is used in the first phase of the prediction model and the performance of model is estimated based on the test set.

Table 3.5 presents the setup for the mobility prediction framework. Specifically, the two proposed PCM selection methods consisting of community interaction similarity-based (CISB) and behavioral similarity-based (BSB) are evaluated and compared with three recent selection approaches: encounter frequency-based method (EFB) [56,70,92], spatial closeness (SC) [8], and spatiotemporal closeness (STC) [8]. Table 3.6 shows a list of acronyms which are used in the manuscript.

Table 3.6: List of Acronyms

Abbreviation	**Meaning**
BSB	Behavioral similarity-based
CISB	Community interaction similarity-based
EFB	Encounter frequency-based
MFL	Most frequent location
PCM	Persons with correlated movements
PLP	PCM-based location prediction
SC	Spatial closeness
STC	Spatiotemporal closeness

Encounter frequency is usually considered as a metric for measuring the friendship between two users [92], especially finding social friends, colleagues, or coworkers. In EFB, PCMs are selected according to the assumption that two members who encounter more frequently are assumed to hold a closer relationship. We first define that a co-cell event occurs when the mobile phones of two users are associated with the same cell tower at the same time slot. The higher number of co-cell events means the more frequently two members stay in the same place, which can indicate the stronger social tie between two users. Therefore, in EFB, members with the highest frequency of co-cell events are selected as PCMs of a considered person.

The EFB method is implemented as follows. In order to represent whether persons p

and q are at the same location during time slot t, an indicator $\boldsymbol{I}_{p,q}(t)$ is defined as follows:

$$\boldsymbol{I}_{p,q}(t) = \begin{cases} 1, & \text{if } p \text{ and } q \text{ are in the same cell} \\ & \text{during time slot } t \\ 0, & \text{otherwise} \end{cases} \tag{3.10}$$

where $p, q \in \mathbb{M}$. $\boldsymbol{I}_{p,p}(t) = 1$ if person p stays at one of $|\mathbb{L}|$ feasible locations at time t. If the location of p or q is undefined at time slot t, then, $\boldsymbol{I}_{p,q}(t) = 0$. The friendship correlation $\rho_{p,q}$ $(0 \leq \rho_{p,q} \leq 1)$, which indicates the link weight between persons p and q, is defined with the consideration of number of co-cell events between them as follows:

$$\rho_{p,q} = \frac{\sum_t I_{p,q}(t)}{\sum_t I_{p,p}(t)} \tag{3.11}$$

Then, among $(m-1)$ remaining members, r people who have the highest $\rho_{p,q}$ are selected as user p's SCs.

Recall that Alhasoun *et al.* [8] constructed a prediction model based on dynamic Bayesian networks which leveraged historical location data of selected PCMs to predict the most probable position of the person of interest. However, the prediction model in [8] requires to know the last visited place of the person of interest, which is not applicable to our considering problem where the historical position data of the person of interest is assumed to be unknown. Note that since there was no existing prediction model designed for the considering problem, the PCM-based location prediction (PLP) model is used in the experiments.

Additionally, we make a performance comparison between our two proposed PCMs selection approaches and counterpart methods in [8]. There were three similarity measurement metrics to select PCMs in [8], and specifically the first method, temporal closeness, compares patterns of communication (e.g., call, sms, and data) between people. Due to the requirement of extra communication information of members, the temporal closeness is not considered as a counterpart method for selecting PCMs in our work. Therefore, we compare the proposed PCMs selection methods with only two approaches in [8]: STC and SC.

For the PLP model, a variety of NN architectures which have a different number of layers, hidden units, and activation functions were examined. Then, among the considered setting, the most appropriate NN architecture was selected by using the validation set. Specifically, there is 1 hidden layer of 150 sigmoid activation units in the NN of the BSB approach. In the NN-based predictor, 500, 150, and 50 sigmoid hidden units are used in the first,

second, and third hidden layers, respectively. The reason of selecting a neural network to construct a prediction model is explained below. Figure 3-5 shows the performance of the prediction model constructed by using different machine learning classifiers, e.g., support vector machine (SVM), Naive Bayes (NB), and logistic regression (LR). Since achieving the highest prediction accuracy, the NN-based model is used in this work.

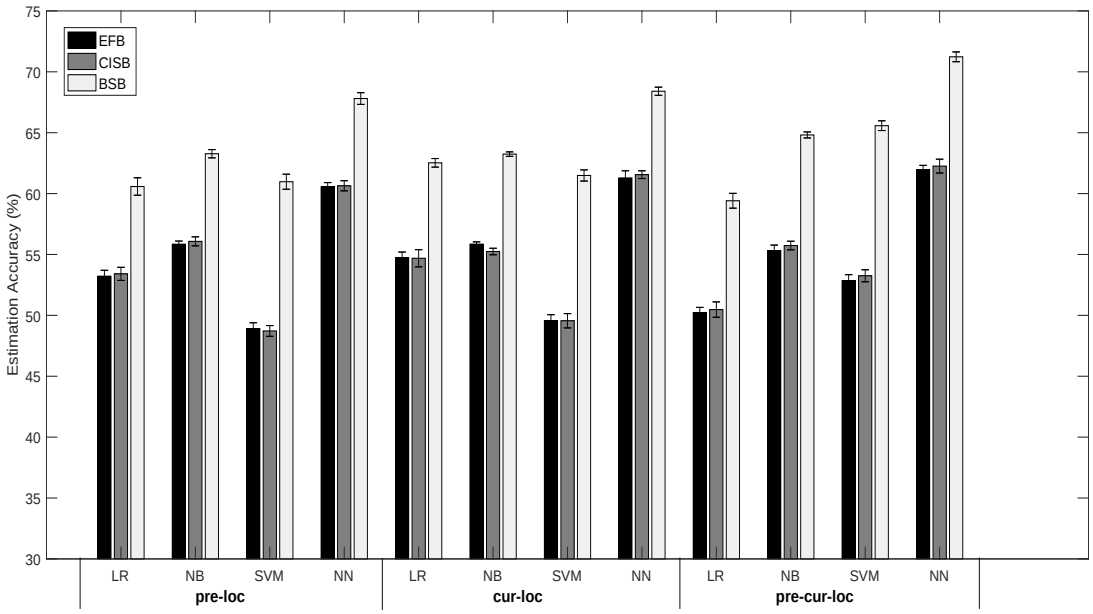


Figure 3-5: The effects of machine learning classifiers on the performance of the prediction model

Weight and bias values are initialized with normal distribution. The L^2 regularization coefficient is set to 0.01. Learning rate in the gradient descent and the momentum coefficient are set to 0.4 and 0.1, respectively. Full batch learning with 1500 iterations is examined. Note that only the training set is used to select the r PCMs in the first phase of the prediction models. For example, in case of the Dartmouth dataset, r PCMs are determined by using the training set of 1,866 samples which are randomly selected from the total 3,776 instances.

Moreover, three types of temporal information are considered. The number of selected PCMs varies from 1 to 5, while top-k accuracy is considered with $k = \{1, 2, 3, 4\}$. In addition, un-grouped and grouped time slot features are compared. The bold text in the second column of Table 3.5 denotes default values of the prediction model. In this work, performance metrics including average and standard deviation of prediction accuracy are obtained over 5 runs. The results are collected via averaging the entire people.

The most noticeable results are summarized as follows. From the conducted experi-

ments, the performance of designed PLP model outperforms that of the baseline method, MFL. With regard to the PCMs selection methods, the proposed BSB shows significantly better performance than other PCM extraction approaches. In addition, the generalization capability of the proposed framework increases as a larger number of PCMs are embedded into the model. In particular, this increase is more clearly shown in case of BSB.

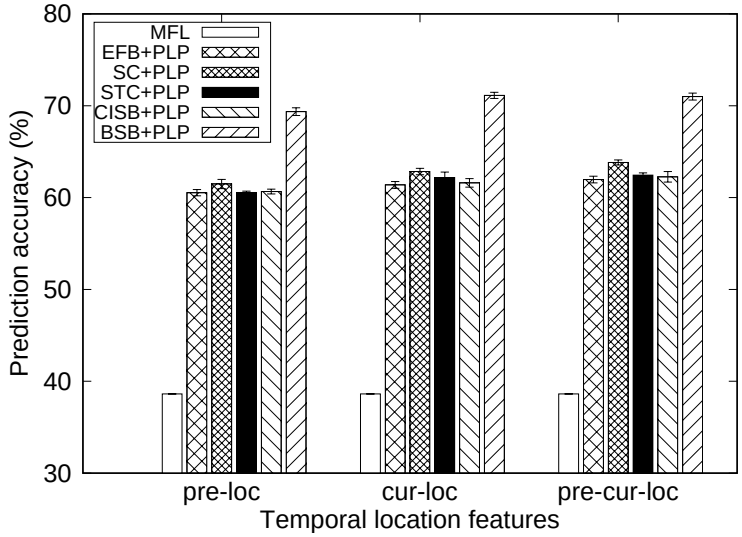
3.5.1 Performance Comparison of PCM Selection Methods

In this subsection, the performance results of PCMs extraction methods are discussed with different temporal features of selected PCMs. The location of a person during time t is predicted with the support of PCMs' positions during the time slot $(t-1)$ (denoted by pre-loc), time slot t (denoted by cur-loc), and both $(t-1)$ and t time slots (denoted by pre-cur-loc). Recall that the time slot lasts 15 minutes. The obtained results are shown in Fig. 3-6.

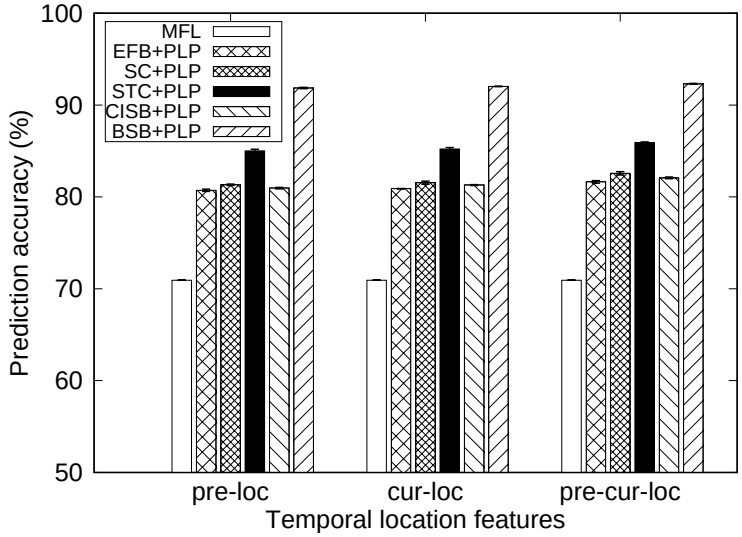
First, we evaluate the predictability of the baseline prediction model, most frequent location (MFL). As can be seen in Fig. 3-6, the MFL model achieves much lower performance than the proposed PLP architecture, and specifically leads to top-1 accuracy of 38.62% and 70.92% with MIT and Dartmouth datasets, respectively. Since the MFL model does not use information of PCMs to predict where person p stays at time t, the temporal feature of PCMs does not affect the performance of MFL model.

Now, the performance of PLP model is analyzed with different PCM selection methods. In EFB and CISB methods, members with the most similar movement paths are selected as PCMs, i.e., people who stay the most frequently with a given person. The main difference between these two methods is that the EFB scheme weights the link of two people based on direct encounters between them, whereas the CISB method compares community interactions between persons to estimate their social correlation. Even though in some cases the CISB method obtains slightly higher accuracy than the EFB method, the two generally achieve similar performance. This indicates that a social relationship between two persons can be mainly reflected by direct interactions between them.

In contrast, the BSB method aims at selecting PCMs with the most correlated behavioral patterns. Our work shows that discovering behavior patterns helps to significantly improve prediction accuracy of the model compared to the friendship-based method. In fact, for members who work in the same university or company, they tend to have similar lifestyles (e.g., behavioral pattern) owing to the common schedules of the workplace. Therefore,



(a)



(b)

Figure 3-6: Effects of temporal location features on the prediction model in case of (a) the MIT dataset and (b) Dartmouth dataset

considering behavioral similarity between people can be helpful in selecting better PCMs for predicting people's locations. Specifically, when embedding pre-cur-loc information of PCMs selected by the BSB scheme into the PLP model, we achieve top-1 prediction accuracy as high as 71.00% and 92.31%, respectively, with MIT and Dartmouth datasets.

In the SC scheme, regardless of encounters between 2 people, members who have similar spatial distribution, or longevity of visiting places are selected as PCMs. The SC method may choose PCMs who have different frequency or regularity of staying in locations with the person of interest. For example, if 2 people p and q spend 4 hours at the library everyday, the SC method is likely to choose person q as a PCM of p. However, if person p stays at the library in the whole afternoon while q spends 2 hours in the morning and 2 hours in the evening, using location information of q may not be appropriate to predict the mobility of person p.

Meanwhile, in case of the STC method, two people have a high closeness score if they move in a synchronous manner or their movement are highly dependent. Note that PCMs in both STC and BSB tend to move in a synchronous way with the person of interest. However, a fundamental difference between BSB and STC is that the STC method does not consider the temporal information when measuring the similarity between two people. Specifically, the STC approach computes the spatial distribution of person p given the location of person q. Meanwhile, the BSB method investigates the spatial distribution of person p given both spatial and temporal information of person q. By considering the temporal data when selecting PCMs, the BSB approach is able to measure movement association with both spatial and temporal aspects. As a consequence, the BSB method achieves significantly higher prediction accuracy than the STC one with both datasets.

However, the performance gap between BSB and STC in the Dartmouth traces is noticeably smaller than that in the MIT one. The gap difference may come from different characteristics of human mobility in 2 datasets. Recall that the STC scheme does not consider the dependence between human movement and the temporal information. Therefore, if human mobility extracted from a dataset (e.g., the Dartmouth) is not highly related to the temporal data, the STC can achieve better accuracy than the case of strong relationship between human movement and the temporal information.

One should note that BSB, SC and STC approaches do not take into account the actual encounter or friendship between people. The experiment results in Fig. 3-6 indicate that friendship-based methods (EFB and CISB) causes the lower performance than other

approaches (BSB, SC, and STC) in which friendship is not considered. In addition, the proposed BSB approach achieves the most accurate prediction among the considered PCMs selection methods. For example, using location data of PCMs at time $(t-1)$ and t, the PLP model predicts human locations at time t with top-1 accuracy of 81.64%, 82.57%, 85.92%, 82.07%, and 92.31% using EFB, SC, STC, CISB, and BSB methods, respectively, in case of the Dartmouth dataset.

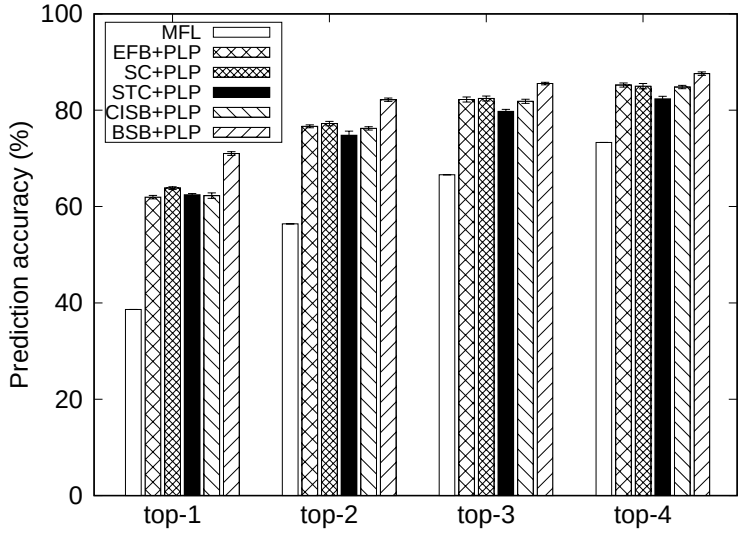
Also, from the observations in Fig. 3-6, among three considered temporal features, pre-cur-loc and cur-loc lead to quite similar performance in both datasets. For example, in case of the MIT traces, the PLP model achieves 71.13% and 71.00% accuracy in predicting the location of person p given pre-cur-loc and cur-loc information of PCMs, respectively. As anticipated, using data of PCMs during time t (cur-loc) can result in a more accurate prediction than the use of $(t-1)$ (pre-loc).

3.5.2 Top-k Accuracy

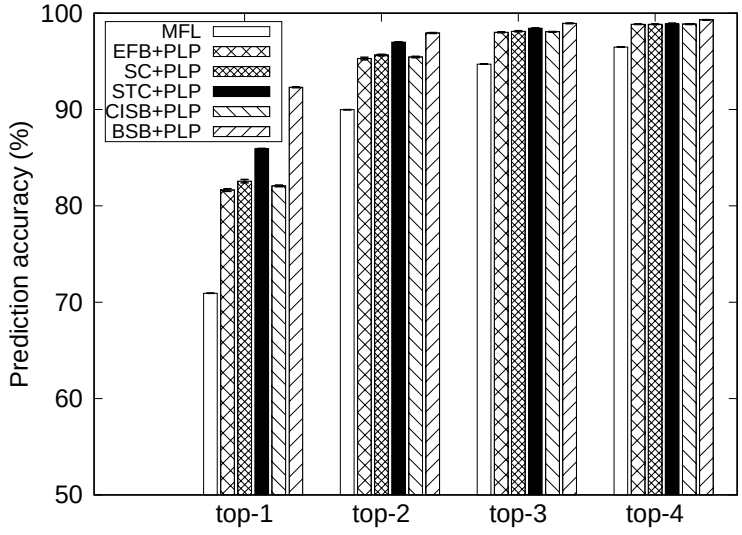
In this subsection, we evaluate the top-k accuracy of the proposed PLP model when $k = \{1, 2, 3, 4\}$. Other parameters are set to the default values shown in Table 3.5. For top-k accuracy, the model outputs a list of k location(s) with the highest probability. If the correct location belongs to the list of k element(s), the prediction is considered accurate.

As expected, Fig. 3-7 shows that the obtained accuracy increases from top-1 to top-4. For example, in case of the MIT dataset, the BSB approach results in 71.00%, 82.17%, 85.53%, and 87.59% accuracy corresponding to top-1, top-2, top-3, and top-4, respectively. In all examined PCM selection schemes, the steepest rising rate is observed when k changes from 1 to 2. Then, the outcomes tend to plateau.

As can also be observed from Fig. 3-7, there is higher prediction accuracy in the Dartmouth dataset than the MIT one. This observation is attributed to the fact that human movement in the MIT dataset was collected within a wider area than that in the Dartmouth traces. More specifically, all APs in the Dartmouth dataset are located in the university campus while movement trajectories of people in the MIT dataset are not bounded in the school area. Therefore, people mobility in the MIT dataset tends to be more divergent and more difficult to predict than that in the Dartmouth traces.



(a)



(b)

Figure 3-7: Top-k accuracy of the prediction model in case of (a) the MIT dataset and (b) Dartmouth dataset

3.5.3 Effects of the Number of PCMs

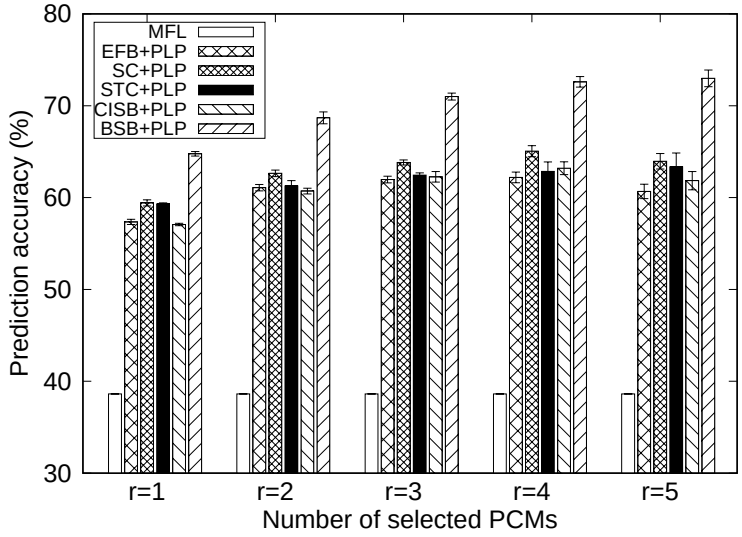
In this subsection, the number of selected PCMs varies from 1 to 5. As seen in Fig. 3-8(b), using location information of one PCM (i.e., $r = 1$), the PLP model obtains top-1 accuracy of 86.53% with the BSB method compared with 70.92% in case of the MFL model. Meanwhile, the PLP model with EFB, SC, STC, and CISB approaches results in 76.58%, 76.94%, 79.57%, and 76.70% accuracy, respectively.

It should be emphasized that when adding more PCMs, the performance of the prediction model is generally enhanced. In friendship-based methods (CISB and EFB), this is attributed to the fact that a person usually interacts with multiple PCMs rather than one person during the day. For example, in the morning, person p encounters co-worker q at the workplace. At lunch, p and r have an appointment at their favorite restaurant. Then, persons p and s go to the fitness center to exercise in the afternoon. In other PCMs selection approaches, the performance gain with more PCMs is due to the fact that a person has movement correlation with various PCMs depending on the time of a day. For instance, assume that person p is a middle-aged man. He may have a similar behavioral pattern with younger co-worker q during the daytime, whereas after work, he may have behaviors similar to other middle-aged men, rather than with the younger co-workers.

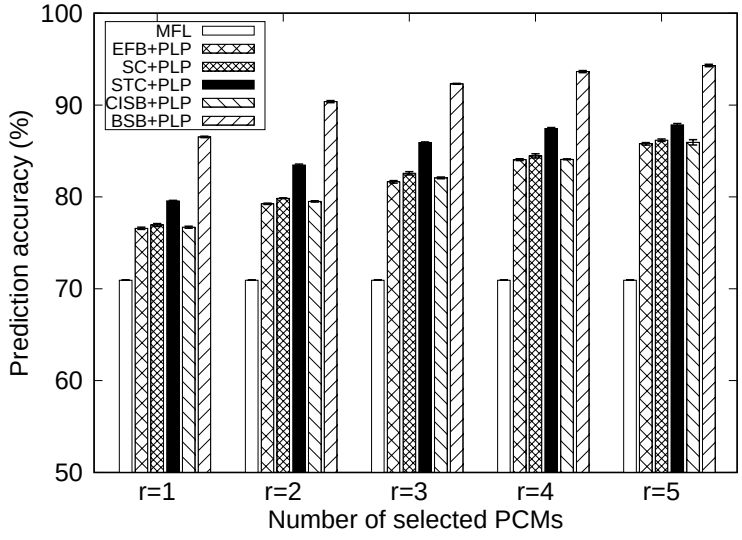
Moreover, the performance gap between the BSB and friendship-based methods (CISB and EFB) becomes larger as the number of PCMs embedded in the model increases especially in the MIT dataset. This is because selected PCMs of CISB and EFB tend to have more similar encounter patterns with person p than under the BSB method. To show this, KL divergence [93] is considered to measure similarity between temporal encounter distributions of PCMs and person p. Vector $\boldsymbol{q}$ denotes the probability distribution of encounters between persons p and q. The i^{th} element of $\boldsymbol{q}$ represents the probability that persons p and q encounter each other during time slot i. Vector $\boldsymbol{r}$ indicates a distribution of encounters between persons p and r. The length of $\boldsymbol{r}$ and $\boldsymbol{q}$ equals the number of samples for each person. Using KL divergence, the difference between the two probability distributions is given by:

$$\phi_{\boldsymbol{q},\boldsymbol{r}} = \sum_i q_i \log \frac{q_i}{r_i} \tag{3.12}$$

where q_i and r_i are the i^{th} elements of vectors $\boldsymbol{q}$ and $\boldsymbol{r}$, respectively. If the score from KL divergence is small, encounter distributions between PCMs and person p are close, which



(a)



(b)

Figure 3-8: Effects of the number of PCMs on the performance of the prediction model in case of (a) the MIT dataset and (b) Dartmouth dataset

indicates similar movement patterns between PCMs.

Assume that a mobility prediction model with r PCMs is considered. If another PCM with highly similar movement patterns to those of the r existing PCMs is added at input, the location information of the added PCM would not be helpful in predicting location of person p. As shown in Table 3.7, in cases where the number of selected PCMs is 5, the average value of KL divergence between PCMs' patterns of encounters with a person of interest is 11.01, 11.03, and 15.19 in EFB, CISB, and BSB methods, respectively. We can also observe from Table 3.7 that KL divergence values of SC and STC approaches are generally higher than friendship-based ones since SC and STC do not consider encounters between people.

Table 3.7: Comparison of KL divergence values between PCMs selection methods

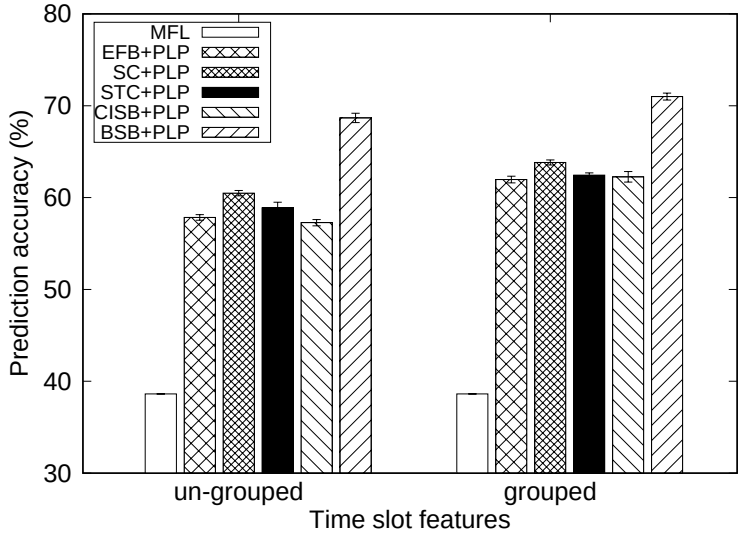
Dataset	**EFB**	**SC**	**STC**	**CISB**	**BSB**
MIT	11.01	11.66	15.07	11.03	15.19
Dartmouth	13.59	13.63	20.32	13.76	20.44

Lower KL divergence values from EFB and CISB methods indicate that they tend to select PCMs with more similar mobility patterns than the BSB method, and accordingly, movement patterns of the selected PCMs under the BSB method are more divergent than those under EFB/CISB methods. This contributes to a large accuracy gap between the BSB and friendship-based methods. The results also agree with the assumption that correlated locations in the BSB method do not have to be the same places.

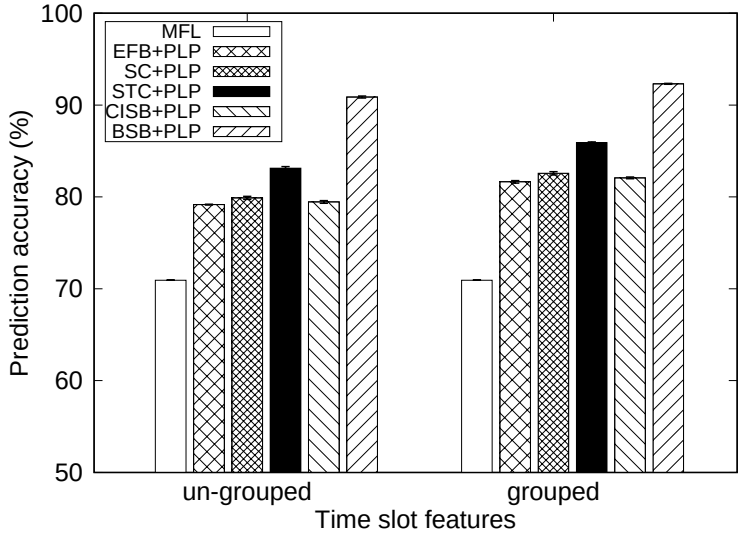
3.5.4 Effects of Time Slot Feature Selection

Figure 3-9 compares the precision of the mobility prediction model with two different time slot features: un-grouped and grouped. For the un-grouped case, 32 time slots in a day are mapped into a 32-dimensional vector. Meanwhile, for the grouped case, one-hot encoding converts a time slot index into a vector in $\mathbb{B}^3$ where each dimension indicates one of the three parts of a day.

In addition, as shown in Fig. 3-9, it is clear that grouping time slot feature achieves higher prediction accuracy than with un-grouped features. The results reflect that people tend to spend time with each other for a relatively long period. This becomes more understandable in environments like university campuses or companies, where people usually stay with their colleagues for a significant amount of time during the day. After work, they may enjoy leisure activities with friends in the evening. Another reason for the low performance with the un-grouped case may be the small number of data samples compared with the large



(a)



(b)

Figure 3-9: Comparison of time slot features in the human prediction framework in case of (a) the MIT dataset and (b) Dartmouth dataset

input dimensions in the un-grouped case.

3.6 Chapter Summary

Human mobility prediction has been a key point for the success of a variety of potential applications. For instance, with the development of device-to-device communication technologies, being able to accurately predict human locations will facilitate the design of an efficient data routing protocol in opportunistic networks. Moreover, the location prediction model inspires further applications such as urban data mining, location-based recommendation services, and contagious disease control. In addition, location estimation is required even when someone may not be willing to share their locations (e.g., geographic profiling of criminals).

Therefore, in this chapter, we address the mobility prediction problem in which the current or next location of a person is estimated, especially without requiring the positional history of that person. Since the movement of a person is highly related to other people, a two-phase framework is proposed in which persons with correlated movements (PCMs) with the person of interest are determined in the first phase. Then, information of these selected PCMs is leveraged in the second phase to estimate the location of the person of interest. For selecting PCMs, the communication interaction similarity-based (CISB) method considers encounter interactions between people, whereas the behavioral similarity-based (BSB) scheme selects PCMs who have similar behavioral patterns, instead of considering co-location between people.

For the considered problem, our approach is robust, because it can reduce overfitting by only using the location information of the selected PCMs to estimate the positions of the person of interest. Low time and space computational complexity is also achieved. Furthermore, geographic locations are not required in our model. Experimental results show that the BSB method gains significant improvement over other PCMs extraction approaches. In addition, the performance generally increases when more PCMs were embedded into the designed prediction model. In particular, a large number of PCMs is more beneficial under the BSB method than others.

However, our PCMs selection methods are centralized approaches because of requiring mobility traces of all people. Therefore, the proposed PCMs extraction methods are restricted to small or medium networks. In addition, information assurance needs to be taken

into account since mobility data of people is exchanged in the network.

Chapter 4

Low Cost Decentralized Human Encounter Prediction Model

4.1 Preliminaries

4.1.1 Dataset

In this work, two datasets of Wi-Fi traces are used to examine the proposed model. The first dataset is the Buffalo/phonelab-wifi logs collected over 5 months from smart phones carried by a group of 284 people, who are faculty members, staffs, and students at University at Buffalo (UB) [20, 94]. The second dataset provides four-month Wi-Fi logs of 13,888 mobile device carriers on the Dartmouth college campus [19]. Whenever a phone connects to a nearby AP, information including time-stamp, device ID, and basic service set identifier (BSSID) was recorded.

People usually carry their phones most of the time and Wi-Fi networks have short ranges of tens of meters. Therefore, human mobility can be represented by a sequence of associated (or scanned) APs which are identified by BSSID [95]. Specifically, in the first dataset, we use the sub-dataset named WifiScanResult which contains Wi-Fi scan logs of 274 anonymous mobile users and 1,193,746 APs. In the second dataset, the Wi-Fi logs including the AP association information of 13,888 mobile device users and 623 APs were used.

In order to extract significant human movement, the most active period in a day, that is, from 9 am to 6 pm, is considered. Moreover, in the UB dataset, we select the 129 most active users who have sufficient Wi-Fi scans and 5,251 APs, which were scanned more than 20,000 times. In addition, the experiment duration was trimmed to 65 days (from January

26th to March 31st, 2015), in which we can observe the most Wi-Fi activities of all mobile users as well as their interactions. Similarly, in the Dartmouth dataset, the 162 most active users and the 118-day period from January 3rd to April 30th, 2004 are extracted. Table 4.1 summarizes the main features of the two datasets.

Table 4.1: Summary of two datasets after user and AP extraction

Dataset	**UB**	**Dartmouth**
Number of people	129	162
Number of APs	5,251	623
Duration (Days)	65	118
Period in a day	9h to 18h	9h to 18h

4.1.2 Encounter Extraction

Since the purpose of this study is to construct an encounter prediction model, the contact information between mobile carriers is extracted using Wi-Fi logs. In particular, time-sliced user position indication data (TSUPID), which represents the temporal location of all users, is created first. Then, we build the meeting table, which contains the time-sliced encounters between individuals. Note that if two people concurrently associate with the same AP deployed at some place, e.g., a class room or cafeteria, they stay in proximity to each other and are likely to have an encounter. This definition of indirect encounter was also introduced in other existing studies [4, 63, 96, 97]. Moreover, the event of indirect encounter between 2 people implies either a direct encounter or they are two-hop neighbors of each other. That means a communication path exists between these 2 people. Therefore, in this work, two individuals are considered to have contact during a time slot if their devices associate with the same AP in that time slot.

Now, we describe how to create TSUPID for each person based on the Wi-Fi association (or scan) traces of that person. Recall that the movement of a person in a given period can be represented by the sequence of associated APs in that period. Specifically, we consider

Table 4.2: TSUPID of person p

Person ID	Day Index	Time Slot Index	Start Time of Time Slot	Number of Associated APs	Set of Associated APs
p	1	1	9:00	3	{17, 256, 300 }
p	1	2	9:15	0	{ $\emptyset$ }
...	...	...	...	...	...
p	1	n_{TS}	17:45	2	{118, 5246}
p	2	1	9:00	1	{5246}
p	2	2	9:15	2	{1, 1299}
...	...	...	...	...	...
p	n_D	n_{TS}	17:45	4	{71, 3527, 4630, 5110}

Table 4.3: Meeting table

	$p_1d_1s_1$	$p_1d_1s_2$	...	$p_1d_{n_D}s_{n_{TS}}$	$p_2d_1s_1$	...	$p_2d_{n_D}s_{n_{TS}}$	...	$p_md_1s_1$	...	$p_md_{n_D}s_{n_{TS}}$
p_1	0	0	...	0	1	...	0	...	1	...	1
p_2	1	1	...	0	0	...	0	...	0	...	1
...	...	...	...	...	...	...	...	...	...	...	...
p_m	1	0	...	1	0	...	1	...	0	...	0

the nine-hour duration from 9 am to 6 pm in a day and divided this duration into 15-min time intervals. Then, a list of associated APs during a 15-minute time slot can be used to reflect human mobility in that slot.

Table 4.2 shows the example of person p's TSUPID, which contains six fields: person ID, day, time slot, start time of time slot, number of APs, and the set of associated APs. For instance, at time slot 1 of day 1, person p associates with a total of three APs (i.e., 17, 256, and 300), while at time slot 2 of day 1, that person associates with no AP. We define m, n_D, and n_{TS} as the number of people, number of days in the experiment period, and number of time slots per day, respectively. In the UB dataset, $m = 129, n_D = 65$, and $n_{TS} = 36$.

Then, given TSUPID of m people, a meeting table is constructed which accounts for the temporal contacts of all people throughout the duration of the experiment period. The meeting table is composed of m rows corresponding to the encounter information of m individuals, and $(m \times n_D \times n_{TS})$ columns, where each column indicates one specific time slot of a day for a person. Specifically, row p of the table represents the time-sliced encounter information between person p and other members. During a specific time slot, if the intersection of associated AP sets of two people contains at least one element, there is considered to be an encounter between them, and the value corresponding to that specific time slot is set to 1. Otherwise, this value is set to 0. Note that a person is considered not to meet himself/herself during the experiment period.

Recall that we aim at predicting indirect encounters, which are inferred from Wi-Fi traces. In the case when a person wants to extract encounter information, access point AP_i, with which that person is concurrently associated, should send a list of persons who are connecting to AP_i. As a result, the problem of privacy leakage can happen. In order to mitigate the privacy issue, people are assumed to be willing to share their association with members who are connected to the same Wi-Fi access point. Moreover, an extra privacy method for Wi-Fi networks should be used, e.g., access point AP_i needs to first encrypt and then send the information about the list of associated people.

Note that encounters can be obtained by using traces from device-to-device communica-

tion networks (e.g., Bluetooth, Wi-Fi direct). In this case, privacy can be better preserved since people communicate directly to each other. However, due to the lack of a large Bluetooth dataset, in this work the Wi-Fi traces are used to infer encounters between people.

4.1.3 The AP Embedding Model

Recall that historical mobility information, as represented by a sequence of associated APs, is used to predict future encounters between people. Therefore, it is necessary to encode the AP indices. Let n_{AP} denote the number of APs. If we apply the one-hot encoding technique, an AP index is mapped to an indicator vector in $\mathbb{B}^{n_{AP}}$. In the case that we process with a large number of APs, e.g., the UB dataset, a fundamental problem is raised: the curse of dimensionality. Specifically, a large number of parameters need to be fitted, thus resulting in a substantially long training process. Therefore, in this section, we propose an embedding model with which to learn low-dimensional representations of the APs which reflect geographical closeness of APs, i.e., two geographically near APs should be represented by vectors with a short distance in the vector space.

The proposed embedding model is inspired by the task of learning a distributed representation for words in natural language processing. Specifically, we consider the Skip-gram architecture [98,99], which shows improved estimation of word representations. In word embedding, given a current word, a Skip-gram model is trained to predict surrounding words, which are within a certain window centered at the current word in the same sentence. During the training process, pairs of current and surrounding words are fed into the model. Consider the following sentence as an example: *a dog is walking in the garden.* Assume that the window size is set to 1, then, the pairs of (current, target) words include (dog, a), (dog, is), and (is, walking). If the number of sentences in the word training dataset is sufficiently large, other words pairs of semantic and syntactic similarities can be found, such as (cat, a), (cat, is), and (was, walking). As a result, after training the Skip-gram model, semantically and syntactically neighboring words (e.g., dog and cat; is and was), which have the similar surrounding words, will be encoded into close embedding vectors.

We construct the AP embedding model which predicts the one-hop neighbor AP given a current AP. Let AP_i denote the i^{th} AP ($1 \leq i \leq n_{AP}$) and v^i_{AP} denote the representative AP vector of AP_i. In this work, an AP_i is said to be a one-hop neighbor of an AP_j if one person stays in the communication range of both AP_i and AP_j at an arbitrary time. Two APs (e.g., AP_i and AP_k) which have the same one-hop neighbor (e.g., AP_j) are likely to be

near each other geographically. Therefore, taking the Skip-gram model as an inspiration, the AP embedding model is trained such that APs (e.g., AP_i and AP_k) with the same one-hop neighbor are represented by close vectors in the vector space.

The Skip-gram based AP embedding model consists of three layers: input, hidden, and output. The input and output layers have n_{AP} units (e.g., $n_{AP} = 5,251$ in the UB dataset), while the number of hidden units equals the size of embedding vector. There is no activation at the hidden layer, whereas the sigmoid units are used at the output layer. Note that there are 5,251 output units corresponding to the outcomes of n_{AP} binary classifiers. The weight matrix connecting the input and hidden layers contains the embedding vectors of n_{AP} considered APs. Let n' denote the length of the AP representation vector ($n' < n_{AP}$). The embedding model outputs an $n_{AP} \times n'$ weight matrix, where row i accounts for the AP_i embedding vector, v^i_{AP}. The length of the representative AP vector is reduced from n_{AP} to n', where n' is set to 12.

The training samples of the embedding model are (AP_i, AP_j) where AP_j is a one-hop neighbor of AP_i. The objective of the embedding model is to maximize the probability $P(AP_j|AP_i)$ or minimize the loss function $-\log P(AP_j|AP_i)$. The embedding model is trained such that neighbor APs are encoded into close vectors with a small distance. In addition, non-neighbor APs are embedded into distant points in the vector space.

In order to accelerate the training process, negative sampling [98] is used, in which we only adjust weights that connect to the neighbor unit AP_j (a positive sample) and a few non-neighbors of AP_i (negative samples), as opposed to updating all of the model parameters every epoch. In this work, the model parameters are adjusted every epoch by using a positive sample and two negative samples which are randomly selected with a uniform distribution.

4.2 The Human Encounter Prediction Model

Assume that there are m people in the network, and that at every time slot, each person records contact data consisting of a list of encountered members, a rendezvous place, and the contact time. Given the recorded encounter data of a specific person p, our work aims at predicting whom person p is likely to meet after k time slots ($k \geq 1$). In the upcoming subsections, the proposed distributed human encounter prediction (DHEP) models based on RNN and FFNN are presented and compared to the centralized human encounter prediction (CHEP) models. For simplicity, DHEP/RNN and DHEP/FFNN are used hereafter to denote

the DHEP models based on RNN and FFNN, respectively. Similarly, we use CHEP/RNN and CHEP/FFNN to represent the CHEP models based on RNN and FFNN, respectively.

4.2.1 DHEP/RNN Model

Let $\boldsymbol{x}_{p,t}$ denote the encounter information of person p at current time slot t. In our work, $\boldsymbol{x}_{p,t} = \{\boldsymbol{v}_{p,t}^{Loc}, \boldsymbol{v}_{p,t}^{Con}, \boldsymbol{v}_{p,t}^{TS}\}$, where vectors $\boldsymbol{v}_{p,t}^{Loc}$, $\boldsymbol{v}_{p,t}^{Con}$, and $\boldsymbol{v}_{p,t}^{TS}$ represent the location of person p, his/her contacts, and the time slot index at t, respectively. Recall that n' and n_{TS}, respectively, are the size of the AP embedding vector and the number of time slots in a day. If person p only has an association with AP_i at time slot t, the AP_i embedding vector is used to reflect the meeting place, i.e., $\boldsymbol{v}_{p,t}^{Loc} = v_{AP}^{i}$. In the case that person p associates with j APs (e.g., $AP_1, AP_2, ..., AP_j$ with $j \geq 2$), a vector which represents j-associated APs is used to reflect the location of that person. Recall that $AP_1, AP_2, ..., AP_j$ are encoded into embedding vectors $v_{AP}^{1}, v_{AP}^{2}, ..., v_{AP}^{j}$, respectively. The median is a common metric used to measure the properties of a set, particularly when the elements of a set exhibit skewed distribution. Therefore, the median of two-dimensional array $[v_{AP}^{1}, v_{AP}^{2}, ..., v_{AP}^{j}]^{\top}$ along the first axis is computed and used as the location representation of person p. Vector $\boldsymbol{v}_{p,t}^{Con}$ contains m elements, in which the q^{th} element denoted by $\boldsymbol{v}_{p,t}^{Con,q}$ is set to 1 if person p encounters q at time t, and otherwise this element is 0. $\boldsymbol{v}_{p,t}^{TS}$ with n_{TS} elements is a vector of 0s, but the element corresponding to time slot t gets activated.

In the proposed model, the sequential contact data during the past r time slots $\{\boldsymbol{x}_{p,t-r+1}, \boldsymbol{x}_{p,t-r+2}, ..., \boldsymbol{x}_{p,t}\}$ is driven into the prediction model, which outputs the probability that person p will meet $(m-1)$ other people at time slot $t+k$. More specifically, for person p, a generalization model is constructed such that the future encounter probability between p and q is maximized given his/her own historical contact information:

$$P\left(\boldsymbol{v}_{p,t+k}^{Con,q} | \{\boldsymbol{x}_{p,t-r+1}, \boldsymbol{x}_{p,t-r+2}, ..., \boldsymbol{x}_{p,t}\}\right), 1 \leq q \leq m \tag{4.1}$$

If the computed encounter probability between persons p and q exceeds a decision threshold value (e.g., 0.5), then p and q are predicted to have contact with each other at time slot $t+k$.

Since recurrent neural networks (RNNs) have shown potentially high performance when modeling sequential data, we propose an RNN-based encounter prediction model for estimating future contacts between people. In particular, our work considers long short-term

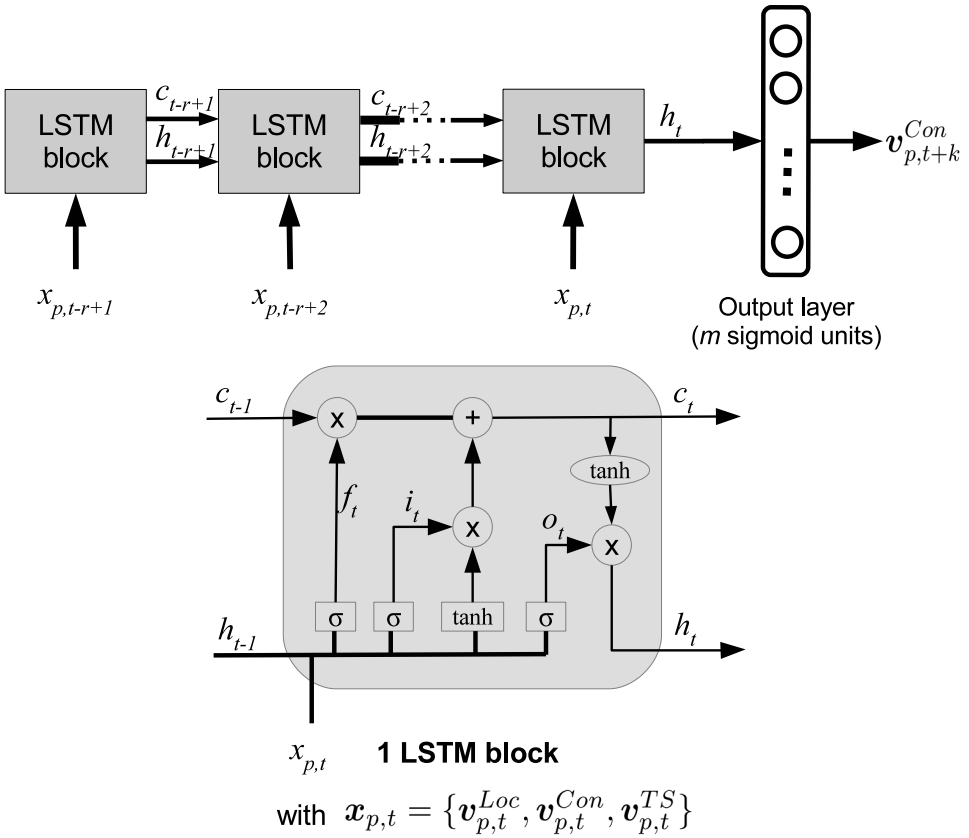


Figure 4-1: The RNN-based distributed human encounter prediction model of person p

memory (LSTM) [100], which is a specific recurrent neural network architecture, to better capture the dependencies of human encounters. As shown in Fig. 4-1, an unrolled LSTM contains a number of memory blocks, where each block stores the temporal cell state of the network, which depends on the previous cell state and current input. In order to control the information flow in the network, three gates are used: input, output, and forget gates. Precisely, the previous cell state is scaled by the forget gate prior to being added to the current cell state. The input gate adjusts to what degree the current input activations contribute to the cell state of the current block. Meanwhile, the flow of cell activations into the network is determined by the output gate.

The architecture of one LSTM block is described in detail. The current encounter data $\boldsymbol{x}_{p,t}$ at time t is fed into the LSTM block. Let $\boldsymbol{c}_t$ and $\boldsymbol{h}_t$ denote the cell state and hidden cell state at time slot t, respectively. Further, let $\boldsymbol{f}_t, \boldsymbol{i}_t, \boldsymbol{o}_t$ denote the outcomes of forget, input, and output gates, respectively. In each LSTM block, the current cell state is computed based on the input and previous cell states as follows:

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{xf}^\top \boldsymbol{x}_{p,t} + \boldsymbol{W}_{hf}^\top \boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \tag{4.2}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{xi}^\top \boldsymbol{x}_{p,t} + \boldsymbol{W}_{hi}^\top \boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \tag{4.3}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{xo}^\top \boldsymbol{x}_{p,t} + \boldsymbol{W}_{ho}^\top \boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \tag{4.4}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tanh(\boldsymbol{W}_{xc}^\top \boldsymbol{x}_{p,t} + \boldsymbol{W}_{hc}^\top \boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \tag{4.5}$$

$$\boldsymbol{h}_t = \tanh(\boldsymbol{c}_t) \odot \boldsymbol{o}_t \tag{4.6}$$

where $\boldsymbol{W}_{xf}, \boldsymbol{W}_{hf}, \boldsymbol{W}_{xi}, \boldsymbol{W}_{hi}, \boldsymbol{W}_{xc}, \boldsymbol{W}_{hc}, \boldsymbol{W}_{xo}, \boldsymbol{W}_{ho}$ denote weight matrices and $\boldsymbol{b}_f, \boldsymbol{b}_i, \boldsymbol{b}_c, \boldsymbol{b}_o$ represent bias vectors.

The outcome of LSTM block $\boldsymbol{h}_t$ is mapped to the output of prediction model $\boldsymbol{v}_{p,t+k}^{Con}$ by a fully connected sigmoid output layer. There are m output units corresponding to the probability of encounter between person p and m people at time $t+k$. The future contacts, $\boldsymbol{v}_{p,t+k}^{Con}$, are predicted as follows:

$$\boldsymbol{v}_{p,t+k}^{Con} = \sigma(\boldsymbol{W}_{hy}^\top \boldsymbol{h}_t + \boldsymbol{b}_h) \tag{4.7}$$

Let $\boldsymbol{y}$ and $\boldsymbol{t}$ denote output $\boldsymbol{v}_{p,t+k}^{Con}$ of the prediction model and the true label, respectively. The number of samples is defined as η. The cross-entropy cost function $J(\theta)$ is used to measure the difference between the estimated and correct labels as follows:

$$J(\theta) = -\frac{1}{\eta} \sum_{j=1}^{\eta} \left(\boldsymbol{t}^{(j)\top} \log \boldsymbol{y}^{(j)} + (1 - \boldsymbol{t}^{(j)})^\top \log (1 - \boldsymbol{y}^{(j)}) \right) \tag{4.8}$$

where $\boldsymbol{y}^{(j)}$ and $\boldsymbol{t}^{(j)}$ are the estimated output and target vectors of the j^{th} instance, respectively. The adaptive moments algorithm [101] is used to minimize the cost function and to train the parameters (i.e., weight matrices and bias vectors) of the model. In addition, in order to construct a generalization model that fits well for not only the training dataset but also the unobserved data, the dropout regularization technique [102] is applied, since dropout showed the substantially reduced overfitting on a variety of tasks.

4.2.2 DHEP/FFNN Model

Although RNNs can capture information of sequential data, learning the RNN-based prediction model demands high computation capability due to a large number of training parameters. Moreover, since people typically carry mobile devices with constrained resources, constructing the RNN-based prediction model can cause the out of memory problem. Therefore, for devices with limited computation capability, we also propose another DHEP model based on a feed-forward neural network (FFNN), which requires fewer resources than the RNN-based model.

The objective of the DHEP/FFNN model is either to predict whether persons p and q are in physical proximity to each other, given person p's current mobility information $\boldsymbol{x}_{p,t}$, or to maximize the following probability:

$$P(\boldsymbol{v}_{p,t+k}^{Con,q}|\boldsymbol{x}_{p,t}), \quad 1 \leq q \leq m \tag{4.9}$$

As shown in Fig. 4-2, the proposed FFNN-based model consists of an input layer, followed by two hidden layers with tanh activation units. The logistic output layer has m units which correspond to the encounter probabilities between person p and m people in the network. Two individuals p and q are said to be in close proximity to each other after k time slots if the value at output unit q exceeds a given decision threshold value (e.g., 0.5).

During the training process, the cross-entropy loss function, which indicates the error between the actual and predicted movements, is minimized by using the adaptive moment estimation [101]. Dropout regularization is applied to the outputs of two hidden layers in order to mitigate the problem in which the prediction model is over-fitted to the training dataset.

Next, we compare the complexity of DHEP/RNN and DHEP/FFNN prediction models. We define n_i and n_o as the number of input and output units of the prediction models, respectively. Let n_c denote the number of cell units in each LSTM block, i.e., $n_c = |\boldsymbol{c}_t| = |\boldsymbol{h}_t|$. In the FFNN-based model, n_{h1} and n_{h2} are defined as the number of activation units in hidden layers 1 and 2, respectively. Ignoring their biases, the number of parameters N_{RNN} in the DHEP/RNN model is computed as:

$$N_{RNN} = 4n_c^2 + 4n_i n_c + n_o n_c + 3n_c \tag{4.10}$$

Meanwhile, the proposed DHEP/FFNN is composed of n_{FFNN} weight parameters, and n_{FFNN} is calculated as:

$$N_{FFNN} = n_i n_{h1} + n_{h1} n_{h2} + n_{h2} n_o \tag{4.11}$$

For example, in the case of the UB dataset, $n_i = 178$ and $n_o = 129$. If $n_c = 900$ and $n_{h1} = n_{h2} = 256$, $N_{RNN} \approx 4$M and $n_{FFNN} \approx 144$k, which means that the RNN-based prediction model needs to train approximately 27 times as many parameters as the FFNN-based model does.

In spite of requiring substantially more computation resources, unlike the FFNN-based model, the RNN-based model can learn long-term temporal information. Due to the self-recurrent connections, the RNN-based model can store historical encounter data in the cell state. Therefore, the RNN-based prediction model is expected to achieve higher performance than the FFNN-based model, at the cost of large computation complexity.

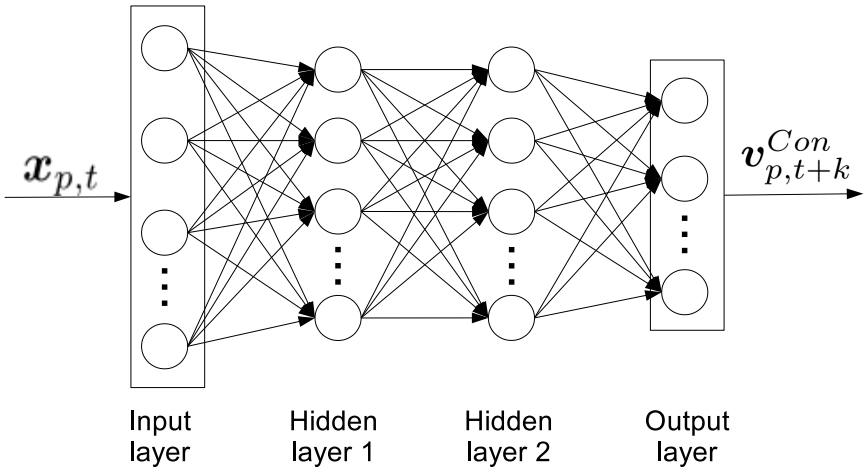


Figure 4-2: Architecture of the FFNN-based encounter prediction model of person p

4.2.3 The Centralized Human Encounter Prediction (CHEP) Models

Note that the distributed models, DHEP, may suffer from a lack of samples, since each person constructs a separate model by solely using his/her own encounter data. In this subsection, we consider the centralized human encounter prediction (CHEP) model which uses the contact data of all people to construct one predictor for the encounter prediction task. CHEP would be useful when all people's data is allowed to be collected.

The centralized model maps the recent encounter data of person p ($1 \leq p \leq m$) to the

future contacts between p and other members. Note that, since the mobility data of all m people is fed into the centralized model, vector $\boldsymbol{v}_p^I \in \mathbb{B}^m$, which indicates the person index, is appended to input $\boldsymbol{x}_{p,t}$. That means $\boldsymbol{x}_{p,t} = \{\boldsymbol{v}_p^I, \boldsymbol{v}_{p,t}^{Loc}, \boldsymbol{v}_{p,t}^{Con}, \boldsymbol{v}_{p,t}^{TS}\}$, where $\boldsymbol{v}_p^I \in \mathbb{B}^m$ is a vector of 0s except that the p^{th} element is set to 1 in order to distinguish person p from other members. The CHEP model outputs $\boldsymbol{v}_{p,t+k}^{Con}$, i.e., encounters of person p at time $t+k$. In the centralized prediction model, the network architecture, loss function, and training algorithm are similar to those of the distributed model.

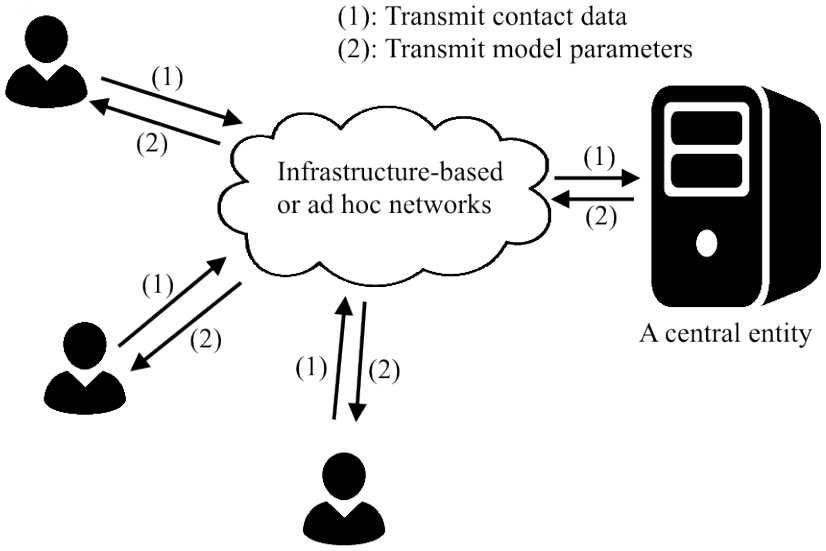


Figure 4-3: Data communication in the centralized encounter prediction model

Figure 4-3 summarizes the data exchange in the network. Each individual needs to send the historical contact information to a central entity via infrastructure-based or ad-hoc networks, as represented by (1) in Fig. 4-3. At the central entity, the CHEP model is trained using the aggregated encounter data. Then, the model parameters (i.e., weights matrices and biases vectors) are distributed to all people, as shown by (2) in Fig. 4-3. Finally, for each person, future encounters are estimated using the trained parameters. People keep transmitting their contact data so that the central entity can update the parameters of the CHEP model. Further, changes in parameters should be regularly noticed to all persons via networks.

The comparison between DHEP and CHEP models is now discussed. Since the centralized model requires encounter data of all people, connections between the central entity and all individuals in the network should be established. By contrast, the DHEP model is more suitable for opportunistic networks with intermittent connections between members in the

network. Because transmitting personal encounter data via networks can result in privacy leakage, information assurance techniques should be used in the case of the CHEP model.

Moreover, since the centralized model suffers from the scale and high overhead problem, it is restricted to the small or medium network. In order to better understand the high overhead problem of the CHEP model, we compute the amount of data exchanged in the network. Recall that each time slot person p transmits time-sliced encounter data, $\boldsymbol{x}_{p,t} = \{\boldsymbol{v}_p^I, \boldsymbol{v}_{p,t}^{Loc}, \boldsymbol{v}_{p,t}^{Con}, \boldsymbol{v}_{p,t}^{TS}\}$, to the central entity. In the UB dataset, $m = 129, n_{AP} = 5251, n_{TS} = 36$, and the number of model parameters $n_{RNN} = 4$M. Assume that a person associates with only one AP during a time slot. If binary coding is used to represent encounter data $\boldsymbol{x}_{p,t}$, we need 8, 13, 129, and 6 bits (i.e., 156 bits in total) to indicate $\boldsymbol{v}_p^I, \boldsymbol{v}_{p,t}^{Loc}, \boldsymbol{v}_{p,t}^{Con}$, and $\boldsymbol{v}_{p,t}^{TS}$, respectively. Since there are 36 time slots in a day and 129 people, the amount of contact data that needs to be sent by people each day is $156\text{bits} \times 36 \times 129 = 90.5\text{kB}$. Additionally, the 4M parameters of the CHEP model should be informed to all people. Assume that four bytes are required to store each parameter value, then the central entity needs to send $4 \times 4 \times 129 = 2.06$ GB data to 129 people whenever model parameters are updated.

On the other hand, due to the fact that it only requires the contact information of the person of interest, the DHEP model can be applied to a network of arbitrary size. However, the centralized model is likely to capture the whole encounter interactions between individuals in the network. Therefore, the centralized model may show a performance improvement over the distributed model.

4.3 Performance Analysis

In this section, an evaluation of the proposed distributed human encounter prediction (DHEP) models is conducted. First, we describe the counterpart methods (i.e., Jahromi's model [63], Jyotish's model [64], and the DHEP model based on Naive Bayes (NB)) and the evaluation metrics in subsection 4.3.1. Then, the predictive performance of DHEP models is analyzed in subsection 4.3.2. Finally, the comparison with regard to performance and applicability between DHEP and CHEP models is discussed in subsection 4.3.3.

4.3.1 Experiment Setup

In this subsection, Jyotish's and Jahromi's models, which are motivated from the NB classification, are summarized. We also describe the NB-based DHEP model (DHEP/NB), which

is proposed to evaluate the predictability of the NB classifier with our input encounter features (i.e., $\boldsymbol{v}_{p,t}^{Con,q}$ and time t). Recall that $\boldsymbol{v}_{p,t}^{Con,q}$ denotes the contact information between p and q at time t. Jyotish's model assumes conditional dependence between the features of day d and time slot t. Accordingly, the probability that p meets q at time t on day d is calculated as follows:

$$P_{\text{Jyotish}}(\boldsymbol{v}_{p,t}^{Con,q}|d,t) = \frac{P(d,t|\boldsymbol{v}_{p,t}^{Con,q})P(\boldsymbol{v}_{p,t}^{Con,q})}{P(d,t)} \tag{4.12}$$

Meanwhile, Jahromi's model assigns weights w_d and w_t to features d and t, respectively. Then, given time slot t and day index d, the encounter probability between p and q is estimated below:

$$\begin{aligned} P_{\text{Jahromi}}(\boldsymbol{v}_{p,t}^{Con,q}|d,t) &= \frac{P(d,t|\boldsymbol{v}_{p,t}^{Con,q})P(v_{p,t}^{Con,q})}{P(d,t)} \\ &\propto P(d|\boldsymbol{v}_{p,t}^{Con,q})^{w_d} P(t|\boldsymbol{v}_{p,t}^{Con,q})^{w_t} P(\boldsymbol{v}_{p,t}^{Con,q}) \end{aligned} \tag{4.13}$$

We then present the proposed DHEP/NB model, which, given temporal contact data at current time t, predicts future encounters at time $t+k$. Inspired by the Naive Bayes classifier, DHEP/NB supposes that the current contacts and time slot t are conditionally independent. More specifically, person p is predicted to meet q after k time slots if the following probability exceeds a decision threshold value (e.g., 0.5):

$$\begin{aligned} P\big(\boldsymbol{v}_{p,t+k}^{Con,q}|t,\boldsymbol{v}_{p,t}^{Con,q}\big) &= \frac{P(t,\boldsymbol{v}_{p,t}^{Con,q}|\boldsymbol{v}_{p,t+k}^{Con,q})P(\boldsymbol{v}_{p,t+k}^{Con,q})}{P(\boldsymbol{v}_{p,t}^{Con,q},t)} \\ &\propto P(t|\boldsymbol{v}_{p,t+k}^{Con,q})P(\boldsymbol{v}_{p,t}^{Con,q}|\boldsymbol{v}_{p,t+k}^{Con,q})P(\boldsymbol{v}_{p,t+k}^{Con,q}) \end{aligned} \tag{4.14}$$

Now, we describe the evaluation metrics. This study evaluates the prediction models in terms of estimation accuracy. In addition, since each person usually contacts a small number of other people during a time slot, the presence of bias in favor of non-encounter cases can clearly be seen. Therefore, other performance metrics are also considered, including sensitivity, precision, F1 score, receiver operating characteristic (ROC), and area under curve (AUC). Sensitivity is defined as the number of accurately estimated contacts divided by the number of actual encounters. Precision indicates the ratio of correctly predicted samples to samples for which models estimate encounters that are likely to happen. Meanwhile, the performance of prediction models with varying decision threshold values can be seen in the

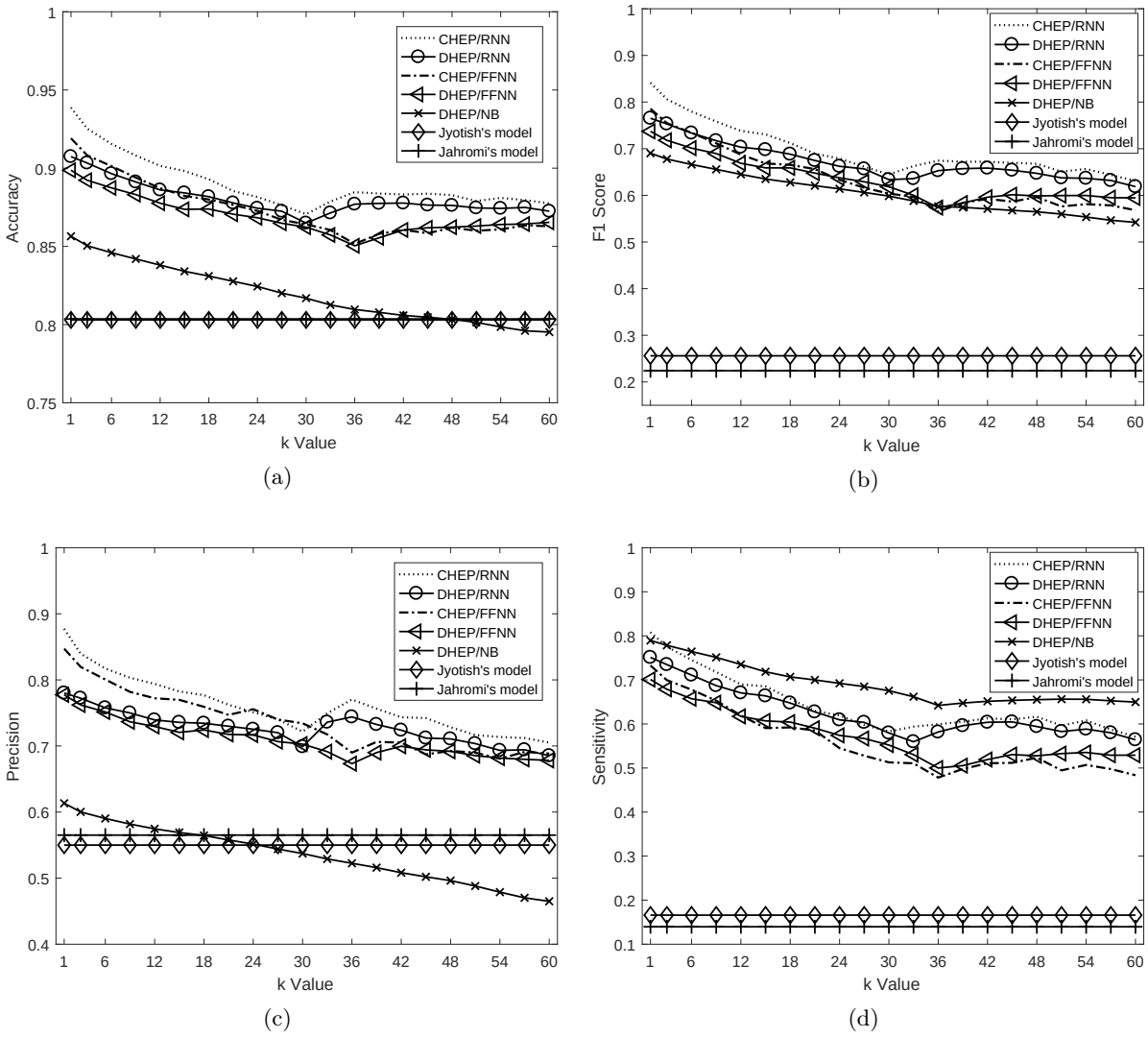


Figure 4-4: Performance of Encounter Prediction Models on the UB dataset: (a) Accuracy, (b) F1 Score, (c) Precision, (d) Sensitivity

ROC curve.

Our work collects the performance of prediction models on both UB and Dartmouth datasets. Each dataset is randomly divided into training, validation, and test sets at the ratio of 6:2:2. The experiment period lasts 65 and 118 days in UB and Dartmouth traces, respectively. The first two sets are used to fit the models and select the optimal hyper-parameters, respectively. Specifically, a variety of RNN and FFNN architectures were evaluated. Then, we selected the RNN-based model with one LSTM layer of 900 cell units and the FFNN-based model with 256 units in each of the two hidden layers. On the other hand, using the test set, the performance of models is measured by averaging the results of the 50 people with the most samples.

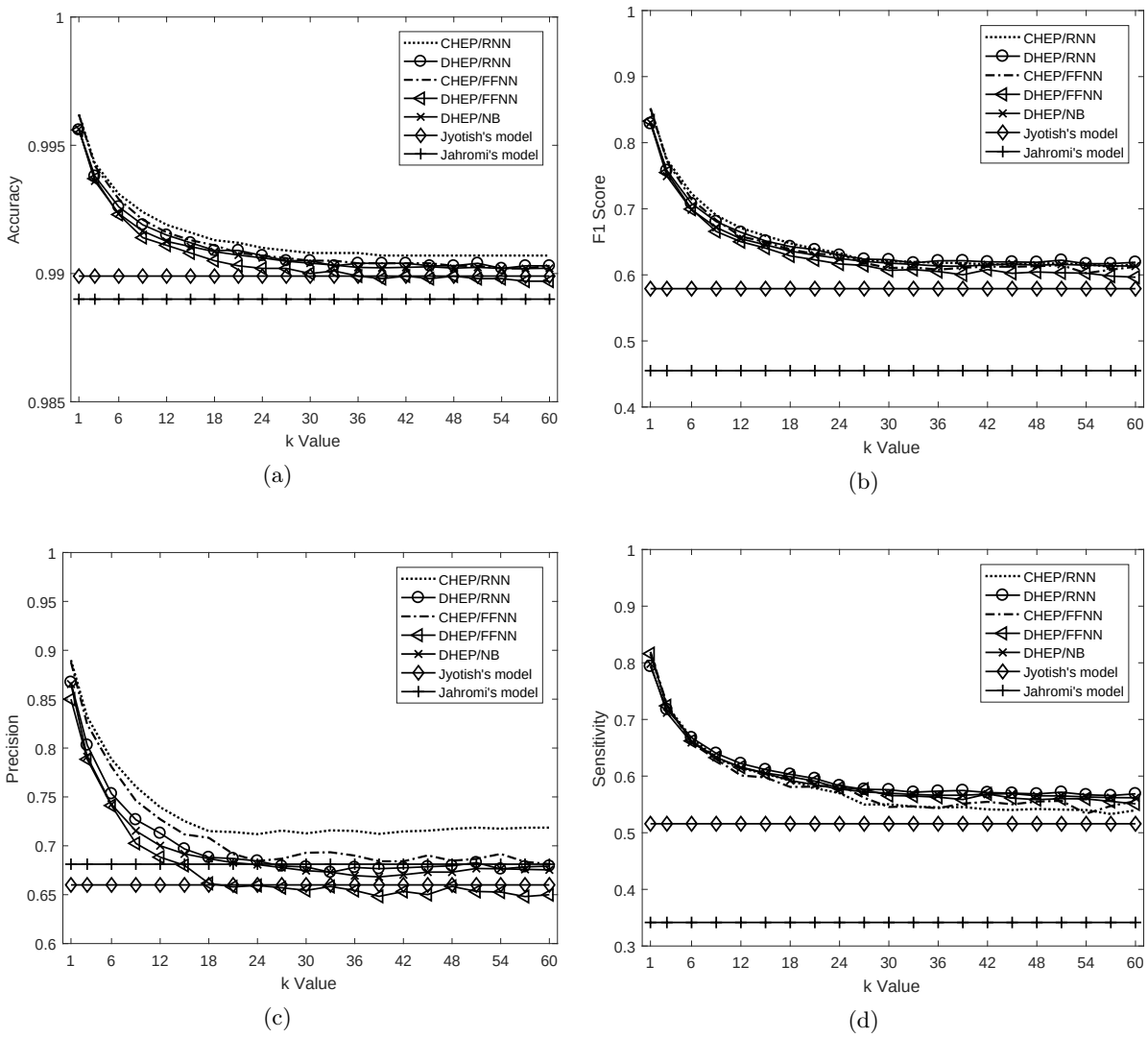


Figure 4-5: Performance of Encounter Prediction Models on the Dartmouth dataset: (a) Accuracy, (b) F1 Score, (c) Precision, (d) Sensitivity

4.3.2 Performance of DHEP Models

In the following subsection, we evaluate the proposed distributed models which predict encounters after k time slots. In this work, k is set between 1 and 60. Recall that we consider 15-minute time slots and a 9-hour period each day. As shown in Figures 4-4 and 4-5, the DHEP/RNN model outperforms DHEP/FFNN, DHEP/NB, Jahromi's, and Jyotish's models in both datasets. In the UB dataset, the DHEP/RNN model predicts people's contacts at the very next time slot (i.e., $k = 1$) with accuracy of 90.74% compared to the values of 89.33%, 85.65%, 80.36%, and 80.3 % in the FFNN-based, NB-based, Jahromi's, and Jyotish's models, respectively. Note that there is a difference in input features between the DHEP/RNN and DHEP/FFNN models. i.e., encounters information during the past r time slots (r is set to 6 and 3 in the UB and Dartmouth datasets, respectively) is fed to the DHEP/RNN model, while only the current encounter data is used for the prediction of future contacts in the FFNN-based model. Due to its ability to handle the temporal dependencies of LTSM, the RNN-based model can capture the human encounter pattern better than the FFNN-based model, thus gaining performance improvement over the other distributed models.

On the other hand, as defined in Equation (4.14), the DHEP/NB model assumes that the input features (i.e., current contacts and time slot t) are conditionally independent, given the output of future contacts at $t + k$. This assumption may not be realistic, as people are likely to follow regular daily movements. For example: person p usually takes a class in the morning on weekdays and has an encounter with classmates during class time. In the afternoon, person p often studies and meets some people at the library. In this example, contacts and time are strongly correlated. Therefore, the DHEP/NB model leads to lower performance than the RNN-based distributed model in both datasets. For example, as shown in Fig. 4-4(c), when $k = 1$, the NB-based model achieves 0.613 precision, i.e., 61.3% of predicted future encounters will be accurate, as compared to the 0.781 precision of the DHEP/RNN model, respectively.

In contrast to other models that leverage the historical human encounter information for the prediction of future contacts, Jahromi's and Jyotish's models only used the temporal context, which includes time slot and day indices. As a consequence, Jahromi's and Jyotish's models achieve the lowest evaluation results among the examined ones. This consequence indicates that using temporal context alone is not sufficient for a future encounter prediction model. Note that the performance of Jahromi's and Jyotish models is independent of the k

value, as their models predict contacts given the time slot and day indices of that period.

It can also be seen in Figures 4-4 and 4-5 that an increased k value generally leads to decreased performance in all of the models evaluated. This is attributed to the fact that human mobility is usually affected more by recent activity. For example, a student takes a class in the morning, goes to a restaurant at lunch time, then visits a café after lunch. It is highly likely that whom the student meets in the café depends more on the members encountered in the restaurant rather than those encountered in the classroom. As a result, there is a tendency of lower encounter predictability in the forecast span.

Note that there is an interesting common pattern of evaluation results among models in both datasets. Specifically, we observe that the performance of the evaluated models starts to improve when k is around 36 time slots (i.e., about one day). Recall that human movement traces were extracted daily during the 9-hour period which is divided into 36 time slots, meaning that a one-day experiment period lasts 36 time slots. This phenomenon is attributed to the fact that human mobility tends to have a temporal periodicity. Specifically, people typically keep consistent daily schedules (e.g., work, school), thus leading to some regularity in human movement.

However, the increased model performance at $k = 36$ is more clearly shown in the UB traces than Dartmouth dataset. This result can be attributed to the fact that the Dartmouth data only captured human movement on a college campus, whereas the human mobility in the UB dataset was not bounded inside the university area. Human movement on a school campus experiences less periodicity because people's schedules at a university are likely to differ between days. For instance, a student might take a Maths class on Mondays while they have a Physics class on Tuesdays. For a larger area, geographically and temporally periodic movements can be observed more clearly. For example, people normally stay at home in the morning and then go to their workplace (e.g., a company or school). In the evening, people usually spend time at home. Therefore, the prediction models exhibit more clear improvement when k is around 36 in the UB traces than in the Dartmouth traces.

Since the UB data provides human movement in a larger region than the Dartmouth traces, we discuss the performance on the UB dataset in further detail. Figures 4-6 and 4-7 show the ROC and AUC of prediction models collected on the UB data, respectively. As shown in Fig. 4-6, the proposed DHEP models with $k = 1$ achieve a higher true positive (TP) rate given a false positive (FP) rate, and result in a lower FP rate with a given TP value. For example, if the FP rate is required to be at most 0.1, Jahromi's, Jyotish's, DHEP/FFNN,

and DHEP/RNN models yield TP rates of 0.345, 0.342, 0.816, and 0.836 respectively. The ROC curves also allow us to select the most appropriate models and decision threshold value for satisfying performance requirements. For instance, suppose that an application requires a human encounter prediction model with TP values of at least 0.8 and FP values of at most 0.1, the four models of CHEP/RNN, CHEP/FFNN, DHEP/RNN, and DHEP/FFNN meet the requirement. If the DHEP/RNN model is used, the decision threshold value should be set from 0.093 to 0.2.

The AUC scores, shown in Fig. 4-7, reflect the performance across the entire range of decision threshold values; an AUC score of 1 implies perfect predictions, whereas a random binary classifier yields a value of 0.5. The results in Fig. 4-7 show that higher AUC scores are produced by the proposed DHEP models than by Jahromi's and Jyotish's methods. Specifically, DHEP/RNN, DHEP/FFNN, and DHEP/NB achieve average AUC values of 0.869, 0.857, and 0.833 in the forecast span, respectively. By contrast, the scores in Jharomi's and Jyotish's models are 0.745 and 0.739, respectively.

Figure 4-8 presents the clock-time-based F1 scores (i.e., the average F1 score over one-hour periods) of the distributed models on the UB traces. Recall that human mobility is extracted in one day over the 9-hour period ranging from 9 am to 6 pm. As shown in Fig. 4-8, when the current time t is set from 9 to 10 am, the DHEP/RNN model predicts encounters in the next hour ($k = 4$) with an F1 score of 0.744. Interestingly, the performance of DHEP/RNN and DHEP/FFNN tends to decrease in the time span from 10 am to 4 pm, which may be attributed to the fact that there is a stable tendency of people's movements in the morning due to their daily schedules. By contrast, people may have greater variation in their activities in the afternoon, such as exercising and meeting friends, thus resulting in more diverse movements. As a result, the predictions of the DHEP models are less accurate in the afternoon than in the morning.

It should also be highlighted, as shown in Fig. 4-8, that the lowest performance of the models with $k = 4, 8$, and 12 is produced when t is in periods (17h-18h), (16h-17h), and (15h-16h), respectively. The reason for this result is that when t is in (17h-18h), the DHEP models with $k = 4$ actually predict future encounters on the next day in (9h-10h), i.e., contacts after 16 hours. Due to the fact that predictions are being made regarding encounters in a distant future, the DHEP models with $k = 4$ yield the lowest evaluation results when t is in the (17h-18h) period. Similarly, the DHEP models with $k = 8$ and 12 show the worst performance when t is in the (16h-17h) and (15h-16h) periods, respectively.

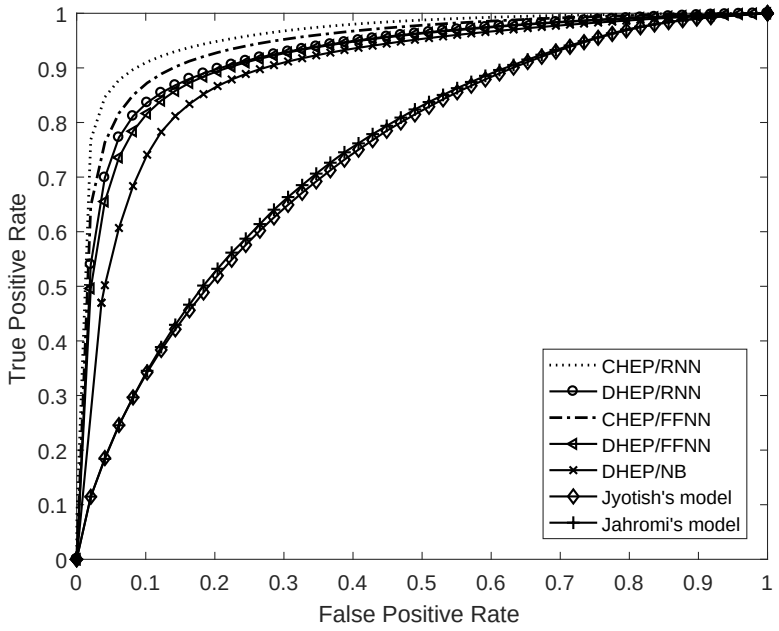


Figure 4-6: Receiver Operating Characteristic (ROC) Curves of Encounter Prediction Models with $k = 1$ on the UB traces

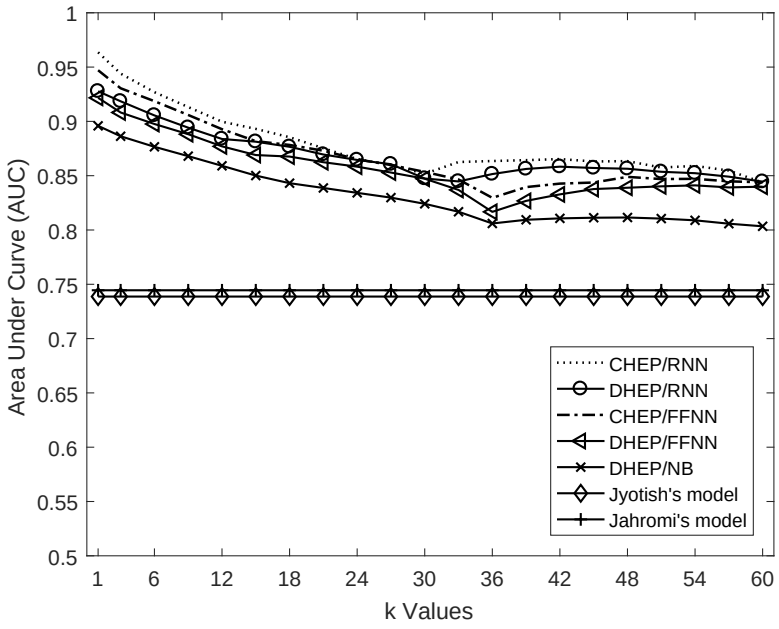


Figure 4-7: Area under Curve (AUC) of Encounter Prediction Models on the UB traces

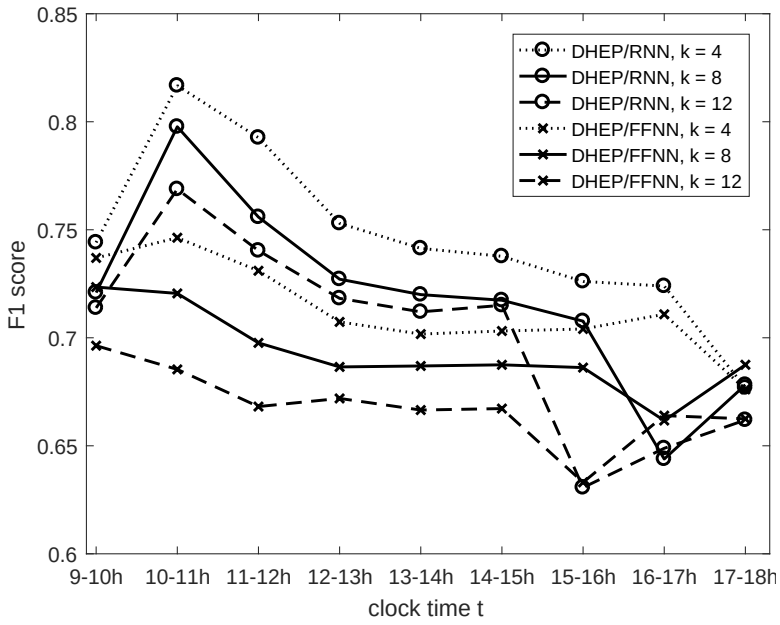


Figure 4-8: Clock-time-based F1 Score of DHEP Models on the UB dataset

4.3.3 Comparison between Distributed and Centralized Models

In this subsection, we compare the performance of the DHEP and CHEP models. Figures 4-4 and 4-5 show that the CHEP models produce a clear decreasing pattern in the time span of interest and the performance increase at k is around 36, which is similar to that of DHEP. However, the centralized models show clearly higher performance than the distributed ones. For example, as shown in Fig. 4-4(a), the CHEP/RNN model with $k = 1$ achieves 93.87% accuracy on the UB traces, as compared with the 90.74% accuracy of DHEP/RNN. More impressively, Fig. 4-4(c) shows that CHEP/RNN with $k = 1$ produces nearly 10% higher precision than DHEP/RNN.

Note that CHEP model makes encounter prediction based on the historical contact data of all users, while DHEP models use only the data of the person of interest. The higher performance of CHEP over DHEP model is attributed to the following facts. First, the DHEP prediction model tries to adjust parameters using samples which contain an encounter pattern of a specific person, thus being susceptible to the overfitting problem. On the other hand, by observing a much higher number of samples from all people, the CHEP model is able to capture common encounter patterns between people and tends to generalize better than the DHEP model.

Secondly, by leveraging the data from all people, the CHEP model can better learn the effect of social aspects on the human mobility patterns (e.g., encounter transitions and hu-

Table 4.4: Applicability comparison of the distributed and centralized prediction models with different factors

Factors	**The DHEP model is preferred if**	**The CHEP model is preferred if**
Number of People	large	small
Data exchange cost	high	low
Model maintenance cost	high	low
Requirement of predictability	medium-high	high
Data security threat	high	low

man interactions) throughout the entire network. For example, from the historical encounter data of people, CHEP discovers that people in a social group tend to encounter each other more frequently than people in different social groups. In addition, the contact between 2 people can be predicted by analyzing the interaction between these 2 people and the rest of the community. More specifically, given the assumption that person 1 meets person 2, the probability that person 1 will encounter person 3 is higher in the case that there is a contact between persons 2 and 3 than the case of no contact between persons 2 and 3.

Even though the CHEP models achieve better performance than the distributed ones, they involve a higher cost for data communication in the network and require the training of a complex prediction model as well as the maintenance of the central entity. In the case of a large number of people, e.g., thousands of participants, this cost becomes extremely large. In contrast, the distributed model can remove the burden of the central entity and the overhead in the network. Therefore, the distributed models can be easily scaled up. Moreover, the identity security level is greatly enhanced since there is no encounter information or model parameters to be exchanged over the network.

It is prominent to discuss which model (distributed or centralized) should be selected for a specific application. Table 4.4 summarizes the factors that need to be considered, including the network characteristics (e.g., the number of members, data communication cost, and model maintenance fee), requirement for the encounter prediction model (e.g., estimation accuracy), and data security threat. Specifically, in the case of a large number of members, high data exchange cost, or high maintenance fee, the DHEP model may be an appropriate choice for reducing the overhead of data communications. Meanwhile, if the application requires high prediction accuracy and data security is not a major concern, we can use the CHEP model to satisfy the required performance. In summary, since both types of prediction models have their own advantages and disadvantages, selecting a suitable

model depends on the specific application and the features of the considered network.

4.4 Chapter Summary

In this chapter, we considered the encounter prediction problem in which human encounters are predicted in the near and distant futures. The ability to accurately predict human encounters allows for improved understanding of human mobility and will facilitate a variety of applications such as context delivery in opportunistic networks and contagious disease control. Therefore, our goal is to design a distributed prediction model with high accuracy, low overhead, and ease of scaling up.

Specifically, the RNN-based and FFNN-based prediction models have been constructed, which both leverage the historical encounter information of a person in order to estimate whom that person is likely to meet in the future. Since a person's position is represented by a list of associated APs, we also propose an embedding model which produces a low dimensional representation of a person's location in order to mitigate the computation complexity of the prediction model. As can be seen from the performance results on two large-scale Wi-Fi traces, the proposed models achieve substantially more accurate encounter prediction than existing schemes. Moreover, with little sacrifice in performance, the low-cost decentralized model has been shown to be much more suitable for large networks than the centralized architectures.

Chapter 5

Concluding Remarks

5.1 Summary of the Contributions

In this dissertation, because of a variety of applications of human mobility prediction (e.g., location-based recommendation systems, geographic profiling, disease spread control, urban planning, and data forwarding in opportunistic networks), we have studied the problem of predicting human movement using large mobility traces (e.g., Wi-Fi and cellular network logs). Specifically, two prediction models are proposed to answer two separate but related questions: where a person-of-interest is most likely to visit and whom that person is most likely to encounter in the future. The proposed prediction models are summarized in a concise way as follows.

First, for the human location prediction problem, we assume that the current position of a person is unavailable due to some reasons (e.g., battery power shortage or the person may not want to share the current position). Taking the inspiration from the facts that the behavior of an individual is highly related to other members in the community, a two-phase framework is proposed, which first selects some persons with greatly correlated movements with the person of interest, and then uses the location information of these selected people to estimate the person's location. For the first phase, two methods are proposed including communication interaction similarity-based (CISB) and behavioral similarity-based (BSB). Specifically, in the CISB method, people who have similar encounters with the rest of the community are selected. Meanwhile, the BSB method finds members with similar behavioral patterns with the person-of-interest even though no direct encounters or co-locations between them are observed. In the second phase, by leveraging the location information of selected members, a prediction model based on a neural network is constructed.

Secondly, we proposed a low cost and high accurate human encounter prediction model which can be applied to large-scale networks. Inspired by the advantages of distributed systems, we leverage the historical mobility data of only the person-of-interest to construct the distributed human encounter prediction (DHEP) model. We proposed the DHEP models based on a recurrent neural network and a feed-forward neural network, the latter which contains a smaller number of training parameters even becomes more suitable for devices with constrained computing resource. Also, in order to speed up the training process, an embedding model is proposed to produce the low-dimensional representation of a person's location.

In summary, the main contributions of the dissertation are listed as follows.

- We summarized studies related to human mobility prediction and described limitations of existing works as well. Then, we formally define two problems of predicting future human location and encounter using large mobility traces.

- For the first problem, human location was estimated in a challenging situation when the recent historical location of the person-of-interest is unknown. Specifically, we proposed the two-phase framework with low time and space complexity, which first finds persons with correlated movement and then estimates the future location of the-person-of-interest by leveraging the location information of these selected people.

- For the second problem, the low cost distributed human encounter prediction (DHEP) model, which uses the movement history of only the person-of-interest, was designed to estimate future contacts of that person. More specifically, the DHEP models based on a recurrent neural network and a feed-forward neural network are constructed, the latter is more preferred for devices with limited computing capability. Moreover, the embedding model which learns the low-dimensional representation of a person's location was also proposed in order to accelerate the training process.

- Extensive experiments have been conducted to evaluate the proposed models and compare with existing methods by using large mobility datasets extracted from traces of communication networks. The evaluation results show that the proposed models can predict future human locations and encounters with higher accuracy while requiring lower cost than existing methods.

5.2 Future Works

For the location prediction model, our PCM selection methods are centralized approaches because of requiring mobility traces of all people. Therefore, the proposed PCM extraction methods are restricted to small or medium networks. In addition, information assurance needs to be taken into account since mobility data of people is exchanged in the network. As a future work, to address the limitations of the current PCM selection methods, we plan to design a better PCM selection method with lower time complexity which can be used in large networks without demanding movement data of all people.

In addition, datasets with more people and samples will be used to examine the proposed prediction model and this case allows us to apply deep neural network architectures (i.e., recurrent and convolutional neural networks) to construct prediction models with higher estimation accuracy. Specifically, deep learning techniques can be used to extract useful features from input data and then these extracted features are leveraged to build a human mobility prediction model. Moreover, in the proposed location prediction framework, each person needs to build and train their own prediction model. However, we can actually think about a centralized model which aims at predicting future locations of all people in the community. As we discussed and compared two types of models in Chapter 4, the centralized model is able to capture common movement patterns of people and tends to generalize better than the distributed one. In the case when we do not have such a large dataset, the centralized location prediction model is a preferred solution.

For the human encounter prediction model, since two people may not have a communication link even though there is an indirect encounter between them, we plan to address the problem of predicting direct encounters between people. Note that direct contacts can be extracted by using traces from device-to-device communication networks (e.g., Bluetooth or Wi-Fi direct). Therefore, the necessary condition is we need to obtain a large dataset of Bluetooth or Wi-Fi direct logs. Then, the encounter prediction model will be constructed and evaluated.

Moreover, we would like to apply the encounter prediction model to enhance the data routing in opportunistic networks. Specifically, influencer who tend to meet many other members should be selected to forward advertising packets in opportunistic networks. By using the encounter prediction model, we can be able to estimate the expected packet delay from the source to the destination, which allows us to design a data forwarding algorithm

with constrained packet delay. Furthermore, in the task allocation problem in crowdsensing, the encounter prediction model can be used to calculate the inter-meeting time between a task requester and workers, which can then allow the requester to select the most appropriate worker to satisfy a given task completion deadline.

Bibliography

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

[2] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási, "Returners and explorers dichotomy in human mobility," in *Nature communications*, 2015.

[3] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities: universal patterns in human urban mobility," *PloS one*, vol. 7, no. 5, 2012.

[4] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, June 2008.

[5] D. Karamshuk, C. Boldrini, M. Conti, and A. Passarella, "Human mobility models for opportunistic networks," *IEEE Communications Magazine*, pp. 157–165, December 2011.

[6] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2011, pp. 1082–1090.

[7] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. P. Huttenlocher, and J. M. Kleinberg, "Inferring social ties from geographic coincidences." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107 52, pp. 22 436–41, 2010.

[8] F. Alhasoun, M. Alhazzani, F. Aleissa, R. Alnasser, and F. González, "City scale next place prediction from sparse data through similar strangers," in *Proceedings of ACM KDD Workshop, Halifax, Canada, August 14, 2017 (UrbComp'17)*, 2017.

[9] Z. Zhao, H. N. Koutsopoulos, and J. Zhao, "Individual mobility prediction using transit smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 19 – 34, 2018.

[10] A. Sadilek, H. Kautz, and J. P. Bigham, "Finding your friends and following them to where you are," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. ACM, 2012, pp. 723–732.

[11] D. Zhang, D. Zhang, H. Xiong, L. T. Yang, and V. Gauthier, "Nextcell: Predicting location using social interplay from cell phone traces," *IEEE Transactions on Computers*, vol. 64, no. 2, pp. 452–463, Feb 2015.

[12] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Mar. 2003.

[13] H. A. Nguyen and S. Giordano, "Context information prediction for social-based routing in opportunistic networks," *Ad Hoc Netw.*, vol. 10, no. 8, pp. 1557–1569, Nov. 2012.

[14] A. Tatar, T. Phe-Neau, M. D. de Amorim, V. Conan, and S. Fdida, "Beyond contact predictions in mobile opportunistic networks," in *2014 11th Annual Conference on Wireless On-demand Network Systems and Services (WONS)*, April 2014, pp. 65–72.

[15] B. Wu, F. Zeng, and W. Li, "A dynamic human contacts prediction method in mobile social networks," *Procedia Computer Science*, vol. 129, pp. 123 – 127, 2018, 2017 International Conference on Identification, Information and Knowledge in the Internet of Things.

[16] Y. Li and S. Zhang, "Combo-pre: A combination link prediction method in opportunistic networks," in *2015 24th International Conference on Computer Communication and Networks (ICCCN)*, Aug 2015, pp. 1–6.

[17] Z. C. Lipton, "A critical review of recurrent neural networks for sequence learning," *CoRR*, vol. abs/1506.00019, 2015. [Online]. Available: http://arxiv.org/abs/1506.00019

[18] N. Eagle and A. (Sandy) Pentland, "Reality mining: Sensing complex social systems," *Personal Ubiquitous Comput.*, pp. 255–268, Mar. 2006.

[19] D. Kotz, T. Henderson, I. Abyzov, and J. Yeo, "CRAWDAD dataset dartmouth/campus (v. 2009-09-09)," Sep. 2009.

[20] J. Shi, C. Qiao, D. Koutsonikolas, and G. Challen, "CRAWDAD dataset buffalo/phonelab-wifi (v. 2016-03-09)," Downloaded from https://crawdad.org/buffalo/phonelab-wifi/20160309, Mar. 2016.

[21] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless Communications and Mobile Computing*, vol. 2, pp. 483–502, 2002.

[22] R. R. Roy, *Random Walk Mobility.* Boston, MA: Springer US, 2011, pp. 35–63.

[23] A. M. Borah and B. Sharma, "A survey of random walk mobility model for congestion control in manet's," *International Journal of Computer Applications*, 2015.

[24] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Physics Reports*, vol. 734, pp. 1 – 74, 2018, human mobility: Models and applications.

[25] A. Ribeiro and R. C. Sofia, "A survey on mobility models for wireless networks," University Lusofona, Portugal, Tech. Rep., Feb, 2011.

[26] M. Fiore, J. Harri, F. Filali, and C. Bonnet, "Vehicular mobility simulation for vanets," in *40th Annual Simulation Symposium (ANSS'07)*, March 2007, pp. 301–309.

[27] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "Slaw: A new mobility model for human walks," in *IEEE INFOCOM 2009*, April 2009, pp. 855–863.

[28] C. Kissling, "Social dimensions of sustainable transport: Transatlantic perspectives, edited by kieran p. donaghy, stefan poppelreuter, and georg rudinger," *Journal of Regional Science*, vol. 47, no. 2, pp. 383–385, 2007.

[29] J. A. Carrasco and E. J. Miller, "Exploring the propensity to perform social activities: a social network approach," *Transportation*, vol. 33, no. 5, pp. 463–480, Sep 2006. [Online]. Available: https://doi.org/10.1007/s11116-006-8074-z

[30] E. R. Dugundji and J. L. Walker, "Discrete choice with social and spatial network interdependencies: An empirical example using mixed generalized extreme value mod-

els with field and panel effects," *Transportation Research Record*, vol. 1921, no. 1, pp. 70–78, 2005.

[31] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proceedings of the National Academy of Sciences*, vol. 106, no. 36, pp. 15 274–15 278, 2009.

[32] M. Picornell, T. Ruiz, M. Lenormand, J. J. Ramasco, T. Dubernet, and E. Frías-Martínez, "Exploring the potential of phone call data to characterize the relationship between social network and travel behavior," *Transportation*, vol. 42, no. 4, pp. 647–668, Jul 2015.

[33] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: Improving geographical prediction with social and spatial proximity," in *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010, pp. 61–70.

[34] M. Musolesi and C. Mascolo, "Designing mobility models based on social network theory," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 11, no. 3, pp. 59–70, Jul. 2007. [Online]. Available: http://doi.acm.org/10.1145/1317425.1317433

[35] C. Boldrini and A. Passarella, "Hcmm: Modelling spatial and temporal properties of human mobility driven by users' social relationships," *Computer Communications*, vol. 33, no. 9, pp. 1056 – 1074, 2010.

[36] P. Pirozmand, G. Wu, B. Jedari, and F. Xia, "Human mobility in opportunistic networks: Characteristics, models and prediction methods," *Journal of Network and Computer Applications*, vol. 42, pp. 45 – 58, 2014.

[37] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM Transactions on Networking*, vol. 19, no. 3, pp. 630–643, June 2011.

[38] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 606–620, June 2007.

[39] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces," in *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, April 2006, pp. 1–13.

[40] J. L. Toole, C. Herrera-Yaqüe, C. M. Schneider, and M. C. González, "Coupling human mobility and social ties," *Journal of The Royal Society Interface*, vol. 12, no. 105, p. 20141128, 2015.

[41] P. A. Grabowicz, J. J. Ramasco, B. Gonçalves, and V. M. Eguíluz, "Entangling mobility and interactions in social media," *PLOS ONE*, vol. 9, pp. 1–12, 03 2014.

[42] D. Van Anh Duong and S. Yoon, "A social relationship-aware mobility model," in *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, Feb 2018, pp. 658–663.

[43] J. E. Doe, "Nodobo capture : Mobile data recording for analysing user interactions in context," 2011.

[44] B. Gianni, N. Marco De, L. Roberto, C. Antonio, C. Cristiana, T. Giovanni, A. Fabrizio, V. Alessandro, P. Alex, and L. Bruno, *Scientific Data*, vol. 2, 2015.

[45] A. Bazzani, B. Giorgini, S. Rambaldi, R. Gallotti, and L. Giovannini, "Statistical laws in urban mobility from microscopic GPS data in the area of Florence," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2010, no. 5, p. 05001, May 2010.

[46] M. McNett and G. M. Voelker, "Access and mobility of wireless pda users," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 9, no. 2, pp. 40–55, Apr. 2005. [Online]. Available: http://doi.acm.org/10.1145/1072989.1072995

[47] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell, "Nextplace: A spatio-temporal prediction framework for pervasive systems," in *Pervasive Computing*, K. Lyons, J. Hightower, and E. M. Huang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 152–169.

[48] L. H. Vu, K. Nahrstedt, S. Retika, and I. Gupta, "Joint bluetooth/wifi scanning framework for characterizing and leveraging people movement in university campus," in *MSWiM*, 2010.

[49] J. K. Laurila, J. Blom, O. Dousse, D. Gatica-perez, O. Bornet, J. Eberle, I. Aad, and M. Miettinen, "The mobile data challenge: Big data for mobile computing research," in proc. mdc workshop, 2012. trinh minh tri do received his phd degree in computer science from pierre and marie curie university, paris, france in 2010. he is working as a postdoctoral r," in *Daniel Gatica-Perez, S'01, M'02 received the Ph.D. degree*

in Electrical Engineering from the University of Washington, Seattle, in 2001. He is the Head of the Social Computing Group at Idiap Research Institute and Maitre d'Enseignement et de Recherche at.

[50] J. Ghosh, S. J. Philip, and C. Qiao, "Sociological orbit aware location approximation and routing (solar) in manet," *Ad Hoc Networks*, vol. 5, no. 2, pp. 189 – 209, 2007.

[51] J. Ghosh, S. Yoon, H. Ngo, and C. Qiao, "Sociological orbits for efficient routing in intermittently connected mobile ad hoc networks," Dept. of Computer Science and Eng., State Univ. of New York at Buffalo, Tech. Rep., 2015.

[52] J. Ghosh, S. J. Philip, and C. Qiao, "Solar: Sociological orbit aware location approximation and routing in manet," in *The 6th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, ser. MobiHoc '05, 2005.

[53] H. Pang, P. Wang, L. Gao, M. Tang, J. Huang, and L. Sun, "Crowdsourced mobility prediction based on spatio-temporal contexts," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[54] W. Mathew, R. Raposo, and B. Martins, "Predicting future locations with hidden markov models," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 911–918.

[55] N. Eagle, A. Clauset, and J. A. Quinn, "Location segmentation, inference and prediction for anticipatory computing." in *AAAI Spring Symposium: Technosocial Predictive Analytics*. AAAI, 2009, pp. 20–25.

[56] L. Cao and J. She, "Can your friends predict where you will be?" in *2014 IEEE International Conference on Internet of Things (iThings), and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom)*, Sept 2014, pp. 450–455.

[57] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "Mining user mobility features for next place prediction in location-based services," in *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ser. ICDM '12. IEEE Computer Society, 2012, pp. 1038–1043.

[58] S. Zeng, H. Wang, Y. Li, and D. Jin, "Predictability and prediction of human mobility based on application-collected location data," in *2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, Oct 2017, pp. 28–36.

[59] A. Al-Molegi, I. Alsmadi, and A. Martínez-Ballesté, "Regions-of-interest discovering and predicting in smartphone environments," *Pervasive and Mobile Computing*, vol. 47, pp. 31 – 53, 2018.

[60] T. Hossmann, T. Spyropoulos, and F. Legendre, "Putting contacts into context: Mobility modeling beyond inter-contact times," in *Proceedings of the Twelfth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, ser. MobiHoc '11. New York, NY, USA: ACM, 2011.

[61] Y. Su, X. Li, W. Tang, J. Xiang, and Y. He, "Next check-in location prediction via footprints and friendship on location-based social networks," in *2018 19th IEEE International Conference on Mobile Data Management (MDM)*, June 2018, pp. 251–256.

[62] M. Sepahkar and M. R. Khayyambashi, "A novel collaborative approach for location prediction in mobile networks," *Wireless Networks*, vol. 24, no. 1, pp. 283–294, Jan 2018.

[63] K. K. Jahromi, M. Zignani, S. Gaito, and G. P. Rossi, "Predicting encounter and colocation events," *Ad Hoc Networks*, vol. 62, pp. 11 – 21, 2017.

[64] L. Vu, Q. Do, and K. Nahrstedt, "Jyotish: Constructive approach for context predictions of people movement from joint wifi/bluetooth trace," *Pervasive and Mobile Computing*, vol. 7, no. 6, pp. 690 – 704, 2011, the Ninth Annual IEEE International Conference on Pervasive Computing and Communications (PerCom 2011).

[65] C. Chilipirea, A. C. Petre, and C. Dobre, "Predicting encounters in opportunistic networks using gaussian process," in *2013 19th International Conference on Control Systems and Computer Science*, May 2013, pp. 99–105.

[66] R. I. Ciobanu and C. Dobre, "Predicting encounters in opportunistic networks," in *Proceedings of the 1st ACM Workshop on High Performance Mobile Opportunistic Systems*, ser. HP-MOSys '12. New York, NY, USA: ACM, 2012, pp. 9–14.

[67] C. Lee, F. Gutierrez, and D. Dou, "Calculating feature weights in naive bayes with kullback-leibler measure," in *2011 IEEE 11th International Conference on Data Mining*, Dec 2011, pp. 1146–1151.

[68] K. Jahanbakhsh, V. King, and G. C. Shoja, "Predicting missing contacts in mobile social networks," *Pervasive and Mobile Computing*, vol. 8, no. 5, pp. 698 – 716, 2012.

[69] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *PNAS*, vol. 99, no. 12, pp. 7821–7826, 2002.

[70] C. Nguyen, S. Yoon, and Y. Kim, "Discovering social community structures based on human mobility traces," *Mobile Information Systems*, vol. 2017, 2017.

[71] D. Boston, S. Mardenfeld, J. S. Pan, Q. Jones, A. Iamnitchi, and C. Borcea, "Leveraging bluetooth co-location traces in group discovery algorithms," *Pervasive and Mobile Computing*, vol. 11, no. Supplement C, pp. 88 – 105, 2014.

[72] N. Eagle and A. S. Pentland, "Eigenbehaviors: identifying structure in routine," *Behavioral Ecology and Sociobiology*, vol. 63, no. 11, pp. 1689–1689, Sep 2009.

[73] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model-based collaborative filtering for personalized poi recommendations," *IEEE Transactions on Multimedia*, vol. 17, no. 6, pp. 907–918, June 2015.

[74] S. Jiang, X. Qian, T. Mei, and Y. Fu, "Personalized travel sequence recommendation on multi-source big social media," *IEEE Transactions on Big Data*, vol. 2, no. 1, pp. 43–56, March 2016.

[75] G. Zhao, X. Qian, and C. Kang, "Service rating prediction by exploring social mobile users' geographical locations," *IEEE Transactions on Big Data*, vol. 3, no. 1, pp. 67–78, March 2017.

[76] C. A. D. Jr., G. L. Pappa, D. R. R. de Oliveira, and F. de Lima Arcanjo, "Inferring the location of twitter messages based on user relationships." *Trans. GIS*, vol. 15, no. 6, pp. 735–751, 2011.

[77] B. Cao, F. Chen, D. Joshi, and P. S. Yu, "Inferring crowd-sourced venues for tweets," in *2015 IEEE International Conference on Big Data (Big Data)*, Oct 2015, pp. 639–648.

[78] W.-H. Chong and E.-P. Lim, "Tweet geolocation: Leveraging location, user and peer signals," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM '17. New York, NY, USA: ACM, 2017, pp. 1279–1288.

[79] T. S. Rappaport, W. Roh, and K. Cheun, "Mobile's millimeter-wave makeover," *IEEE Spectrum*, vol. 51, no. 9, pp. 34–58, Sept 2014.

[80] W. Zeng, C.-W. Fu, S. M. Arisona, S. Schubiger, R. Burkhard, and K.-L. Ma, "A visual analytics design for studying rhythm patterns from human daily movement data," *Visual Informatics*, vol. 1, no. 2, pp. 81 – 91, 2017.

[81] B. Jedari, L. Liu, T. Qiu, A. Rahim, and F. Xia, "A game-theoretic incentive scheme for social-aware routing in selfish mobile social networks," *Future Gener. Comput. Syst.*, vol. 70, no. C, pp. 178–190, May 2017.

[82] X. Zhang and G. Cao, "Transient community detection and its application to data forwarding in delay tolerant networks," *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 2829–2843, Oct 2017.

[83] T. Mo, S. Sen, L. Lim, A. Misra, R. K. Balan, and Y. Lee, "Cloud-based query evaluation for energy-efficient mobile sensing," in *2014 IEEE 15th International Conference on Mobile Data Management*, vol. 1, July 2014, pp. 221–224.

[84] K. Karamat Jahromi, M. Zignani, S. Gaito, and G. P. Rossi, "Predicting encounter and colocation events," *Ad Hoc Netw.*, vol. 62, no. C, pp. 11–21, Jul. 2017.

[85] I. O. Nunes, C. Celes, P. O. V. de Melo, and A. A. Loureiro, "Groups-net: Group meetings aware routing in multi-hop d2d networks," *Computer Networks*, vol. 127, pp. 94 – 108, 2017.

[86] Y. Shi, S. Chen, and X. Xu, "Maga: A mobility-aware computation offloading decision for distributed mobile cloud computing," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 164–174, Feb 2018.

[87] S. Kosta, A. Mei, and J. Stefa, "Large-scale synthetic social mobile networks with swim," *IEEE Transactions on Mobile Computing*, vol. 13, no. 1, pp. 116–129, Jan 2014.

[88] K. K. Jahromi, M. Zignani, S. Gaito, and G. P. Rossi, "Predicting encounter and colocation events," *Ad Hoc Networks*, vol. 62, pp. 11 – 21, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1570870517300768

[89] D. S. Goldberg and F. P. Roth, "Assessing experimentally derived interactions in a small world," *Proceedings of the National Academy of Sciences*, vol. 100, no. 8, pp. 4372–4376, 2003.

[90] P. Golik, P. Doetsch, and H. Ney, "Cross-entropy vs. squared error training: a theoretical and experimental comparison." in *INTERSPEECH*. ISCA, 2013, pp. 1756–1760.

[91] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[92] 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, Dec 2012.

[93] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*. New York, NY, USA: Cambridge University Press, 2002.

[94] A. Nandugudi, A. Maiti, T. Ki, F. Bulut, M. Demirbas, T. Kosar, C. Qiao, S. Y. Ko, and G. Challen, "Phonelab: A large programmable smartphone testbed," in *Proceedings of First International Workshop on Sensing and Big Data Mining*, ser. SENSEMINE'13. New York, NY, USA: ACM, 2013, pp. 4:1–4:6.

[95] J. Shi, L. Meng, A. Striegel, C. Qiao, D. Koutsonikolas, and G. Challen, "A walk on the client side: Monitoring enterprise wifi networks using smartphone channel scans," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, April 2016, pp. 1–9.

[96] W. Hsu and A. Helmy, "On nodal encounter patterns in wireless lan traces," *IEEE Transactions on Mobile Computing*, vol. 9, no. 11, pp. 1563–1577, Nov 2010.

[97] K. K. Jahromi, F. Meneses, and A. Moreira, "Impact of ping-pong events on connectivity properties of node encounters," in *2014 7th IFIP Wireless and Mobile Networking Conference (WMNC)*, May 2014, pp. 1–8.

[98] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *CoRR*, vol. abs/1310.4546, 2013.

[99] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[100] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *INTERSPEECH*, 2014.

[101] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[102] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *CoRR*, vol. abs/1409.2329, 2014.

Author's Publications

International Journals

1. **Dao, Thi-Nga**; Le, Duc V.; Yoon, Seokhoon. "Predicting Human Location Using Correlated Movements." *Electronics*, Jan. 2019

2. Vu, Duy-Son; **Dao, Thi-Nga**; Yoon, Seokhoon. "DDS: A Delay-Constrained Duty-Cycle Scheduling Algorithm in Wireless Sensor Networks." *Electronics*, Nov. 2018

3. **Dao, Thi-Nga**; Yoon, Seokhoon. "A Node Deployment Algorithm for Maximizing Network Lifetime in Delay-Constrained Duty-Cycled Wireless Sensor Networks." *International Journal of Distributed Sensor Networks*, Apr. 2018

4. **Dao, Thi-Nga**; Yoon, Seokhoon; Kim, Jangyoung. "A Deadline-Aware Scheduling and Forwarding Scheme in Wireless Sensor Networks." *Sensors*, Jan. 2016

International Conferences

1. **Dao, Thi-Nga**; Yoon, Seokhoon, "An Encounter-based Social Link Prediction Method Using Bluetooth Traces," *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, 2017

2. **Dao, Thi-Nga**; Yoon, Seokhoon, "A Sub-interval-based Scheduling Algorithm in Duty-cycled Wireless Sensor Networks," *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)*, Vienna, 2016