From Mobile Phone Sensing to Human Geo-Social Behavior Understanding

Bin Guo[1], *Yunji Liang*[1,2], *Zhiwen Yu*[1], *Minshu Li*[1], *Xingshe Zhou*[1]

[1] School of Computer Science, Northwestern Polytechnical University, P. R. China
[2] Department of Management Information System, University of Arizona, USA
guob@nwpu.edu.cn

Abstract

The study of geo-social behaviors has long been a scientific problem. In contrast to traditional social science, which suffers from the problems such as high data collection cost and imported user subjectivity, a new approach is presented to study social behaviors based on mobile phone sensing data. Different from other similar studies on mobile social sensing, three different types of geo-social behaviors, including online interaction, offline interaction, and mobility patterns, are characterized based on a newly released Nokia mobile phone dataset. We further discuss the impact factors to these behaviors as well as the correlation among them. The findings in this paper are crucial for many different fields, ranging from urban planning, location-based services, to social recommendation.

Keywords: geo-social behaviors, online/offline interaction, mobility patterns, mobile phone sensing, correlation analysis

1. Introduction

Man is by nature a social animal —*Aristotle*. In physiology or sociology, social behavior is a behavior directed towards society, or taking place between, members of the same species[1]. Numerous concepts are included in the scope of social behavior, such as online interactions, offline or face-to-face interactions, human mobility, and so on. Therefore, one of the most important problems in social behavior study is to extract the associated geographical or social properties (i.e., geo-social properties) from human daily behavior history. Many applications can benefit from human social behaviour study. For instance, understanding of human interaction patterns is important to provide social recommendations (Guo et al. 2012a) and control the spread of diseases (Eubank

[1] http://en.wikipedia.org/wiki/Social_behavior

et al. 2004), whereas accurate models of human mobility are essential for urban planning or targeted advertising (Gonzalez, Hidalgo, and Barabasi 2009).

Many efforts have been made to benefit the study of social behaviors. To study human social behaviors, data collection is of most importance. In traditional social science, data is collected majorly by self-reporting and interviewing (Lazer et al. 2009). The data obtained in this manner is often sparse and subjective, which limits the scale and accuracy on social study. With the development of pervasive sensing and wireless communication techniques, it is now possible to collect various "digital footprints" left by people while interacting with cyber-physical spaces. Leveraging the capacity to collect and analyze the "digital footprints" at community scale, a new research field called "social and community intelligence (SCI)" (Zhang, Guo, and Yu 2011a) is emerging, which aims at revealing the patterns of individual, group and societal behaviours. The scale and heterogeneity of the multimodal, mixed data sources present us an opportunity to compile the digital footprints into a comprehensive picture of individual's daily life facets, radically change the way we build computational models of human behaviour, and enable completely innovative services in areas like human health, public safety, city resource management, environment monitoring, and so on. There have been studies that aim to extract social behaviors from human digital footprints (refer to Section 2 for details). However, none of them studies three types of behaviors (online/offline/mobility) as a whole; the impact factors to the three types of social behaviors and the correlation between them are also not well addressed.

This paper aims to understand human social behaviors by analyzing the large-scale data collected from off-the-shelf smart phones. Our work is based on the Nokia open project "Mobile Data Challenge"[1]. It consists of mobile phone sensing data collected from 200 subjects over one year. Compared to the mobile phone dataset used in previous studies, the Nokia dataset contains more comprehensive information that can be sensed in off-the-shelf smart phones, such as GPS logs, Bluetooth & WLAN records, accelerometer readings, call logs, short messages, and so on. The rich dataset presents an unprecedented opportunity to study human social behaviors from different perspectives. In this paper, we aim to study human geo-social behaviors with some new

[1] http://research.nokia.com/page/12000

insights and assumptions. Specifically, the main contributions of this paper can be summarized as follows:

- ***Extracting human geo-social properties from mobile phone data***. By analyzing the large-scale mobile phone data, we attempt to understand human geo-social behaviors from three different aspects: *online interaction*, *face-to-face interaction*, and *mobility patterns*.
- ***Exploring the impact factors to social behaviors***. We analyze the impact of different factors to social behaviors, such as age, occupation, time, distribution of public infrastructure, and so on. The results indicate that different communities of users have obvious diversities in social interaction and activity patterns.
- ***Studying the correlation of geo-social properties***. The relationship between human friendship and interested visiting areas is studied; the correlation between online and offline behaviors is also studied.
- ***Using advanced techniques for social behavior analysis***. To achieve the above goals, new techniques and approaches are leveraged. First, we propose a generic framework for human-behavior understanding using the heterogeneous data collected from mobile phones. It presents a systematic approach to mobile phone-enabled social behavior study. Second, new methods are proposed for human behavior analysis. For example, to understand human mobility, we propose a clustering-based algorithm to detect the global and individual crucial regions, and leverage a combination of geo-interaction matrix and social network analysis methods to study human mobility patterns.

The rest of this paper is structured as follows. Section 2 gives a summary of the related work. A generic framework of mobile phone sensing-based social behavior study and the problems addressed in this paper are presented in Section 3. In Section 4 we analyze online and offline behaviors. The impact factors to human behaviors and the correlation between geographical and social behaviors are investigated in Section 5. Additional findings and discussions are given in Section 6. We conclude our work in Section 7.

2. Related Work

The study of human geo-social behaviors through mobile phone sensing is tightly related with the following research issues: *mobile social sensing*, *human mobility patterns*, and *geo-social property association*.

2.1. Mobile Social Sensing

Forging social connections with others is the core of what makes us human. Mobile social sensing aims to understand social behaviors and improve social connectivity in physical communities by leveraging the information detected by mobile devices.

By logging various aspects of physical interactions and communication among people (e.g., co-location, conversations, and call logs) and mining user behavioral patterns (e.g., places of interest), mobile phones can be used to analyze and predict social relationships among people. For example, the Reality Mining project of MIT can infer 95% of friendships on the basis of observational data from mobile phones (Eagle , Pentland, and Lazer 2007). Social Serendipity matches the interests among nearby people and cues informal, face-to-face interactions to like-minded ones (Eagle and Pentland 2005). Li et al. (2008) proposed a friend recommendation approach that mines similarities among users (e.g., points of interest) on the basis of their location histories, collected from GPS-equipped mobile phones. Social Contact Manager (Guo et al. 2013) has demonstrated how to merge the complementary features of online and opportunistic social networks, to automatically collect rich information (e.g., user profile, face-to-face interaction events) about their contacts. GroupMe (Guo et al. 2012a) is a group formation and recommendation tool that aims to facilitate social activity organization in the real world using mobile phone sensing.

The above studies have demonstrated how to leverage mobile sensing techniques to facilitate social interaction and enhance social connection. Different from these studies, our paper focuses on the spatial-temporal patterns of social interaction, and it particularly addresses the correlation between online and offline interaction behaviors.

2.2. Human Mobility Patterns

Observing and modeling human mobility in urban environments are essential for the planning and management of urban facilities and services. However, a key difficulty confronting urban planners and social scientists includes the challenge and cost of obtaining large-scale observational data on human mobility. The massive volume of sensing data collected from sensor-equipped smart objects (e.g., smart phones, smart vehicles, smart cards), however, paves the way for studying large-scale human mobility patterns (e.g., where people often go at 9:00 pm in Tokyo). For example, an interesting study based on the monitoring of 100,000 mobile phone users, conducted by Northeastern University in US, revealed that human mobility has a high degree of spatio-temporal regularity (Gonzalez, Hidalgo, and Barabasi 2009). Liu et al. (2009) reported the use of multiple real-time data sources (GPS data from taxis and smart card data from buses) to understand daily urban mobility patterns and traffic dynamics (e.g., hotspot detection). Although some models have been proposed to present human mobility patterns, the factors leading to those patterns are still not clear. To address this problem, Rinzivillo et al. (2013) proposed a general method to evaluate the influence of administrative borders to human mobility. Our work differs from this work at two aspects: (1) heterogeneous data is utilized to characterize user online and offline activities; (2) we find that human mobility is impacted by several factors, such as user age, occupation, friendship, and the distribution of public infrastructure.

2.3. Geo-Social Property Association

There have been several works that exploit the association between geographical and social properties. For example, by analyzing the data from Flickr, Crandall (2010) found that geographic distance affects online social interactions, where the probability of friendship decreases with distance. Sadilek, Kautz, and Bigham (2012) studied the correlation between geography and social topology by leveraging the data from four popular online social networks (OSNs). Others have investigated co-location and friendship and the possibility of predicting location using friendship information (Sadilek, Kautz, and Bigham 2012). In our work, the correlation between social ties and human mobility patterns has also been studied. However, different from existing studies that are mainly based on OSN data (only check-in points are available), our work is based on a smartphone sensing dataset which contains multi-dimensional human interaction data (e.g., human mobility trajectories, call logs, Bluetooth) collected from the real world.

3. Mobile Phone Enabled Human Behavior Understanding

Using large-scale data collected from mobile phones to understand human behaviors is a promising research area. This section we firstly give a generic framework for mobile phone-enabled social behavior study. The data that supports this work and the problems tackled will then be described.

3.1. The Generic Framework

A generic framework for mobile phone-based human geo-social behavior understanding is shown in Fig. 1. It consists of four layers, *mobile phone data collection*, *geo-social property analysis*, *geo-social property fusion*, and *applications*.

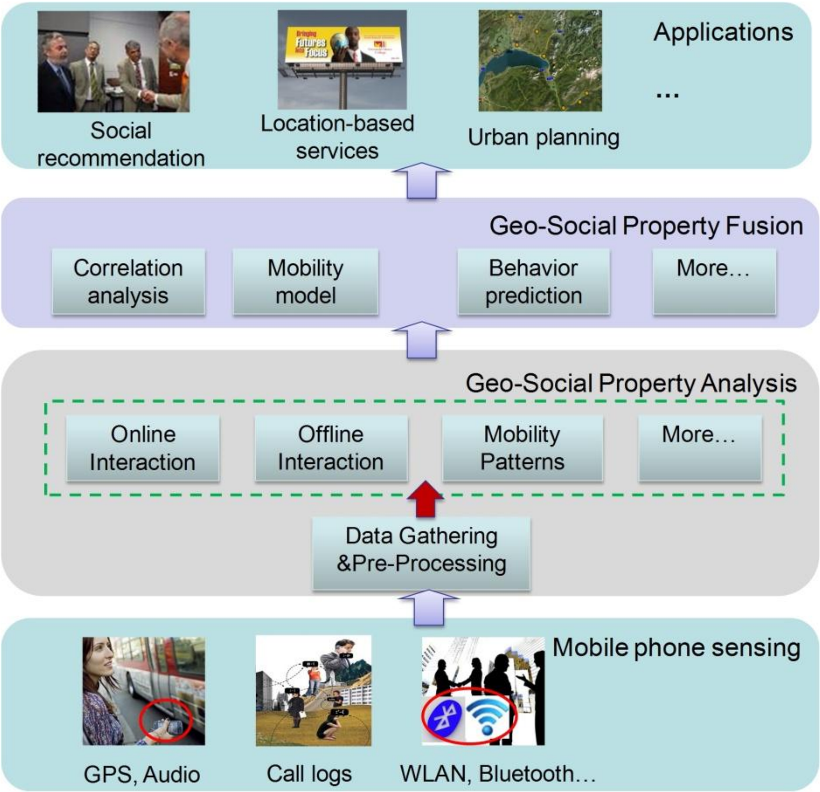


Fig. 1. A generic framework for geo-social behavior study based on mobile phone data.

The *mobile phone sensing* layer is responsible for collecting data from mobile phones using various sensors. The *geo-social property analysis* layer applies diverse data mining techniques to convert the low-level, single-modality sensing data into high-level properties, characterizing human behaviors from different dimensions, including online interaction, offline interaction, and mobility patterns. The *geo-social property fusion* layer studies the correlation among various properties (e.g., online/offline behavior interaction), and leverages the aggregated power of properties to predict human behaviors (e.g., human mobility prediction). Finally, the *application layer* includes a variety of potential services of mobile phone sensing.

3.2. Data Preparation and Problem Description

Our work is based on the Nokia open project "Mobile Data Challenge". Nokia launched a data collection campaign wherein Nokia N95 phones were allocated to nearly 200 participants in Lausanne, Switzerland. The data collection software runs on the background of the phones in a non-intrusive manner, yielding data on modalities such as social interaction and geographical behaviors. In addition, a questionnaire is included which provides user profiles such as gender, age, occupation, and so on. The dataset released in the challenge contains the data of 38 users in nearly two years. To understand the three types of social behaviors, the following data are leveraged, as summarized in Table 1.

Table 1. Data sources for geo-social feature extraction

Behavior	***Data source***	***Data type description***
Online interaction	Phone calls	Income/outgoing calls, Time (Weekday, Weekend)
	SMS messages	Income/outgoing messages
	Contact book	Contact name/friends
Offline interaction	Bluetooth	Number of people meet
Mobility	GPS	Outdoor activities
	WLAN	Indoor activities

Based on the Nokia dataset, we want to answer the following questions: (1) whether online/offline interactions are affected by various physical factors or personal profile information, such as time, age, occupation, and so on? To answer this question, we carefully choose and specify the parameter settings in Table 2. The time factor is studied at two dimensions: "workday/holiday" and "daytime/night". We choose three user groups from the Nokia dataset, one group for students, age ranging from 22-32, and the other two for full-time workers, age ranging from 22-32, and 33-44. In other words, we have two occupations and two different age groups (young and middle-aged) for a comparative study on human social behaviors. (2) What are the human mobility patterns and what factors affect human mobility? (3) Are there correlations between different types of social behaviors? We will answer these questions in the following sections.

Table 2. Parameter settings in the mobile phone data analysis

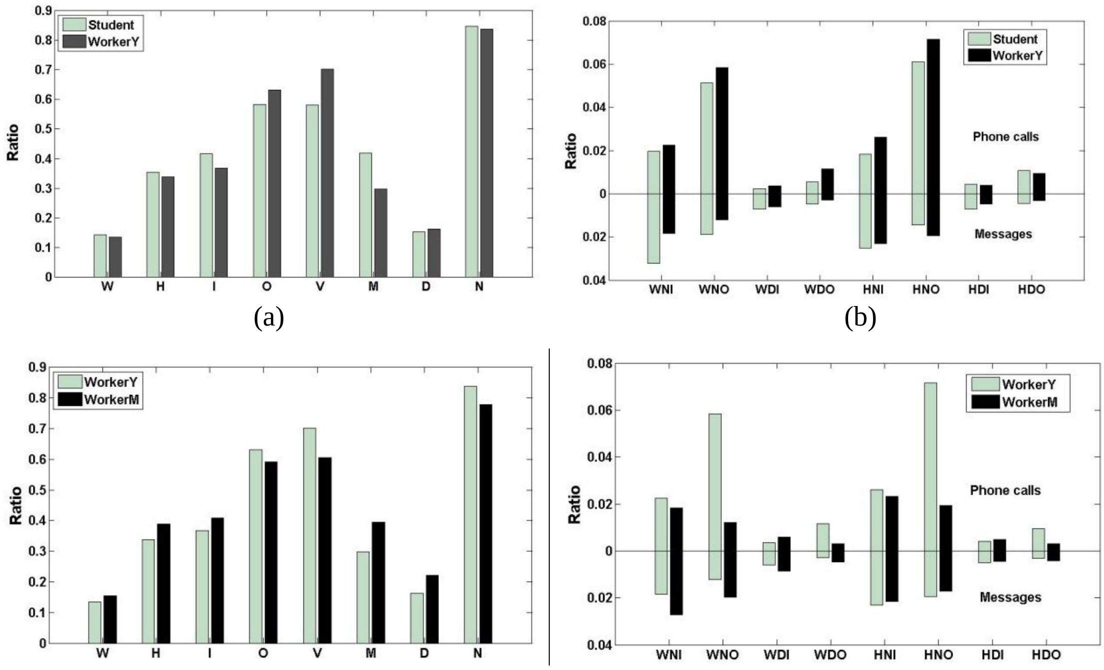
Element	***Abbreviation***
Time	W (Workday), H (Holiday), D (Daytime), N (Night)
Mobile service	I (Incoming), O (Outgoing), V (Voice call), M (Message)
Age and occupation	Student (22-32), WorkerY (22-32), WorkerM (33-44)

4. Understanding Social Interaction Behaviors

Based on the framework and dataset described in Section 3, in this section we present our observations and findings about online and face-to-face interaction behaviors. The impact of different factors to human interaction, including time, age and occupation, is also studied.

4.1. Online Interaction

We analyze the call log data to characterize users' online communication/interaction with others. The results shown in Fig. 2(a)(c) describe the distribution of daily online interaction of the three user groups under different parameter settings, while Fig. 2(b)(d) shows the detailed results by combining these parameters.



(a) (b)

(c) (d)

Fig. 2. Online communication analysis for different user groups.

There are shared observations for the three groups. For example, we find that people have more online communications at night than in the daytime. The communication in holidays is more than that of workdays. Taking consideration of the impact of occupation, we find that the communication modes for students and workers are somewhat different. For instance, the phone calls in holiday daytime of students increase significantly compared with workday daytime, while for workers there is no evident change. It reveals that in the daytime of weekends, there are more communications and possibly associated social activities (e.g., appointments, gatherings) among the buddies in the student group. When the age factor is considered, we can find that compared to young workers, middle-aged workers have more incoming calls than outgoing calls. It is possibly because they have a relatively higher position in the organization. Compared to middle-aged workers, young workers have more phone calls but fewer short messages.

The above findings by its nature reveal the correlation between human profile and daily behaviors. These findings make it possible to 'label' a user (e.g., young people, students) when we have his/her mobile sensing data. The labels can be further used for target marketing and item recommendation services.

4.2. Offline Interaction

We leverage historical Bluetooth scanning data from mobile phones to study offline, face-to-face interactions. Bluetooth proximity sensing allows us to quantify time spent in face-to-face proximity for individuals, which makes it perfect for offline interaction analysis. For example, if a user spends a lot of time at home or in the office, the number of Bluetooth devices he/she meets usually remains constant. On the contrary, if he/she has lots of outdoor social activities, the number of Bluetooth devices may increase (because he/she has chance to meet more Bluetooth device users outdoors). A further statistical analysis can be conducted over the scanned data records to extract user offline interaction patterns.

The CDF (cumulative distribution function) of Bluetooth-scanned devices of the three user groups during different periods of time is given in Fig. 3. The horizontal axis refers to the number of scanned devices and the vertical axis stands for the ratio to the total number of scanned devices. It is obvious that the number of scanned Bluetooth devices at night is significantly larger than during daytime, which indicates that people have more face-to-face interactions and outdoor activities at night. We can also find that people are more active in workdays than in holidays. Compared to young workers, it is obvious that middle-aged workers have fewer offline interactions in any time slots. The most significant difference can be found at holiday night. Workers interact more than students except in workday daytime. Maybe it is because that students can meet many classmates in the classroom.

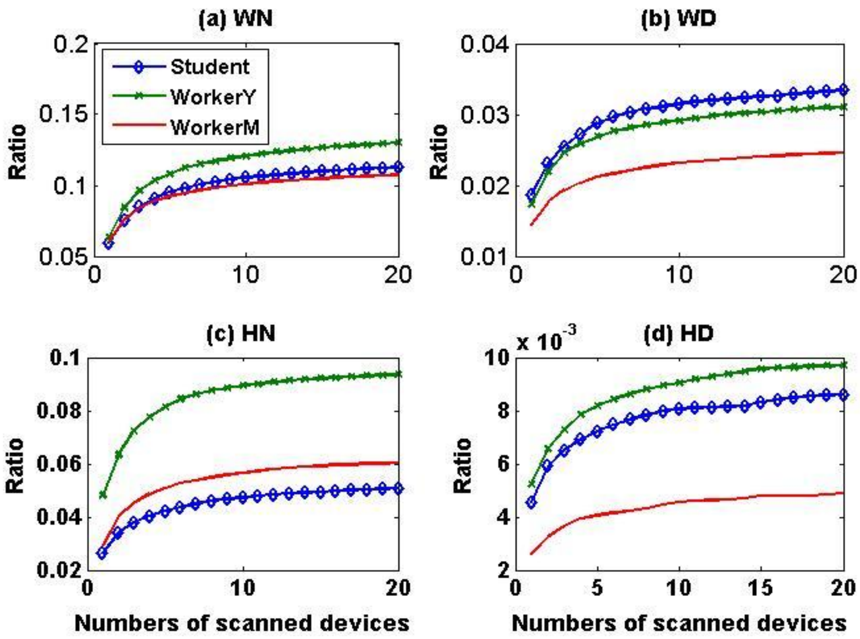


Fig. 3. The CDF distribution of Bluetooth scans.

5. Understanding Human Mobility Patterns

Having described the social interaction patterns, in this section we investigate human movement and mobility patterns in the urban area, characterizing the association of human mobility with other factors, such as user age, occupation, time, geographical limitation, and so on.

5.1. The Approach for Human Mobility Analysis

To understand human mobile patterns over a given large area, a general way is to segment the target area into crucial regions by establishing the socio-geographic boundaries (Lee, Wakamiya, and Sumiya 2011). It on one hand facilitates the detection of frequently visited areas of human, and on the other hand, enables the study of human mobility patterns based on the transformation among the crucial regions. In the following, we present our method for socio-geographic boundary detection and mobility pattern analysis.

(1) Socio-Geographic Boundary Detection

The problem can be formulated as follows. The input is the original geo-trajectories which consists of a set of spatial-temporal points distributed over a large area *S*, formulated as $(x_1, y_1, t_1), (x_2, y_2, t_2), ..., (x_n, y_n, t_n)$, where t_i represents the timestamp of a location point (x_i, y_i), $1 \le i \le n$. The output is a set of crucial regions (or point of interests (Davies et al. 2001)) by partitioning *S*, denoted by $L = \{L_1, L_2, ..., L_m\}$. The problem is how to obtain crucial regions from *S*.

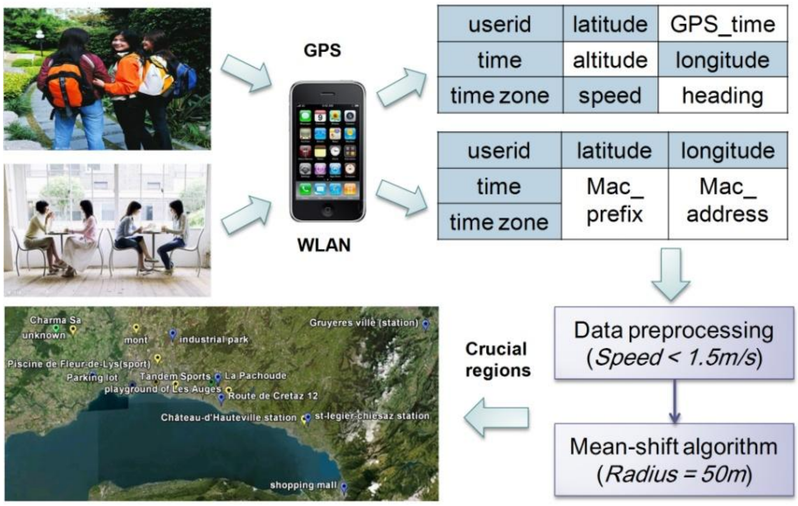


Fig. 4. The approach for mobility analysis.

The grid-based method was traditionally used to partition geo-trajectories (Zhang et al. 2011b). However, the number and size of grid cells are difficult to determine (Lee, Wakamiya, and Sumiya 2011), and it may suffer from information loss when using grid cells (Cao, Mamoulis, and Cheung 2005). In our solution, we propose a clustering-based approach to identify crucial regions from human geo-trajectories. As shown in Fig. 4, because the volume of the original geo-trajectory is extremely large, a preprocessing step is introduced to filter noisy data. In general, people often stay at crucial regions with low speed and long time span. For example, in the park, people usually stay for a long time period with low speed. Therefore, in the filter, two factors – *speed* and *time span* are considered to filter the noisy data from the original geo-dataset. According to the results of Browning (2006), the average walking speed for young pedestrians is around 1.5 m/s. Therefore, in the filter, the points whose velocity is above 1.5m/s are removed (leaving low speed points). We also calculate the mobility range during a given time period, symbolized by *t*. If the mobility range is above 1.5×*t*, the relevant geo-sequence is removed as well (it means the observed object keeps moving during that period).

The mean-shift clustering algorithm (Comaniciu and Meer 2002) is used to find the number and the centers of crucial regions. The strengths of mean-shift algorithm include: *(i) suitability for real data analysis*, *(ii) no assumption of any prior shape of data clusters*, and *(iii) robustness for arbitrary feature spaces*. More importantly, the mean-shift algorithm can detect the number of clusters automatically with the adaptive gradient ascent, without the need to predefine its value. To find fine-grained clusters with less computational complexity, we set the radius of a cluster to '50m'. It is also a reasonable setting considering the general boundaries of the crucial regions (e.g., a station, a supermarket) discussed in this work. The centers of clusters can also be calculated based on the clustering result, which indicate the estimated coordinate points of the crucial regions obtained.

(2) Geo-Interaction Matrix Extraction

Based on the segment of socio-geographic boundaries, the most frequently visited regions are detected. According to those crucial regions, we formulate the interaction among crucial regions by the geo-interaction matrix (GIM). The process of GIM extraction is illustrated in Fig. 5. Trajectory transformation projects the original geo-trajectory sequences into the sequence of crucial spatial regions according to the socio-geographic boundaries. Wij refers to the interaction strength between L_i and L_j, which is simply calculated as the interaction frequency between the two crucial regions.

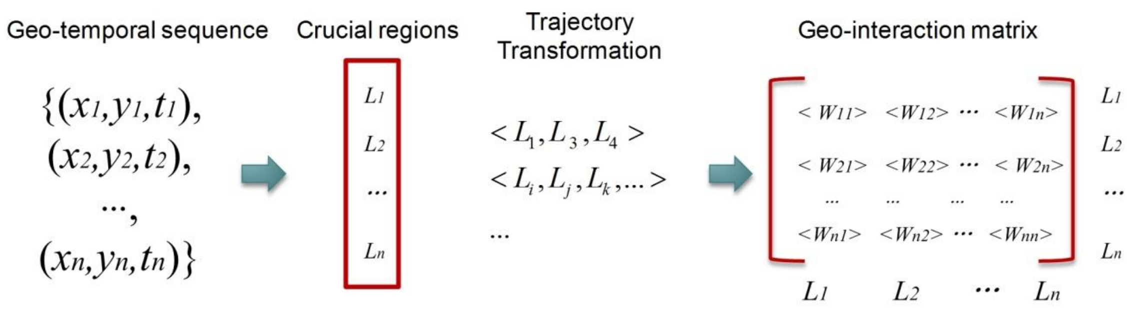
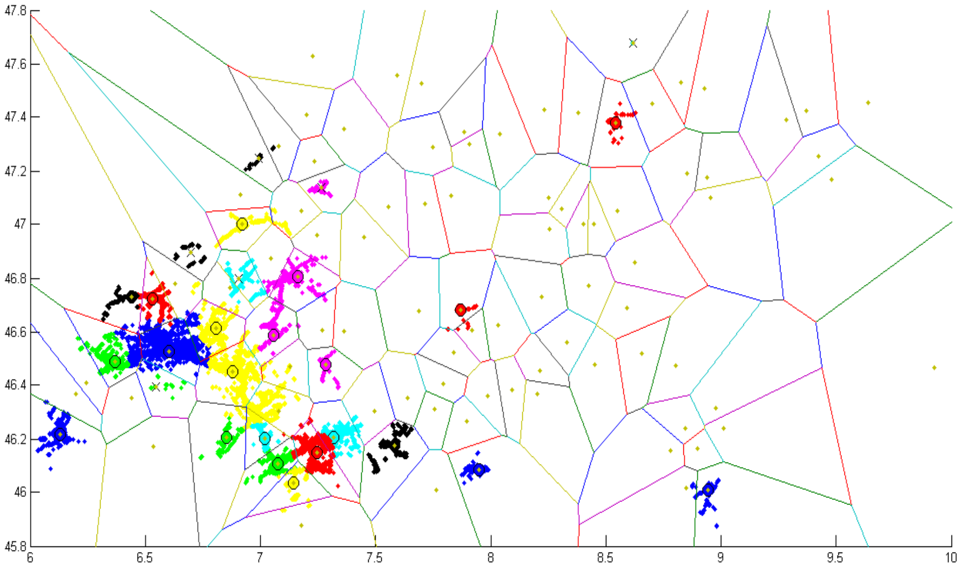


Fig. 5. GIM Extraction Process.

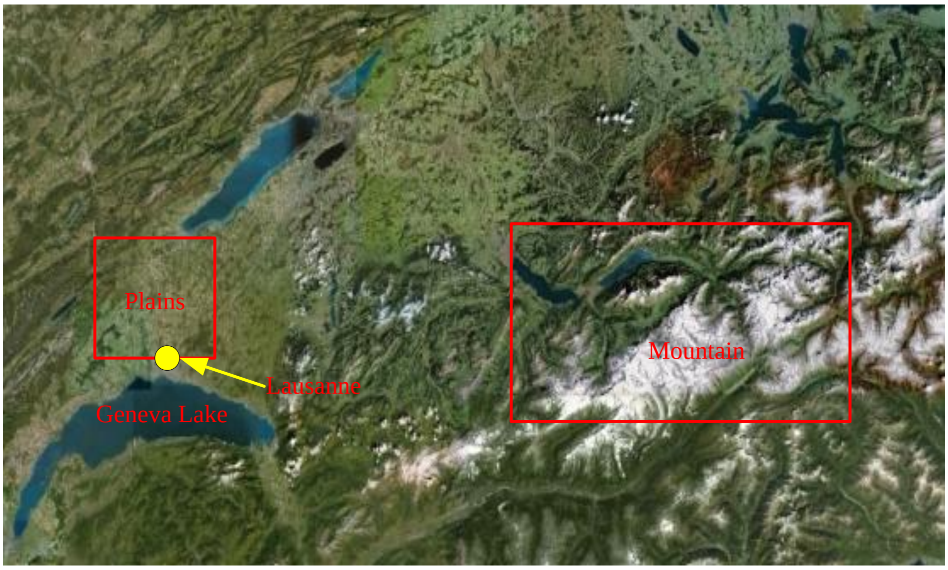
5.2. Global Mobility Patterns

To reveal human mobility patterns, we process all the geographic trajectories according to the proposed solution. Based on our proposed solution, the original trajectories are partitioned into 106 crucial regions. As shown in Fig. 6a, the partitioned regions can be presented by a voronoi diagram (Kise, Sato, and Iwata 1998) using the centers of the clusters. We find that the distribution of crucial regions is not even, with most of them concentrated on the southwest region of Lausanne while few in the north and east area.

We find that for the uneven distribution of spatial regions, the nature terrain plays a significant role. As shown in Fig. 6b, Lausanne is situated on the north coast of Geneva Lake (Lac Leman). The east and north of Lausanne are mountainous terrain with the famous ‘Jungfraujoch’ in Switzerland. However, there are open plains in the southwest of Lausanne. It is obvious that most of the spatial regions are situated on those open plains and off the coast of the Geneva Lake, which provide more farmlands, water resource and developed transportation systems.



(a)



(b)

Fig. 6. Detection of socio-geographic boundaries.

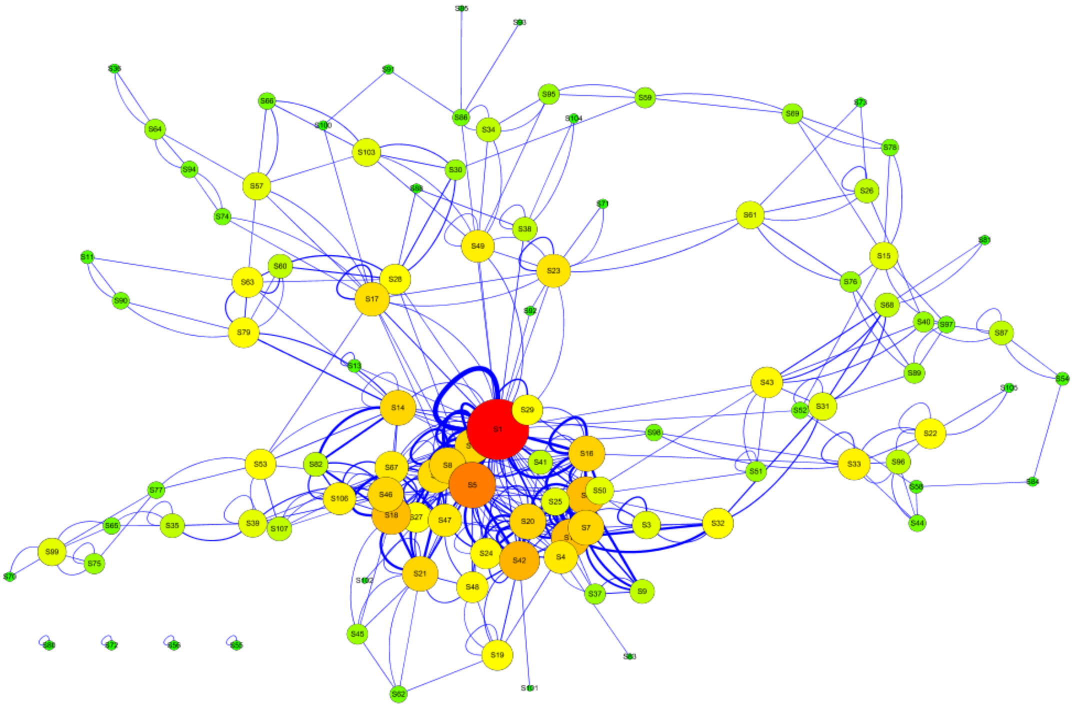


Fig. 7. Visualization of the geo-interaction matrix

We construct the GIM based on the method proposed in Section 5.1, which indicates the transfer patterns among spatial regions. We analyze and visualize the GIM with Cytoscape[1], a powerful tool for social network analysis and visualization. In Fig. 7, each

[1] http://www.cytoscape.org/

node represents a crucial region; the arc between nodes means an interaction between them. The size of a node indicates its degree, and the width of an arc represents the interaction strength. The degree distribution of the GIM is illustrated in Fig. 8. As shown in Fig. 8, most nodes in the interaction matrix have low degree, and only a few nodes have high degree, which partly shows a power law distribution.

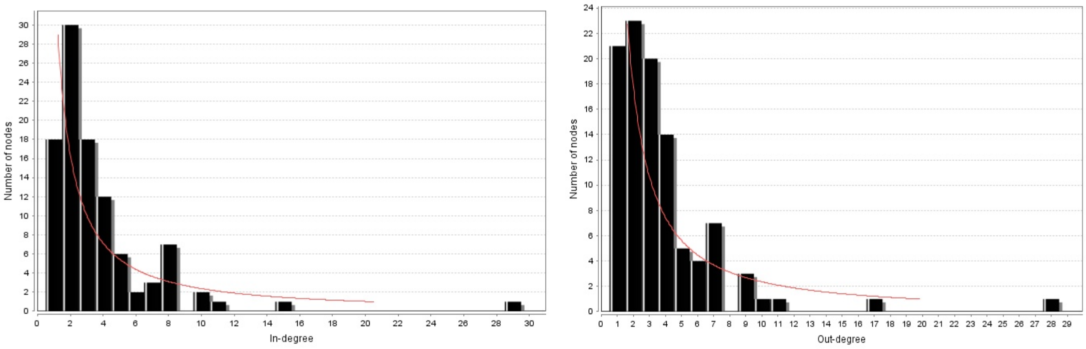


Fig. 8. The degree distribution of geo-interaction matrix

To understand the formation of nodes with high degree, we select the top eight nodes with the highest degree and project them onto the Google Earth (see Fig. 9). Table 3 shows the information of the cluster centers. It is interesting that most of those nodes with the highest degree are close to the *bus* or *metro* stations. For example, the *cluster 1* with the highest degree is very close to Prilly-Malley, which is a busy metro station with the passing of more than 500 trains each day. This phenomena indicates that the distribution of public infrastructures, such as transport junctions, supermarkets, is also an impact factor for human mobility.

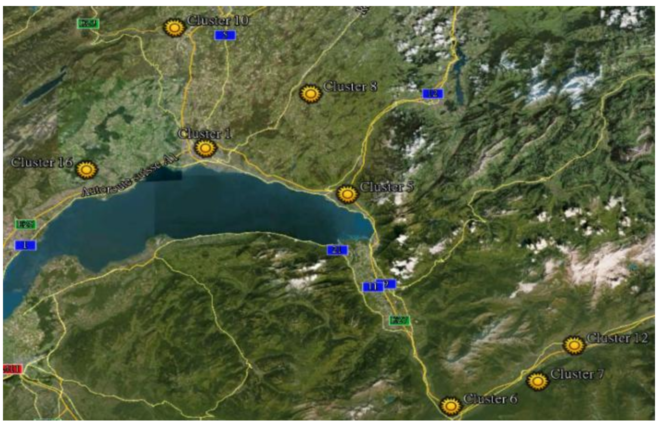


Fig. 9. Distribution of the top eight nodes ranked by its degree.

Table 3. Parameter settings in mobile phone data analysis

Cluster ID	Cluster center	Physical location	GPS point
1	6.602997E 46.525801N	Prilly – Malley	46°31'35.96''N 6°36'9.56''E
5	6.877764E 46.450483N	La Tour-De-Peilz	46°27'4.65''N 6°52'38.10''E
6	7.076414E 46.106336N	Martigny	46°6'20.90''N 7°4'44.87''E
7	7.244203E 46.148461N	La Taoumaz	46°8'37.88''N 7°14'14.59''E
8	6.807135E 46.613154N	Ecublens-Rue	46°36'37.59''N 6°48'39.84''E
10	6.531976E 46.720782N	St-Eloi	46°43'13.08''N 6°31'57.38''E
12	7.316969E 46.203881N	Aproz	46°12'22.87''N 7°18'57.39''E
16	6.366956E 46.490508N	Croix de Luisant	46°29'27.62''N 6°22'6.44''E

5.3. The Profile-Mobility Correlation Analysis

Besides analyzing the mobility patterns at the global level, we also cluster the crucial regions at the individual level. It can be used to analyze the interested activity places for individuals. From the results shown in Fig. 10, we can find that most people visit more places at night than that of daytime. The number of daytime activity places is at an average of five, which indicates that daytime activity place is relatively fixed. Further, compared to workday nights, there are more activity places at holiday nights.

We also project the centers of clustered activity places onto the Google Earth and examine the semantics/categories of them. As shown in Table 4, differences exist on the categories of activity places with respect to user age and occupation. We find that for middle-aged workers, they prefer to stay in workplaces on workdays, and have outdoor activities during holidays. In contrast, young workers often have indoor amusements during holidays. For students, their activity places are mostly centralized in the school and surrounding areas for both workdays and holidays.

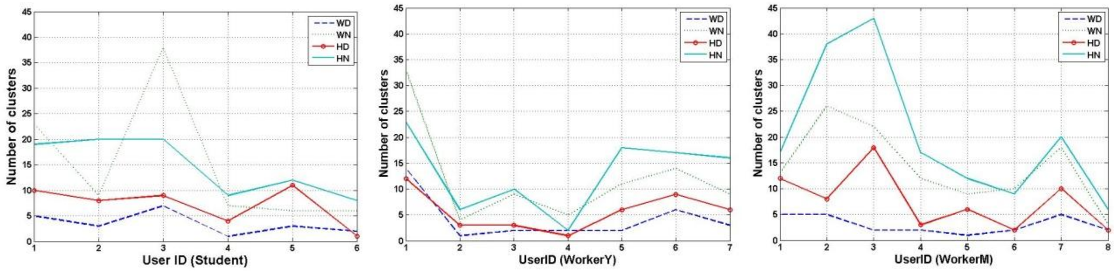


Fig. 10. Mobility patterns of different user groups (each horizontal axis point refers to a user that belongs to that group).

Table 5. Statistics of users' activity places

Group	Time	Place	Time	Place
WorkerM	WN	uptown, scenic spot, etc.	HN	beach, park, etc.
	WD	station, industrial park, etc.	HD	gym, shopping centre, etc.
WorkerY	WN	shopping centre, chapel, etc.	HN	leisure square, dining-hall, etc.
	WD	company, library, etc.	HD	shopping centre, station, etc.
Student	WN	school, industrial park, etc.	HN	school, industrial park, etc.
	WD	school, Station, etc.	HD	school, chapel, outskirts, etc.

5.4. The Geo-Social Property Correlation Analysis

To reveal the correlation between social ties and human mobility, we investigate the similarity of mobile trajectory between friends and non-friends. In Section 5.1, we have presented the work about crucial region extraction. Based on it, the *spatial matrix*, symbolized by ***S***, is introduced to present the interaction frequency between crucial regions and individual user. As shown in Fig. 11, the size of ***S*** is $m\times n$, where m is the volume of users in the dataset; n is the size of the unique crucial regions; and S_{ij} indicates the frequency that user i interacts with crucial region j.

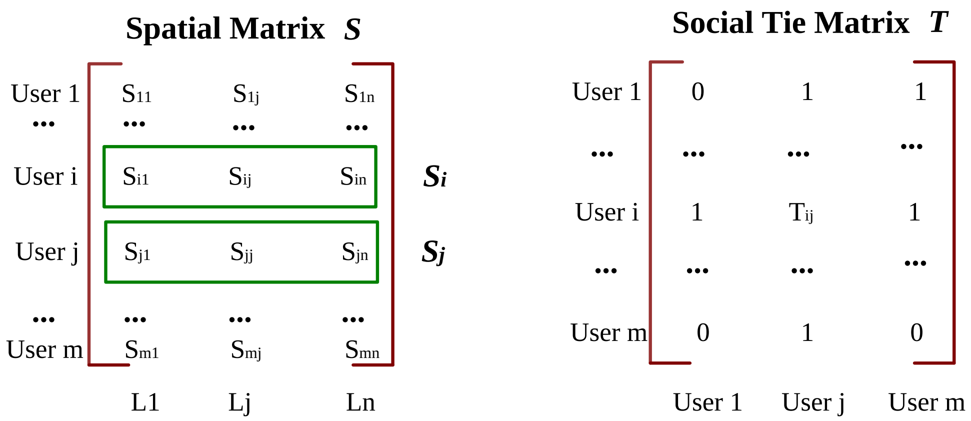


Fig. 11. Construction of spatial matrix and social tie matrix

Meanwhile, we construct the *social tie matrix* ***T*** based on a variety of mobile communication records, including call and short message logs, and Bluetooth & WLAN records. To detect the social ties among users, we make an assumption that friends are more likely to meet each other or co-located frequently. Therefore, besides the contact records, we also analyze the Bluetooth and the WLAN records. The method presented in our previous work (Guo et al. 2012b) is leveraged, where *contact frequency* and *duration* are used for friendship recognition. An example of the social tie matrix is shown in Fig. 11, where ‘0’ indicates non-friend, and ‘1’ means friendship.

Based on ***S*** and ***T***, we measure the similarity of mobile trajectories among users according to *consine* similarity as shown in Eq. (1), where $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$ are the row vectors in ***S*** respectively. By checking the relationship between user *i* and *j* in social tie matrix ***T***, we can classify the trajectory similarity into two groups: *friends* and *non-friends*. We compare the trajectory similarity between the two groups.

$$cos(s_i, s_j) = \frac{s_i}{\| s_i \|} \cdot \frac{s_j}{\| s_j \|} \tag{1}$$

We analyze the similarity of mobile trajectories between friends and non-friends. Based on the digital footprints including communication records, Bluetooth and WLAN logs, totally 37 social ties are labeled as friendship and the others as non-friendship. As shown in Fig. 12, the average of similarity among non-friendship is illustrated in a blue line, while the similarity among friends is presented by red circles. Meanwhile, the inset in Fig. 12 shows the average of friendship and non-friendship as well. It is obvious that the average similarity among friends is higher than that of non-friends, which means that the stronger the social tie is, the higher likelihood of mobility patterns. This result reveals the correlation between social ties and human mobile patterns, which makes it possible to evaluate social tie strength based on mobility patterns.

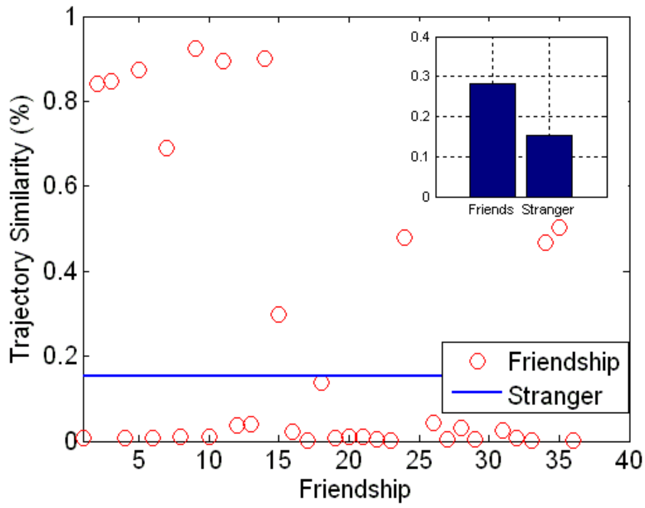


Fig. 12. Similarity of mobile trajectories between friends and non-friends.

6. Discussions

In addition to the understanding of online/offline interactions, we are also interested about the correlation between them. For instance, if a person is active in the virtual space, is he/she also active in the physical space? The prediction of human mobility based on our research results will also be discussed. We finally discuss about the fusion of heterogeneous sensing data.

6.1. Correlation between Online and Offline Behaviors

We make a correlation analysis based on the call log data, Bluetooth data and GPS/WLAN data, which reflects the three behavior dimensions: online interaction, face-to-face interaction, and mobility pattern. Since the scale of different data types varies, the data is firstly normalized into the same scale space. As shown in Fig. 13, user activities at the three different behavior dimensions show a high positive correlation. In other words, if a person is active in one dimension, it is more likely that he/she is active in other dimensions. This conclusion is particularly true for the student group (left of Fig. 13). For example, for User 3 in the student group, he/she has the most GPS&WLAN records among the six people in the same group, and the numbers of his/her Bluetooth and Call Log records are also high compared to most other people. We further calculate the Pearson's correlation coefficient among them. The results shown in Table 6 prove the positive correlation among the three behavior properties.

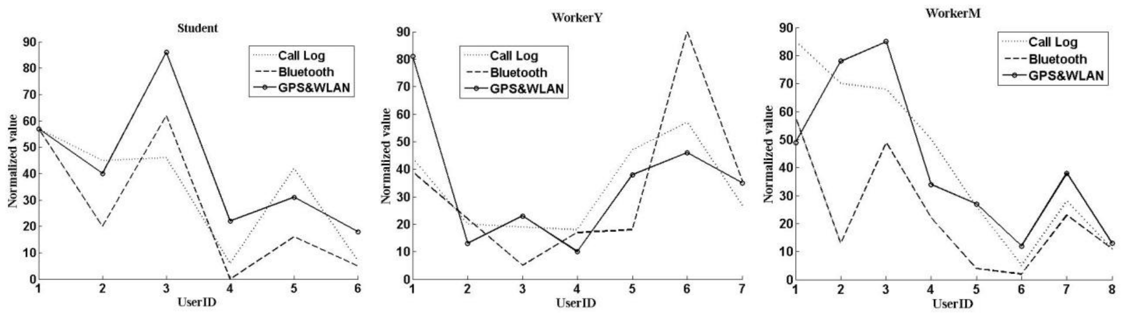


Fig. 13. The three types of human behaviors (each horizontal axis point refers to a user that belongs to that group).

Table 6. The Pearson correlation of the three human behavior properties.

	Student			WorkerY			WorkerM		
	Call log	Blue-tooth	GPS & WLAN	Call log	Blue-tooth	GPS & WLAN	Call log	Blue-tooth	GPS & WLAN
Call logs	1	-	-	1	-	-	1	-	-
Blue-tooth	0.812	1	-	0.814	1	-	0.794	1	-
GPS & WLAN	0.744	0.946	1	0.657	0.453	1	0.758	0.538	1

6.2. Towards a Mobility Prediction Model

Location-Based Services (LBSs, e.g., FourSquare) become popular in recent years. The large-scale user-generated data collected from LBSs can be used for mobility pattern analysis, urban sensing, and travel planning. All these services depend on the establishment of the mobility prediction model. However, due to the diversity and complexity of human movement, the built of effective mobility prediction models is still a big challenge.

Even though human movement have a high degree of freedom and variation, they also exhibit structural patterns due to geographic and social constraints. One of the aims of this paper is to understand what basic laws/factors govern human movement behaviors. By the analysis of both *global* and *individual* patterns, it is obvious that although the geographic terrain plays a significant role to restrain human mobility, the distribution of public infrastructure such as metro or bus stations can impact human mobility patterns. It indicates that the regularity of human mobility is controllable through urban planning. On the other hand, we analyze the impact factors leading to the variability of individual patterns. There are several reasons leading to the variability of human mobility, such as temporal factors, occupations and age. Further, as investigated in (Sadilek, Kautz, and Bigham 2012; Cho, Myers, and Leskovec 2011), humans experience a combination of periodic movement that is temporal-spatial limited and seemingly random jumps correlated with their social networks. For example, they find that people are more likely to visit a distant place if it is in proximity of an existing friend (Cho, Myers, and Leskovec 2011). We also observe that people with social ties are more similar in movement patterns than strangers, as depicted in Section 5.4.

Future work is to build a prediction model that can be used to predict human movement by taking consideration of different factors discussed above. Existing human mobility models, such as CMM (Musolesi, Hailes, and Mascolo 2004) and HCMM (Boldrini and Passarella 2010), do not consider all these factors. The introduction of multi-dimensional factors may benefit the mobility prediction model from the following aspects. (1) It enables the discovery of fine-grained patterns (knowledge), which can improve the prediction accuracy in domain-related applications. For example, we can train distinct prediction models for users of different occupation and age. (2) The introduction of social ties in mobility prediction can nurture novel applications. For example, we may predict if two people will meet in the next hour and recommend a lunch activity for them; we can control the diffusion area of information over cyber-physical spaces by considering both online social ties and offline movement (Guo et al 2012b). We plan to investigate these applications in the near future.

6.3. Fusion of Heterogeneous Sensing Data

The sensing capabilities of mobile phones keep increasing and mobile sensing data is accumulating with explosive rate. In Section 3, we propose a generic framework to gather, analyze, aggregate, and consume the data from mobile phone sensing. The proposed framework is also scalable to involve other data sources (e.g., wearable devices, static sensor networks, Internet of Things), given the gateways are built in the "data gathering" component for accessing data from other data sources. The aggregation and fusion of heterogeneous data presents new opportunities for rich context learning. For example, we can use a combination of microphone and Bluetooth sensors to characterize the ongoing social activity in a room, i.e., using the Bluetooth to identify nearby people, and using the audio data from mobile phones to recognize the activity type, such as 'meeting' or 'party' (Guo et al. 2013).

Our framework presents a general method for heterogeneous sensing data processing, including data preprocessing, geo-social feature extraction, multi-feature fusion, and application development. We have also presented our practice and methods to heterogeneous data analysis and fusion. Nevertheless, more efforts should be explored at the methodology level. One potential direction is to explore the usage of multidisciplinary knowledge, such as social network analysis, complex networks, data mining, and so on. For example, to characterize the properties of crucial regions in a city, social network analysis methods can be introduced (as presented in Section 5.2); to extract trustable information from large-scale sensing data, the method for establishing trust-based social networks (built based on interaction activities) can be explored (Mazzara, Hailes, and Mascolo 2013); to leverage mobile phone sensing for public security purpose, we should investigate the geo-social patterns regarding security-specific scenarios using advanced data mining techniques.

7. Conclusion

We have presented our research on geo-social behavior analysis using mobile phone sensing data. Three types of behaviors, including online interaction, face-to-face interaction and mobility pattern, are studied based on the newly released Nokia dataset. The impact factors to these geo-social behaviors, including age, occupation, geographical limitations are also analyzed and discussed. We have further characterized the correlation among different behavior properties and discuss the way to build an effective mobility prediction model. The results and observations presented in this paper help us understand human social behaviors via a novel way: computation-based social sensing. We are particularly interested about the interaction between online and offline human behaviors, and will exploit the aggregation and association of the geo-social data extracted from cross-space environments to nurture new mobile social sensing applications.

Acknowledgements

This work was partially supported by the National Basic Research Program of China (No.2012CB316400), the National Natural Science Foundation of China (No. 61373119, 61222209), the Program for New Century Excellent Talents in University (No. NCET-12-0466), and the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20126102110043).

References

Boldrini, C. and A. Passarella. 2010. HCMM: Modelling spatial and temporal properties of human mobility driven by users' social relationships. Computer Communications, **33**:1056-1074.

Browning, R.C., E.A. Baker, J.A. Herron, and R. Kram. 2006. Effects of obesity and sex on the energetic cost and preferred speed of walking. Journal of Applied Physiology, **100** (2): 390-398.
Cao, H., N. Mamoulis, D. W. Cheung. 2005. Mining frequent spatio-temporal sequential patterns. In Proceedings of IEEE International Conference on Data Mining (ICDM'05), pp.82-89.
Cho, E., S.A. Myers, and J. Leskovec. 2011. 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 (KDD'11), pp. 1082-1090, San Diego, USA.
Comaniciu, D., and P. Meer. 2002. Mean shift: a robust approach toward feature space analysis. IEEE Transaction on Pattern Analysis and Machine Intelligence, **24** (5): 603-619.
Crandall, D.J. 2010. Inferring social ties from geographic coincidences. Proceedings of the National Academy of Sciences (PNAS), **107**(52):22436-22441.
Davies, N., K, Cheverst, K. Mitchell, and A. Efrat. 2001. Using and Determining Location in a Context-Sensitive Tour Guide. IEEE Computer, **34**(8): 35-41.
Eagle, N., and A. Pentland. 2005. Social Serendipity: Mobilizing Social Software. IEEE Pervasive Computing **4**(2): 28–34.
Eagle, N., A. Pentland, and D. Lazer. 2007. Inferring Social Network Structure using Mobile Phone Data. Proceedings of the National Academy of Sciences (PNAS), **106**(36):15274-15278.
Eubank, S., H. Guclu, A. Kumar, M.V. Marathe, A. Srinivasan, Z. Toroczkai, N. Wang. 2004. Modeling disease outbreaks in realistic urban social networks, Nature, **429**: 180–184.
Gonzalez, M.C., C.A. Hidalgo, and A.-L. Barabasi. 2009. Understanding individual human mobility patterns. Nature, **453**:779–782.
Guo, B., H. He, Z. Yu, D. Zhang, and X. Zhou. 2012a. GroupMe: Supporting Group Formation with Mobile Sensing and Social Graph Mining. In Proceedings of 9th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous'12), Beijing, China.
Guo, B., D. Zhang, Z. Yu, and X. Zhou. 2012b. Hybrid SN: Interlinking Opportunistic and Online Communities to Augment Information Dissemination. In Proceedings of the 9th IEEE International Conference on Ubiquitous Intelligence and Computing (UIC'12), Fukuoka, Japan.
Guo, B., D. Zhang, D. Yang, Z. Yu, and X. Zhou. 2013. Enhancing Memory Recall via an Intelligent Social Contact Management System. IEEE Transactions on Human-Machine Systems, **44**(1): 78-91.
Kise, K. A. Sato, and M. Iwata. 1998. Segmentation of page images using the area voronoi diagram. Computer Vision and Image Understanding, **70**:370-382.
Lazer, D., A. Pentland, et al. 2009. Social science computational Social Science. Science, **323**(5915), 721-723.
Lee, R., S. Wakamiya, and K. Sumiya. 2011. Discovery of unusual regional social activities using geo-tagged microblogs. World Wide Web, **14** (4):321-349.
Li, Q., Y. Zheng, X. Xie, and W.Y. Ma. 2008. Mining User Similarity based on Location History. In Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS'08).
Liu, L., A. Biderman, and C. Ratti. 2009. Urban Mobility Landscape: Real Time Monitoring of Urban Mobility Patterns. In Proceedings of the 11th International Conference on Computers in Urban Planning and Urban Management (CUPUM'09).
Mazzara, M., L. Biselli, P. P. Greco, A. Marraffa, N. Qamar, S. D. Nicola. Social networks and collective intelligence: A return to the Agora. in "Social Network Engineering for Secure Web Data and Services", Luca Caviglione, Mauro Coccoli and Alessio Merlo Eds., IGI Global, 2013.
Musolesi, M., S. Hailes, and C. Mascolo. 2004. An ad hoc mobility model founded on social network theory. In Proceedings of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM'04), pp. 20–24, New York, NY, USA.
Rinzivillo, S., S. Marinardi, F. Pezzoni, M. Coscia, D. Pedreschi, and F. Giannotti. 2013. Discovering the geographical borders of human mobility. Künstliche Intelligenz, **26**(3):253-260.
Sadilek, A., H. Kautz, and J.P. Bigham. 2012. Finding your friends and following them to where you are. In Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM'12).
Scellato, S., C. Mascolo, M. Musolesi, and V. Latora. 2010. Distance matters: geo-social metrics for online social networks. In Proceedings of the Third International Conference on Online Social Networks.
Zhang, D., B. Guo, and Z. Yu. 2011a. The Emergence of Social and Community Intelligence. IEEE Computer, **44**(7):21-28.
Zhang, D., N. Li, Z. Zhou, C. Chen, L. Sun, and S. Li. 2011b. iBAT: detecting anomalous taxi trajectories from GPS traces. In Proceedings of the 13th International Conference on Ubiquitous Computing, 2011, Beijing, China.