The Places of Our Lives: Visiting Patterns and Automatic Labeling from Longitudinal Smartphone Data

Trinh Minh Tri Do and Daniel Gatica-Perez, *Member*, *IEEE*

Abstract—The location tracking functionality of modern mobile devices provides unprecedented opportunity to the understanding of individual mobility in daily life. Instead of studying raw geographic coordinates, we are interested in understanding human mobility patterns based on sequences of place visits which encode, at a coarse resolution, most daily activities. This paper presents a study on place characterization in people's everyday life based on data recorded continuously by smartphones. First, we study human mobility from sequences of place visits, including visiting patterns on different place categories. Second, we address the problem of automatic place labeling from smartphone data without using any geo-location information. Our study on a large-scale data collected from 114 smartphone users over 18 months confirm many intuitions, and also reveals findings regarding both regularly and novelty trends in visiting patterns. Considering the problem of place labeling with 10 place categories, we show that frequently visited places can be recognized reliably (over 80 percent) while it is much more challenging to recognize infrequent places.

Index Terms—Smartphone data, human mobility, place extraction, place visit, place labeling, prediction

1 Introduction

Location is a key feature for context-aware mobile services. In particular, the places in everyday life are the anchors around which social networks like FourSquare and location sharing services like Facebook Places have been built and are exploited, enabled by the widespread use of smartphones, which allow to provide location explicitly (via check-ins) or to infer it from sensors [1], [2]. Given the importance that places play in our lives, it is not surprising that current research is examining methods to automatically characterize places and understand their functions—from private to professional spaces and from transportation hubs to leisure sites [3], [4].

The availability of various forms of geolocation data coming from mobile phones has allowed researchers to study human mobility at large scale in recent years. Human trajectories were used to characterize a law of human motion from mobile cell-tower data in [5]; however, the differences in travel distances and the inherent anisotropy of each trajectory must be corrected to observe recurrent travel patterns. In another study using Foursquare data [6], it was shown that trips are not explicitly dependent on physical distance but on the set of places satisfying the objective of the trip. These and other findings confirm that the concept of place—which has a central role in this paper—is key for studying individual mobility patterns.

- *T.M.T. Do is with the Idiap Research Institute, Martigny, Switzerland. E-mail: do@idiap.ch.*
- *D. Gatica-Perez is with the Idiap Research Institute, Martigny, Switzerland, and Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland. E-mail: gatica@idiap.ch.*

Manuscript received 1 May 2012; revised 28 Aug. 2012; accepted 18 Jan. 2013; published online 31 Jan. 2013.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-2012-05-0227. Digital Object Identifier no. 10.1109/TMC.2013.19.

When considering a user's location traces as sequences of place visits, several questions arise for characterizing the place visit patterns such as how often the user visits new places and how this compares to the frequency with which he goes to places in general. Beyond these basic questions, we are also interested in identifying general place visiting patterns of a population and what affects place visiting patterns. For instance, do woman and men have different patterns? or how the visiting patterns change with respect to time? Moreover, as user behavior (including place visiting patterns) is strongly dependent on the function of the places themselves, one could expect some place semantics to be inferable from sensed data. Besides characterizing the place visits, we are also interested in the problem of classifying places into categories (e.g., home, restaurant, etc.) that we call automatic place labeling in this paper.

In the context of place labeling research [7], the connections between physical locations and their semantic meaning (their place category) are strong, and thus useful to infer the meaning of locations, for example, by using web data [8]. However, disclosing one's physical location is clearly sensitive from the perspective of privacy [9], [10], and it has been often argued that a generalized acceptance of many location-based services is limited by negative perceptions of the potential implications of location sharing [11]. Privacy-sensitive approaches to location sharing have been explored to alleviate this problem [9], including cases where physical location, as a cue for high-level recognition, is either degraded in precision or not provided [12].

We are interested in studying the place labeling problem in a *location privacy-sensitive* setting, i.e., when by design physical geolocation is not available as a feature. Labeling places in this setting is possible because all places are not created equal: our needs, obligations, and preferences

impose patterns on the places we go to [5]. In other words, few places represent the routine in our daily life [13], but a variety of places (sometimes quite significant in number) are visited too. Furthermore, smartphone users are known to follow certain patterns of phone usage based on the places they are in [14]. These two aspects are valuable in the context of privacy-sensitive place characterization from smartphone data, as places are labeled not from physical location but from contextual cues available on smartphones.

This paper presents a study on 1) characterization of real-life place visiting patterns from smartphone data; and 2) automatic place labeling in a location privacy-sensitive setting. Our work uses large-scale data collected from a population of 114 smartphone users over 18 months [15]. Our paper has three contributions. We first conduct an analysis of place visits in daily life, where places are inferred continuously from phone sensor data. This is unlike previous literature that uses cell-tower data (which has limitations of spatial resolution) and location-based social network data (which has limitations of check-in frequency). We demonstrate that in practice, beyond the few places that represent an individual's routine structure, people tend to visit new places on a regular basis, resulting in large number of places that are visited infrequently. In the second place, we demonstrate that this aspect of human behavior has key implications, showing (through an experiment involving manual labeling of visited places) that infrequently visited places are significantly harder to remember and label accurately. In the third place, we addressed the problem of automatic place labeling without using raw geolocation coordinates. Our system achieves an accuracy of 75 percent in a privacy-preserving setting, and further analysis shows that the accuracy is bounded by the frequency with which a place is visited: while the few frequently visited places in phone users' daily life can be recognized reliably, the largest fraction of places are more challenging to label. This result suggests important implications for design of mobile services.

The paper is organized as follows: Section 2 reviews related work. The data collection framework and the place extraction method are presented in Section 3. In Sections 4 and 5, we present a detailed characterization of the automatically extracted place visiting patterns. Section 6 is dedicated for the automatic place labeling in a location privacy-sensitive setting. Finally, Section 7 provides concluding remarks.

2 Related Work

Thanks to the rise of techniques for estimating people location [16], [17], the study of human mobility has emerged as a relevant topic in recent years. A large part of self-reported data used in traditional studies (e.g., [18]) could be replaced by electronic diaries generated by sensing systems [19], [20]. With a large number of built-in sensors, smartphones can record quality data without the need of additional devices. Furthermore, compared to recent efforts to collect mobility data from location-based social networks (LBSNs) like Foursquare [6], smartphones offer a definite advantage due to their ability to record data continuously if efficient systems for battery consumption are put in place [21].

Mobile phone location data can be sensed using several techniques. With the assumption that WiFi access points are fixed, WiFi traces can be used to extract places [22], which are basically represented as WiFi access points fingerprints. Other works have studied the predictability of human mobility from GSM tower data [23], whose location accuracy varies depending on the region, and is relatively coarse for locating many urban places such as cafes, restaurants, and so on. In this work, we consider location data with state-of-the-art accuracy extracted from GPS and WiFi sensors, allowing us to extract meaningful places that people visit in their daily life.

Previous works on human mobility understanding differ from our work on the variables under study. Besides seminal works on individual mobility [5], [23], there are recent works which focus on urban environments. In [24], it was shown that social relationships can explain a significant fraction of all human movement on data from LBSNs. In [4], location data were transformed into activity data to study daily activity patterns. These studies share a limitation: the lack of continuous mobility traces due to the fact that location is only available either when connections to a cellular network are made (through voice, text, or data) or when users explicitly check-in within a LBSN. Using a continuous sensing framework, Eagle was an early proponent of the identification of daily mobility patterns from simplified cell-tower data, in which each cell-tower ID was mapped to three semantic categories: *home*, *work*, and other [13]. Similar tasks were also addressed by other authors [25].

The automatic extraction of places that people visit has been addressed in previous works, with different ways of defining places [3], [26], [27]. For example, in [26] a place is defined as a segment of consecutive coordinates which satisfy a upper bound distance and a lower bound duration. This definition corresponds to what we called a *stay point* or *visit* of a given place in this paper. While many recent works have considered the place extraction task [28], [29], relatively fewer attempts have been made to infer the semantic meaning of the extracted places. In the Reality Mining data set [19], cell tower IDs for home and work were labeled, and an incomplete list of other places were labeled too but often treated as belonging to a single group (Other). The semantic annotation of places was also studied on location based social networks [30], where place category is inferred from check-ins. Closely related to our goals, the works in [7], [31] made a first attempt to recognize user activity from location traces, conducted on a small data set involving five people for one week. Addressing the problem at a much larger scale and in a daily life setting, we face multiple challenges such as noisy data recorded in real-life conditions; obtaining human annotation of places and self-reports of place visits; and performing automatic place recognition without knowing the geographic location.

Our work is conducted on a large-scale mobile phone data with state-of-the-art quality of location trace using GPS and WiFi [15]. Importantly, the longitudinal, continuously recorded location traces allowed us to characterize many individual mobility aspects that cannot be done with other data based on GSM tower IDs or LBSN check-ins in

previous works due to temporal sparsity. Instead of relying on raw geolocation or manually annotated data, our analysis is based on an automatic place extraction framework that transforms the raw location trace into a sequence of place visits. Finally, we use annotations collected for a subset of extracted places, which are used both to understand place visiting patterns and to infer the place category from smartphone data.

3 Data and Preprocessing

3.1 Data Collection

The data set comes from the Lausanne Data Collection Campaign (LDCC) [15], which was collected using Nokia N95 smartphones on a 24/7 basis in French-speaking Switzerland. The recording software is designed to run in the background, uploading recorded data automatically once a day via a user-defined WiFi connection. Since activating all sensors will wipe out the battery within a few hours, the sensing software was optimized with a state machine [15], which allows dynamic sampling rates (e.g., turn GPS off if the phone is detected to be indoors). At the end, users can record data continuously with the only restriction of charging the phone once a day. Also, participants were given additional battery chargers to charge the phone on the car or in the office if necessary. Finally, location data are not available when people go outside Switzerland due to the need of internet connection of the GPS module.

The data come from a period of 18-months started in late 2009. The data features 114 volunteer users who carry the smartphone as their main and unique mobile phone. Most of users are 20-40 years old, distributed between professionals and students from two universities. On average, each user contributed 14 months of data including non-recording time for which the phone was off (roughly 17 percent nonrecording user days). The data correspond to 20M geographic coordinates, 768K app log events, and 26M Bluetooth records, among several other sensor types.

3.2 Place Extraction

The raw location traces were represented as sequences of geographic coordinates obtained from GPS sensors or localized WiFi access points (based on co-occurrence of the AP and GPS data). In our framework, a place is defined as a small circular region (radius $= 100$ meters) that has been visited for a significant amount of time. Our choice of region size was motivated by the existence of noisy data at some places. If a smaller radius (e.g., 50 meters) is used, then actual visits risk being segmented into multiple short visits. Note that the chosen region size is similar to the one reported in previous work on place recognition [28] with GPS data, which studied three different sizes: $200\,\mathrm{m} \times 200\,\mathrm{m}, 300\,\mathrm{m} \times 300\,\mathrm{m}, \ 400\,\mathrm{m} \times 400\,\mathrm{m}$ in which 300×300 was regarded as a reasonable choice.

We use a recent place extraction approach [29] which consists of two steps. In the first step, the raw location trace is segmented into stay points and transitions. A stay point corresponds to a subsequence of the location trace for which the user stayed within a small circular regions (radius $= 100$ meters) for at least 10 minutes. Note that a place (e.g., a restaurant) that the user visited multiple times corresponds to multiple stay points, having similar geographic regions but differ in the time stamp of the visit. In the second step, a grid clustering algorithm is applied on the centers of these stay points, which results in a list of places. The clustering algorithm divides the space with a uniform grid, where each cell is a square region of side length equal to 30 meters. It starts with all stay points in the working set and an empty set for stay regions. At each iteration, the algorithm looks for the 5×5-cell region that covers most stay points and removes the covered stay points from the working set. This process is repeated until the working set is empty. Finally, the centers of 5×5-cell regions are then used to define circular stay regions that we called places.

In our framework, the place extraction is done for each user separately, therefore places are user specific. The place extraction step outputs more than 10,000 distinct places for the set of 114 users. By mapping the raw trajectory data between these places, we obtain a sequence of check-ins and check-outs on the set of places. After filtering out short duration visits of less than 10 minutes, the whole data contains 107,000 visits with a total stay duration of 618,000 hours, covering 65 percent of the time when the sensing software was active.

3.3 Data Annotation

Place annotation process. To validate the place extraction framework and collecting ground truth for place labeling task, users were asked to label their places as shown on an online map at the end of the recording period. We first described this process in [14]. Due to the large number of discovered places, annotation was obtained for only a small subset of discovered places. Each participant was asked to annotate a set of eight automatically selected places (among many others), consisting of the five most frequently visited locations of the user during the recording period, and three infrequent places that were randomly chosen from the lowest tenth percentile (in terms of total time spent). Each of the eight places was presented to the user, one by one, in random order, on an online map. The user then answered a few questions about the place by selecting one of the set of possible responses.

The first question asks if the user remembers the place, *"This is a location that I have been to in the course of the campaign,"* with three possible answers: *agree*, *disagree*, and *not sure*. For the second question, *"This location is highly relevant to me,"* the user chose a score from *1 (totally disagree)* to *5 (totally agree)*. The third question concerns the visit frequency, *"I visit this location ...,"* where the user selected one among six options: *Once a day or more*, *4-6 times per week*, *1-2 times per week*, *1-3 times per month*, *Less than once a month*, and *Never*. Finally, the user was asked to select the most appropriate place label from a list of 22 mutually exclusive predefined labels. Besides the list of labels, there is a special category named *"I don't know"* which was introduced for places that the user could not remember or did not want to provide annotation for. In practice, 17 percent of the selected infrequent places were labeled as *"I don't know"* while only 3 percent of the selected top places were assigned the same label. The total number of places annotated in this way was 912.

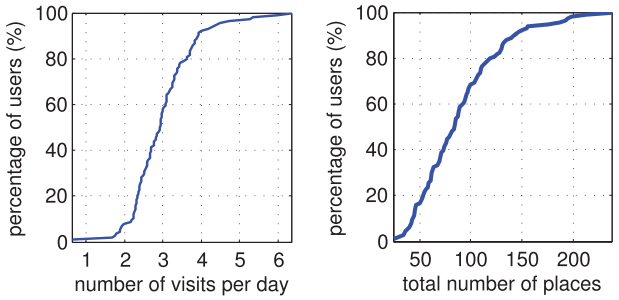

Fig. 1. Cumulative distribution of users with respect to the average number of visits per day and the number of distinct places visited during the data collection campaign.

Demographic attributes. To study the dependencies between mobility patterns and demographic attributes, additional information about the user such as age or gender was also collected. Each user filled in a questionnaire regarding demographic attributes. Besides age and gender, users also declared their marital status and their job position. There are three categories of marital status: *single or divorced*, *in a relationship*, and *married or living with partner*. Regarding job, there were four categories in increasing order of position: *training* (e.g., university student, apprentice, trainee), *PhD student*, *nonexecutive employee*, and *executive employee*.

4 Place Visiting Patterns

Our analysis starts with basic statistics of places and their dynamics. We address the following questions: How many places do people go to in everyday life? How often are these places visited? How often do people visit new places? What are the effects of demographics and calendar in the dynamics of place visits? Each key finding is highlighted below in italic font.

How many places do people visit? Fig. 1(left) illustrates the cumulative distribution of users with respect to the average number of visits per day, showing that *a large fraction of people visited from two to four places per day*. Note that the typical home-work-home daily routine corresponds to two visits per day because we only count the check-in time of visits: one at Work in the morning, and one at home in the evening. Compared to previous studies on human dynamics, our data are more complete and seems to reasonably reflect actual user mobility trends. For example, the foursquare data in [32] is highly sparse, with one check-in every five days on average, while in [5], [23], location data are only available when people make calls or send SMS .

We estimated that, on average, each person visited 90 distinct places during the data collection campaign. To get clearer understanding on the variation among people, Fig. 1(right) shows the cumulative distribution of users with respect to the number of distinct places visited during the study. Most people visited 50-150 places, furthermore there are 8 percent of users who visited more than 150 places, and 16 percent of users visited less than 50 places during the recording period. We remark that the recording time varies from 4 to 18 months depending on the user, so the results presented here are influenced by that fact. More reliable measures related to the number of places people visited will be considered later in the section.

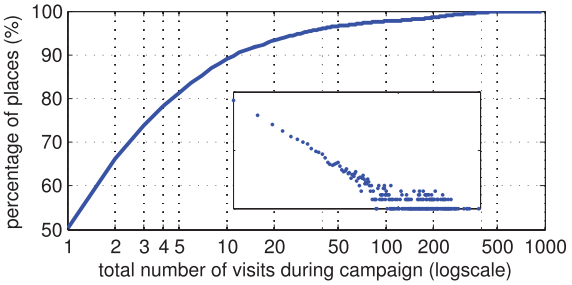

Fig. 2. Cumulative distribution of places with respect to the total number of visits. The inset shows the probability function on log-log scale.

How often are places visited? While people can easily provide a list of frequently visited places in their lives, it would be hard to exactly recall the list of places visited only a few times, even for those that correspond to valuable experiences. Studying the set of places that people visited during one year and a half, we found a simple explanation to this observation: *there are a huge number of infrequently visited places compared to a few places that people usually go*. In Fig. 2, we show the cumulative distributions of the number of places with respect to visit frequency. Note that the linear shape of the log-log plot in the inset of Fig. 2 suggests that the distribution follows the Zipf's law, which is popular for distributions of frequency over rank. One half of the places were visited only once during the campaign, and the fraction of places that were visited more than once a month is less than 10 percent, while only about 3 percent of places were visited at least once a week. This also means that over the set of places that people visited for at least 10 minutes, the largest fraction corresponds to places that were only occasionally visited.

At what rate do people visit new places? To study how often people visit new places, we computed the number of distinct places in each week and the number of new places that were discovered in that week (i.e., places that had not been visited since the data collection started). Then, the average values over active weeks (i.e., weeks with location data) were used as two basic mobility features for each user. As can be seen in Fig. 3, the average value across users for the first feature (horizontal axis) is roughly 7.5. Among the set of distinct places in a week, a few (1.7 ± 0.8) of them are new places, while others are probably frequently visited places such as home, work, or a favorite coffee shop. Note that the age group does not influence much these two

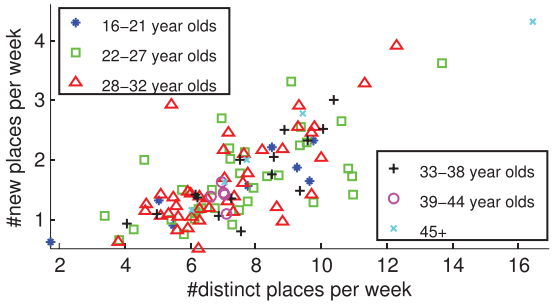

Fig. 3. Number of distinct places visited per week versus number of new places visited per week for each participant. Marker represents age group.

TABLE 1
Average Weekly Number of Visits, Average Weekly Number of Distinct Visited Places, and Average Weekly Number of New Visited Places for Each Group of People

Gender	#visits	#distinct	#new
Female	19.9±5.0	7.5±1.9	1.7±0.7
Male	22.0±6.5	7.5±2.4	1.7±0.8
Marital status	#visits	#distinct	#new
Single or Divorce	21.2±7.1	7.3±2.4	1.6±0.7
In a relationship	22.9±6.5	8.8±2.4	2.1±1.0
Married or living with partner	19.8±4.3	6.8±1.6	1.5±0.6
Job	#visits	#distinct	#new
Training	19.7±7.6	7.2±2.5	1.5±0.5
PhD student	21.8±5.7	7.4±2.0	1.6±0.7
Non-executive Employee	20.7±5.3	7.5±2.1	1.7±0.8
Executive Employee	22.2±7.8	7.8±3.1	1.9±1.0

Standard deviations are also reported.

mobility features as the distribution are mixed. There is a strong correlation ($\rho = 0.744, p < 10^{-20}$) between the number of distinct places per week and the number of new places per week. In other words, *people who visit many places a week also regularly visit many new places*. Moreover, while the number of new places per week generally decreases over time, we found that for some users, the number of places keeps increasing linearly until the end of the recording period (up to 79 weeks). Compared to a previous study [33] which suggests that the number of distinct places at time t follows $S(t) = t^{\mu}$ where $\mu = 0.6 \pm 0.02$ indicates a slow-down at large time scales, our results show that the value of μ varies depending on the person: human location traces are not created equal and *some people keep visiting new places on a regular basis* (i.e., $\mu = 1$), at least for the 18-month scale.

Do demographics affect place visiting patterns? We continue the analysis by studying the dependencies between the mobility statistics and some demographic attributes (see Table 1). Interestingly, we observe *no difference between male and female in the number of distinct places per week and the number of new places per week*. However, men are slightly more mobile than women if we look at the average number of visits per week (first column in Table 1). To judge if the difference is statistically significant, we use the Tukey-Kramer method together with ANOVA, in which two estimates of mean value being compared are significantly different if their Tukey confidence intervals are disjoint. The posthoc testing of ANOVA shows that the difference between male and female in the average number of visits per week is statistically significant with 90 percent confidence intervals. Among the set of studied demographic attributes, marital status was found to have the strongest relation to mobility features. *While previous studies based on self-reported data shows that there is a difference in mobility patterns between married and unmarried people [34], [35], we found that the difference between married people and people who are in a relationship is even more significant*. In Table 1, people who declared to be in a relationship have the highest mean values for the three reported statistics, and people who were married or living with their partner have the lowest values. Using the same posthoc testing of ANOVA, we found that the number of distinct places and the number of new places are significantly different between the group of people in relationships and the other two. For the average number of visits per week, only the difference between the group of people in relationships and the group of married people is statistically significant. Finally, we observe that the higher job position people have, the higher average number of distinct places they visit. Note, however, that the difference is relatively low compared to the variance in each job group. Statistical tests on this data did not return any significant difference.

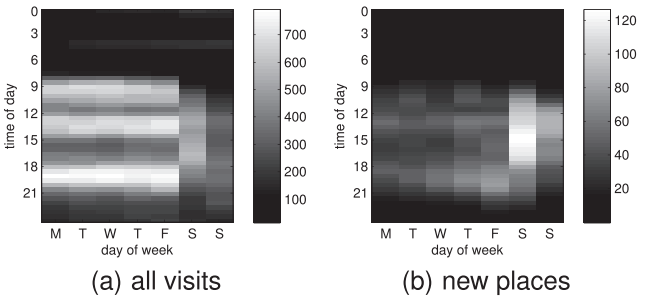

Fig. 4. Arrival time distribution over days and times of day.

Does the calendar influence place visiting patterns? Fig. 4a shows the number of visits in each time slot of 30 minutes in a weekly calendar. We remark that there is a small peak at *4 am* which corresponds to a daily reset of the phone client software that might generate unknown location states. (Despite a filtering step, a number incorrectly generated breaks remain as they are not distinguishable with real breaks.) The plot shows that *the calendar influences significantly human mobility*. Most of the visits occurred in weekdays, corresponding to daily patterns such as arriving at work (9 am), going out for lunch (12 pm-1 pm), returning to work (1 pm-2 pm), and so on. While nonnegligible on weekdays, human mobility is expected to be repetitive and predictable [5]. To verify this claim, we study the time when the phone detects a new place in Fig. 4b. People typically travel between known places on weekdays and discover new places on weekends. In both plots, we observe that people are more active on Saturday than on Sunday. We believe that this finding is due to the typical European lifestyle (e.g., commercial shops are closed on Sunday), and the results could be different for other cultures.

In this section, we presented findings obtained from the fully automatic sensing framework. The analysis is deepened in the next section by exploiting users' annotations on the set of extracted places.

5 Analyzing the Semantics of Places

We continue the study by analyzing the users' answers from the data annotation process described in Section 3.3. Although it was not feasible to obtain annotation for the several thousands of extracted places, we were able to get annotation of places that overall cover 90 percent of the total stay time in the full data set. Based on the set of annotated places, we address the following questions: How well does the place extraction algorithm work? To what degree do the selected places cover the mobility history? Does the reported visiting frequency match the estimates based on mobile phone data? How do people spend time in the main place categories? Again, we highlight the answers to these questions in italic font.

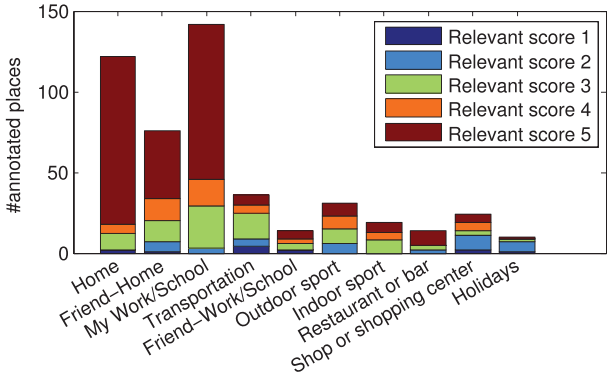

Fig. 5. Distribution of categories over the annotated places.

Validation of discovered places. Participants were asked to give relevance scores (from 1 to 5) assessing how the selected places were relevant to them. On the one hand, 95 percent of the top places were confirmed to have been visited and to be relevant to the user. About 70 percent of the top places got positive scores (4 or 5) for the question on relevance, and about 21 percent of the top places got a borderline score (3). Surprisingly, about 9 percent of the top places in terms of time spend on them are not highly regarded by their users. Focusing on these places, we found about one-third of them are frequently visited places such as home, work, restaurant, home of a friend, or train station, which means that *not all significant places in terms of amount of location traces are significant to the users*. Besides frequently visited places, there are also several insignificant places (e.g., a park next to home) and infrequently visited places in which people stayed for long time (e.g., a hotel) which were in the top 5 but deemed as not relevant. On the other hand, about one-fifth of the infrequent places are not remembered by the user, and these places are much less relevant to the user than the top places. Forty-three percent of these places got relevance score of 1 or 2. After a close inspection, the lowest tenth percentile of the discovered places contains many one-time, short-visit places such as a break spot during a hiking trip, a short discussion on the street, or waiting at an unusual train station. This somewhat unexpected effect also influenced the distribution of categories on the set of annotated places, resulting in some categories with very few annotated places. Finally, note that we could not compute the false negative rate (visited but undetected) of the place discovery method; however, a study [29] with a similar place discovery algorithm on a small subset (eight users over five months) of the data shows that the false negative rate is about 15 percent due to missing data (e.g., phone off, sensing failure).

Statistics of annotated places. For further analysis, all annotations were inspected manually to avoid ambiguous regions (e.g., a restaurant near a train station), filtering out annotations with low certainty (e.g., places that people were unsure they had visited). Due to the imbalance among categories, we redefined a new set of 10 main categories as shown in Fig. 5. *Home* stands for the user's main home and *friend-home* stands for home of a friend or relative. The category *work/school* consists of four raw labels that were presented to users: main working place, part-time working place, main school, and other school. Similarly, *friend-work/school* consists of three raw labels: working place of a friend or relative, school of a friend or relative, and school or daycare of my child. The *transportation* category corresponds to transport-related places (e.g., bus stop, metro station, etc.) but not to being on a transport. *Outdoor sport* corresponds to outside activities such as walking, hiking, skiing while *indoor sport* is the category for gym, badminton, and so on. The last three main categories (i.e., *restaurant or bar, shop or shopping center, and holiday*) correspond to raw labels that were presented to users. Other raw labels that were presented to users are: *my free-time home, other location my child visits, cultural or entertainment location, night club, nonsports related hobby, place to hang out or relax, and Other & I don't know*. At the end, we got 488 annotated places in the 10 main categories as shown in Fig. 5, and a special category called *Other* consisted of annotated places that do not belong to the 10 main categories or places with low certainty annotation. The set of top-five places was dominated by the three categories *home* (26 percent), *work/school* (15 percent) and *friend-home* (30 percent), while the set of selected infrequent place are dominated by *transportation* (24 percent), *shopping* (15 percent), and *other* (28 percent). Looking for places that are not relevant to their users, we found that *most holiday places are not highly regarded by their users, while about half of the Shopping places are marked as not important*. Among the remaining categories, *transportation* and *outdoor sport* also have high fractions of irrelevant places, which probably correspond to infrequently visited places such as a stop in the mountain or a bus stop. Finally, Table 2 reports the exact number of annotated places and other statistics for each main category. As can be seen, *while the main 10 annotated categories consist of few number of places (4.7 percent of the total number of places), it covers more than half of the visits, corresponding to 78 percent of the total stay time in both annotated and unannotated places*. *Friend-home* is the third most popular category after *home* and *work/school*, reflecting the fact that many people visited their friends or relatives on a regular basis.

TABLE 2
List of 10 Main Place Categories and Some Basic Attributes

Label	#places	#visits	time(hours)
Home	122	30343	350814
Friend-Home	76	3388	23681
Work/School	142	22638	105721
Transportation	36	208	114
Friend-Work/School	14	571	1125
Outdoor sport	31	478	1317
Indoor sport	19	669	1030
Restaurant or bar	14	432	676
Shop or shopping center	24	408	399
Holiday	10	28	212
Total of main categories	488	59163	485090
Others or Unlabeled	9799	48183	132977
Total	10287	107346	618067

While the set of places that were asked for annotation cover 90 percent of staying time, the set of annotated places that belong to the 10 main category covers 78 percent of staying time.

Comparing self-reports and mobile sensing. Based on the set of annotated places, we can study how often people visit a given place category. Note that we have access to both self-reported data and the recorded location data to answer this question. These two sources of data are plotted in Fig. 6,

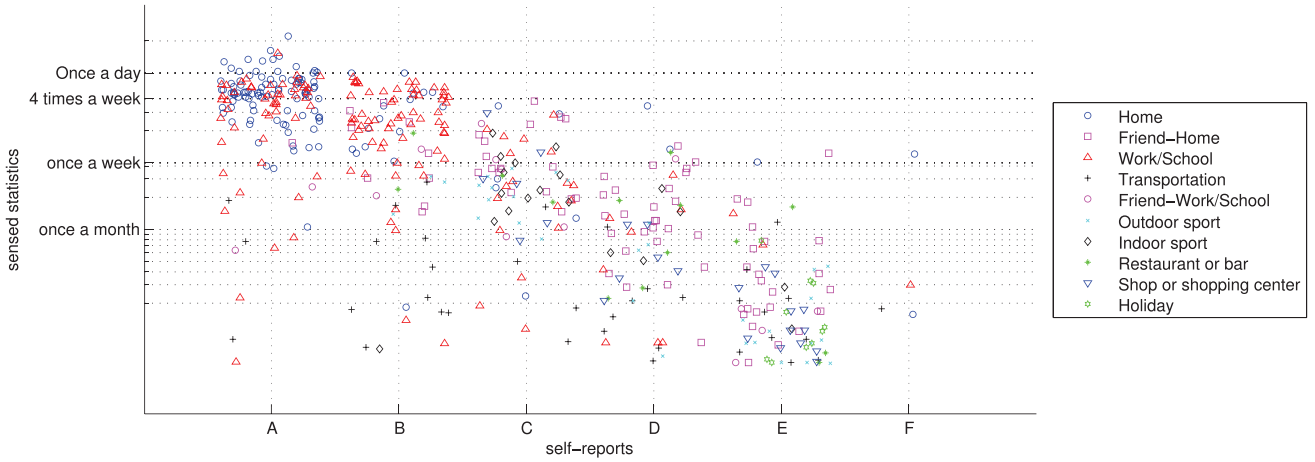

Fig. 6. Self-reported and estimated visit frequencies during the data collection campaign for the set of annotated places. Place categories are represented by distinct symbols. Note that the data points along the horizontal axis have arbitrary displacements for visualization purposes. The six self-reported frequency categories are: A) Once a day or more, B) 4-6 times per week, C) 1-2 times per week, D) 1-3 times per month, E) Less than once a month, and F) Never.

where the horizontal axis corresponds to self-reported data and the vertical axis correspond to the computed visit frequency in log-scale. At first glance, we observe that the computed frequency and self-reported frequency are highly correlated. However, *the computed frequency is lower than the self-reported data on average*. For example, for places that were reported as being visited "once a day or more," the average computed frequency is 4.3 times per week (the median is 4.4). This might come from multiple factors that affect the automatic estimation of visit frequency, such as sensing failures, nonrecording periods, or vacation periods in which daily routines were drastically changed. Given these factors, the automatic estimation is reasonable for most of the places, but there are some outlier cases where the reported data and the estimated frequency are contradictory. Focusing on the set of places with high self-reported frequency (first and second columns) but low estimated frequency (once a week or once a month in y-axis), we see a considerable number of work/school and transportation labels. On the one hand, since we do not keep short visits of less than 10 minutes in the analysis, most waiting time at transportation-related places are not considered as a visit in our framework. The difference between self-reported data and automatically computed frequencies for places in the Transportation category could be explained in this way. On the other hand, there were large places which cannot be covered by a single region of 100 m in radius, resulting in multiple discovered places (e.g., multiple amphitheaters or libraries on the university campus) that could be considered as the same place in people's minds. For these places, the difference between self-reported frequency and estimated frequency might come from the divergence of the concept of place itself. Finally, we also see that *the self-reported data are not always correct as the estimated visit frequency is sometimes much higher than the reported one* (e.g., the highest place in the fourth column, "1-3 times per month"). Other authors have also commented on inaccuracies of self-reports as compared to mobile sensing measurements [36]. Fortunately, this seems to happen only for a low number of annotated places, and the largest fraction of annotations seems to be trustworthy.

We also observe that some *home* and *work* places were reported to have low visit frequency, reflecting the fact that some people have multiple homes and working places. Hence, while *home* and *work* are relatively separable from the rest in general, there are some special cases which are much more challenging to recognize. Finally, other categories have relatively wide ranges of visit frequency. A typical example is the *friend-home* category, for which the computed visit frequency varies from four times per week to less than once a month.

Time spent in different places. We conclude this analysis by examining at how people spend time in the main categories. Without counting transitions, people in the studied population spent 62 percent of the time at *home* (median = 67 percent) and 20 percent of the time at *work/school*. *We also found that highly mobile people (in terms of number of visits per week) spent more time at friends' places and outdoor activities.* Interestingly, the time spent at work is not correlated with neither the number of visits per week nor the number of new places per week.

Overall, the analysis of annotated places contributes to the understanding of everyday life patterns. Frequently visited places are usually relevant to the user, but 9 percent of the top five places are not deemed so. Besides *home* and *work*, friends' places are a popular place category in which people spend a significant amount of time on a regular basis. Finally, we found a not entirely surprising dependency between the number of distinct places per week that a person goes to and how his time is distributed in different place categories: people who visit a large number of distinct places per week spend less time at home and more time at friend's places and other venues.

6 Automatic Place Labeling

The set of annotated places allow us to study the task of automatic place labeling in a supervised learning framework. We consider the place labeling task as a multiclass classification task with 10 place categories. Our place labeling systems employ a random forest [37] as basic classifier. Feature selection was done using greedy forward

TABLE 3
Mobility Features for a Given Place

Feature or group of features	Description
#visits per month	Average monthly visit frequency of the considered place.
Fraction #visits	Fraction of visits at the considered place.
Average/Max stay duration	Average (or the max) stay duration of visits of the considered place.
Average/Max trusted stay duration	Similar to *Average/Max stay duration* but considers only visits with trusted start and end time.
Fraction stay time	Fraction of time that the user spent at the considered place
Time per month	Total amount of time that the user spent at the considered place per month.
%Arrive(timeslot $\mathcal{T}$)	Percentage of visits having start time in the timeslot $\mathcal{T}$.
%Leave(timeslot $\mathcal{T}$)	Percentage of visits having end time in the timeslot $\mathcal{T}$.
%Int(timeslot $\mathcal{T}$)	Percentage of visits intersecting with the timeslot $\mathcal{T}$.
Time-per-month(timeslot $\mathcal{T}$)	Average time that user visited the considered place per month during the timeslot $\mathcal{T}$.
P(timeslot $\mathcal{T}$)	Probability for timeslot $\mathcal{T}$ according to the distribution of stay time at the considered place.
Rank Descend (feature $\mathcal{F}$)	#places that has larger value of $\mathcal{F}$ than the considered place. $\mathcal{F}$ is one of the above features.
Rank Ascend (feature $\mathcal{F}$)	#places that has smaller value of $\mathcal{F}$ than the considered place.

Time slot $\mathcal{T}$ could be the time of the day (e.g., 7-8 am), a larger time slot (morning, afternoon, evening, night), day of the week (monday, tuesday,...,sunday) or a longer period (weekdays, weekend).

search with cross-validation accuracy as criterion. While random forests can deal with multiclass classification tasks, we observed that the multiclass random forest is biased by popular categories (e.g., home or work) and does not focus on discriminating rare categories. For this reason, we also trained a one-versus-all random forest for each category, and then combined the votes of one-versus-all random forests to decide the winner class. In this setting, feature selection was run separately for each one-versus-all problem. All evaluation measures are computed in a leave-one-user-out setting, i.e., the system is trained on annotated places of 113 users, then tested on the annotated places of the remaining user.

6.1 Extracting Location Privacy-Sensitive Features

We focus on four categories of features derived from four data types:

1. mobility data;
2. application usage;
3. Bluetooth; and
4. WiFi.

As discussed, our system is placed in the context that privacy has high priority. Sensitive information (e.g., absolute position or Bluetooth MAC address of observed devices) are neither stored nor sent to any external database in form of queries.

Mobility features. Mobility history can disclose the purpose of a visit to a place in many cases. For example, a place that a user visits only at noon for about one hour is probably a restaurant. We extracted a large number of mobility features for each place:

Visit frequency. The number of visits per month and the fraction of visits at the considered place are used.

Visit duration. The average stay duration and the longest stay duration in a place are used. Due to sensing failures, the sequence of visits of a place might have erroneous start/end times. Therefore, we also characterize visits whose start/end times can be "trusted" by the presence of location data right before the starting time and right after the end time. In our data, we found that 50 percent of the visits have trusted start/end times.

Total time spent. We use both the fraction of stay time that the user spent at the considered place, and the normalized time spent per month considering only months when the place was visited.

Visiting time in the weekly calendar. We consider different time slots in the weekly calendar and estimate mobility features for each time slot, such as the percentage of visits intersecting with night time (0-6 am). We also computed the average time per month that the user spent at the place in a given time slot, and the conditional probability of a time slot given that the place is visited.

Note that no geolocation information is used by our place labeling system as this data type is intrusive in terms of privacy [9]. This is one key difference between our work and previous analysis using external databases (e.g., Microsolf Map Point [8], [7]) to query for types of nearby businesses and other information. The full set of mobility features is described in Table 3. Some features (e.g., number of visits) are more relevant if we consider the relative rank among all places that a user visited (e.g., the most visited place, the second most visited place, etc.) rather than the absolute number. For this reason, we also use the rank of the value as a feature.

Application usage. The set of features includes the usage statistics of some frequently used apps such as e-mail, multimedia player, web browser, voice call, SMS, and so on. For each application, we compute the average number of uses, the hourly usage frequency during visits to a given place, and the percentage of visits where the application was used (see Table 4).

Bluetooth data. We compute BT features using the number of distinct BT devices observed during a visit (see Table 4). As places may be visited multiple times, we use mean, max, and standard deviation values. As the visit durations are variable, we also compute the number of distinct BT devices for fixed time windows, like the number of distinct BT devices in the first 10 (or 20) minutes of the visit.

WiFi data. Similarly to Bluetooth data, we used the number of distinct WiFi access points to characterize the place to define WiFi features (Table 4).

Although accelerometer data are also promising for the place labeling task, we did not experimentally observe the benefit of using this data. This may be due to the sparsity of the recorded accelerometer data. At the end, we obtained 178 mobility features, 54 features for application usage, and both Bluetooth and WiFi data provide nine

TABLE 4
List of Features Extracted from Application Usage, Bluetooth, and WiFi Data

Application usage features	Description
#use(application $\mathcal{A}$)	Average number of uses of application $\mathcal{A}$ during the visits of the considered place
%use(application $\mathcal{A}$)	Percentage of visits in which application $\mathcal{A}$ was used at the considered place.
HUF(application $\mathcal{A}$)	Hourly usage frequency of application $\mathcal{A}$ during the visits of the considered place.
Bluetooth features	**Description**
Avg/Max/Std #BT	Statistics of the number of distinct BT devices detected during the visits of the considered place.
Avg/Max/Std #BT 10min	Statistics of the number of distinct BT devices detected during the first 10 minutes of visits.
Avg/Max/Std #BT 20min	Statistics of the number of distinct BT devices detected during the first 20 minutes of visits.
WiFi features	**Description**
Avg/Max/Std #WiFi	Statistics of the number of distinct WiFi APs detected during the visits of the considered place.
Avg/Max/Std #WiFi 10min	Statistics of the number of distinct WiFi APs detected during the first 10 minutes of visits.
Avg/Max/Std #WiFi 20min	Statistics of the number of distinct WiFi APs detected during the first 20 minutes of visits.

extracted features. Note that the number of mobility and application usage features were multiplied by the number of time slots or by the number of applications. In addition to numerical data, we also discretized each feature with three bins using all extracted places. Each place is finally represented by a 500-dimensional vector of numerical features and nominal features.

6.2 Place Labeling Results

Place labeling accuracy. Using leave-one-user-out cross validation for evaluation, we get an overall accuracy of 73.8 percent with multiclass system and an accuracy of 72.1 percent with one-versus-all system. Interestingly, the combination of the two systems reach an improved accuracy of 75 percent. Table 5 reports the confusion matrix between place categories. As expected, *home* and *work* are the easiest to recognize with accuracy over 90 percent. Friends' places are also recognized reliably with accuracy of 82.9 percent. The last category that has reasonable accuracy is *transportation* with 80.6 percent. At first glance, we see that *the system can only recognize reliably these categories corresponding to frequently visited places, while the recognition rate for infrequent categories is low*. While the high recognition rates for *home*, *work*, and *transportation* categories are not surprising, the result for *friend-home* is interesting. Using only mobility features, we found that recognition rate for *friends-home* is 68.4 and 20 percent of the places were confused to one of the last five categories. Adding WiFi and BT feature to the system, the recognition rate is increased to 82.9 percent by reducing the number of confusions. This points out that WiFi and BT are highly useful to recognize friend-home categories.

TABLE 5
Confusion Matrix of the Place Labeling Task and Accuracy (%TP) for Each Category

	H	FH	W/S	T	FW	O	I	R	S	Hld	%TP	%FP
Home	112	7	2	1	0	0	0	0	0	0	91.8	2.5
Friend-Home	5	63	3	3	0	0	0	1	1	0	82.9	6.1
Work/School	3	5	128	2	1	1	0	0	1	1	90.1	9.0
Transportation	0	2	3	29	0	0	0	0	1	1	80.6	4.9
Friend-Work/School	0	2	7	2	1	0	1	0	1	0	7.1	0.2
Outdoor sport	1	3	4	2	0	12	3	0	5	1	38.7	2.4
Indoor sport	0	1	5	0	0	6	5	1	1	0	26.3	0.9
Restaurant or bar	0	1	5	3	0	0	0	3	1	1	21.4	0.6
Shop or shopping center	0	1	2	7	0	3	0	1	10	0	41.7	2.6
Holiday	0	3	0	2	0	1	0	0	1	3	30.0	0.8

Rows represent the true label and columns represent the predicted label. False positive (%FP) rates are reported.

Why many infrequent categories have low recognition rate? Six of the 10 categories have low accuracy. The low recognition rate for infrequent places could be explained by multiple factors. First and foremost, *the distribution of categories is highly imbalanced and some categories only have a few annotated places*. For categories with low number of examples, it is hard to find reliable decision rules from the set of features. Due to this generalization issue, the random forest system could not exploit efficiently all features that were extracted. We believe that the recognition rates of infrequent categories could be improved if more training data are available, but highlight that data imbalance is a natural characteristic of everyday life. Second, given the low number of visits on these places, *the discriminative features cannot be accurately estimated*. For example, the number of Bluetooth devices at a given place could significantly change over time. If the place is visited just once or twice, then the estimation of the feature *Avg #BT* (average number of BT devices) could be inaccurate. Intuitively, the more frequently a place is visited, the more information there is about the place for making any prediction. Finally, one could expect that *in some cases, the place category cannot be recognized from location privacy-sensitive features*. For example, the *friend-work* category is highly confused with *work/school*, *friend-home*, or even *transportation*, which are much more popular in terms of number of annotated samples. Despite the low recognition rate, we also see promising results for some categories such as *outdoor sport*, *shop or shopping center*, or *holiday*. One could expect that with more data and data types, the recognition rate would be improved.

How does each data type improve accuracy? Finally, we study how each feature set contributes to the final accuracy. Table 6 reports the accuracy with different feature subsets. As can be seen, reasonable accuracy can be achieved with mobility features only. We found that the most important mobility features are the visiting time of a given place (e.g., the percentage of visits intersecting with the time slot

TABLE 6
Place Labeling Accuracy with Different Feature Sets

Feature sets	Acc (%)	Feature sets	Acc (%)
(M)obility	70.3		
(W)iFi	50.4	**M+W**	71.7
(B)T	51.0	**M+W+B**	74.6
(A)PP	48.0	**M+W+B+A**	75.0

5-6 am, or the fraction of departure for the time slot 8-9 pm), the fraction of time spent at the place, and the amount of time spent on weekends per month. The use of each of the three remaining feature sets separately results in modest performance. However, WiFi and Bluetooth features are relatively useful for the final system as they improve the accuracy by about 4 percent. WiFi contributes two features to the system: the average number of APs for the first 20 minutes, and the maximum number of APs observed during a single visit of the place. For the statistics on Bluetooth data, we found that the maximum number and the standard deviation of BT devices during the first 20 minutes are the most important features, which suggests that the BT statistics on 20-minute intervals is more robust than the one with 10-minute intervals or with the total stay time. Finally, the accuracy with only APP features is about 48.0 percent, with the most important applications (in descending order) being clock, profiles, search, telephone, web, maps. However, the contribution of application usage features to the final system is relatively low, which may due to the low level of usage of applications on the Nokia N95 [14]. As phone app usage is much more active for modern smartphones, application usage features should still be promising for place labeling.

7 Final Discussion and Implications

This study contributes to our understanding of places in people's daily lives and to the possibility of inferring place categories from smartphone data.

Our analysis is based on the LDCC mobile sensing framework that extracts automatically places that people visit. While people usually follow simple routines involving a few frequent places, we observe that most of them keep exploring new places, resulting in a large number of places for many individuals. If people could receive relevant recommendation for new places (e.g., restaurants, leisure-related places, etc.), the number of visited places could be even higher. The study suggests that personalized recommendation could be a plausible approach. For instance, the number of distinct places per week has significant correlation with other variables such as the frequency at which people visit new places, and this might suggest what users might find such service useful.

Together with annotation, we also used feedback on the meaning and the quality of some of the discovered places (five top frequent places and three infrequent places). Highly positive feedback was obtained for the set of frequent places, with 95 percent of them confirmed by the users. Interestingly, one-fifth of the set of infrequent places were not remembered by people. While some of these places might actually not be important for users, we believe that if the phone was able to remind the user about events that might have been forgotten, they could elicit positive personal uses related to memory and self-reflection.

Along with this analysis on place labeling, the prediction task was proposed as part of the Mobile Data Challenge [38]. Our results in this paper show that it is not easy to recognize the semantic meaning for a large number of places in our lives if the actual physical location is not known. Frequently visited places such as home, work or the home of a friend can be reliably recognized using only location-sensitive smartphone data. As people spend most time in these frequent places, the place category can be recognized accurately. However, the recognition rate for infrequent places is relative low due to many factors including the current privacy-sensitive data we use. This is a valuable result if privacy is the main concern (i.e., if users wanted that the category of certain infrequent places could not be inferred automatically).

If the category of infrequent places is not a privacy concern, one could think of adding more sensors to the smartphone to improve the recognition rate of infrequent places, besides collecting more training data. For example, a light sensor and a thermometer could help recognizing indoor-outdoor environments without heavy battery consumption. Acoustic noise sensors could be a privacy-sensitive solution for adding audio data. We remark that while the absolute recognition rate is low for many infrequent categories, the results are still promising and our current system could be useful even for complex situations. For example, in the case of confusion among a few categories, the system could present the most likely categories to the user asking for annotation. This could be a reasonable solution if the required amount of interaction was minimal and there were clear incentives for users to tag their own data. This direction might be investigated in the future.

Acknowledgments

This work was funded by Nokia Research Center Lausanne (NRC) through the LS-CONTEXT project. The authors thank Jan Blom (Nokia Research) for sharing the user surveys, Juha Laurila (Nokia Research) for discussions, and Olivier Bornet (Idiap) for technical support.

References

[1] L. Barkhuus, B. Brown, M. Bell, S. Sherwood, M. Hall, and M. Chalmers, "From Awareness to Repartee: Sharing Location within Social Groups," *Proc. 26th SIGCHI Conf. Human Factors in Computing Systems (CHI '08)*, pp. 497-506, 2008.

[2] H. Cramer, M. Rost, and L.E. Holmquist, "Performing a Check-in: Emerging Practices, Norms and 'Conflicts' in Location-Sharing Using Foursquare," *Proc. 13th Int'l Conf. Human Computer Interaction with Mobile Devices and Services (MobileHCI '11)*, pp. 57-66, 2011.

[3] J. Hightower, S. Consolvo, A. LaMarca, I. Smith, and J. Hughes, "Learning and Recognizing the Places We Go," *Proc. Seventh Int'l Conf. Ubiquitous Computing (UbiComp '05)*, pp. 159-176, 2005.

[4] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti, "Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data," *Proc. First Int'l Conf. Human Behavior Understanding*, pp. 14-25, 2010.

[5] 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.

[6] 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, p. e37027, 2012.

[7] L. Liao, D. Fox, and H. Kautz, "Location-Based Activity Recognition Using Relational Markov Networks," *Proc. 19th Int'l Joint Conf. Artificial Intelligence (IJCAI '05)*, pp. 773-778, 2005.

[8] R. Hariharan, J. Krumm, and E. Horvitz, "Web-Enhanced GPS," *Proc. Int'l Workshop Location and Context-Awareness*, pp. 95-104, 2005.

[9] K.P. Tang, P. Keyani, J. Fogarty, and J.I. Hong, "Putting People in Their Place: An Anonymous and Privacy-Sensitive Approach to Collecting Sensed Data in Location-Based Applications," *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI '06)*, pp. 93-102, 2006.

[10] J. Krumm, "A Survey of Computational Location Privacy," *Personal Ubiquitous Computing,* vol. 13, pp. 391-399, Aug. 2009.
[11] J. Krumm, "Inference Attacks on Location Tracks," *Proc. Fifth Int'l Conf. Pervasive Computing (Pervasive '07),* pp. 127-143, 2007.
[12] M. Duckham and L. Kulik, "Location Privacy and Location-Aware Computing," *Dynamic & Mobile GIS: Investigating Change in Space and Time,* pp. 34-51, CRC Press, 2006.
[13] N. Eagle and A. Pentland, "Eigenbehaviors: Identifying Structure in Routine," *Behavioral Ecology and Sociobiology,* vol. 63, no. 7, pp. 1057-1066, May 2009.
[14] T.M.T. Do, J. Blom, and D. Gatica-Perez, "Smartphone Usage in the Wild: A Large-Scale Analysis of Applications and Context," *Proc. 13th Int'l Conf. Multimodal Interfaces (ICMI '11),* pp. 353-360, 2011.
[15] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila, "Towards Rich Mobile Phone Data Sets: Lausanne Data Collection Campaign," *Proc. Seventh Int'l Conf. Pervasive Services (ICPS),* 2010.
[16] T. Liu, P. Bahl, and I. Chlamtac, "Mobility Modeling, Location Tracking, and Trajectory Prediction in Wireless ATM Networks," *IEEE J. Selected Areas in Comm.,* vol. 16, no. 6, pp. 922-936, Aug. 1998.
[17] R. Bajaj, S.L. Ranaweera, and D.P. Agrawal, "GPS: Location-Tracking Technology," *Computer,* vol. 35, no. 4, pp. 92-94, Apr. 2002.
[18] J. Krumm and A.J.B. Brush, "Learning Time-Based Presence Probabilities," *Proc. Ninth Int'l Conf. Pervasive Computing (Pervasive '11),* pp. 79-96, 2011.
[19] N. Eagle and A. Pentland, "Reality Mining: Sensing Complex Social Systems," *Personal Ubiquitous Computing,* vol. 10, no. 4, pp. 255-268, 2006.
[20] M. Raento, A. Oulasvirta, R. Petit, and H. Toivonen, "Context-phone: A Prototyping Platform for Context-Aware Mobile Applications," *IEEE Pervasive Computing,* vol. 4, no. 2, pp. 51-59, Jan.-Mar. 2005.
[21] D. Kim, Y. Kim, D. Estrin, and M. Srivastava, "Sensloc: Sensing Everyday Places and Paths Using Less Energy," *Proc. Eighth ACM Conf. Embedded Networked Sensor Systems (Sensys '10),* pp. 43-56, 2010.
[22] L. Vu, Q. Do, and K. Nahrstedt, "Jyotish: A Novel Framework for Constructing Predictive Model of People Movement from Joint Wifi/Bluetooth Trace," *Proc. IEEE Int'l Conf. Pervasive Computing and Comm. (PerCom '11),* pp. 54-62, 2011.
[23] C. Song, Z. Qu, N. Blumm, and A.-L. Barabsi, "Limits of Predictability in Human Mobility," *Science,* vol. 327, no. 5968, pp. 1018-1021, 2010.
[24] E. Cho, S.A. Myers, and J. Leskovec, "Friendship and Mobility: User Movement in Location-Based Social Networks," *Proc. 17th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD),* pp. 1082-1090, 2011.
[25] K. Farrahi and D. Gatica-Perez, "Probabilistic Mining of Socio-Geographic Routines from Mobile Phone Data," *IEEE J. Selected Topics in Signal Processing,* vol. 4, no. 4, pp. 746-755, Aug. 2010.
[26] J.H. Kang, W. Welbourne, B. Stewart, and G. Borriello, "Extracting Places from Traces of Locations," *Proc. Second ACM Int'l Workshop Wireless Mobile Applications and Services on WLAN Hotspots (WMASH '04),* pp. 110-118, 2004.
[27] M. Kim, D. Kotz, and S. Kim, "Extracting a Mobility Model from Real User Traces," *Proc. IEEE INFOCOM,* Apr. 2006.
[28] V.W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Collaborative Location and Activity Recommendations with GPS History Data," *Proc. 19th Int'l Conf. World Wide Web (WWW),* 2010.
[29] R. Montoliu and D. Gatica-Perez, "Discovering Human Places of Interest from Multimodal Mobile Phone Data," *Proc. Ninth Int'l Conf. Mobile and Ubiquitous Multimedia (MUM),* pp. 12:1-12:10, 2010.
[30] M. Ye, D. Shou, W. Lee, P. Yin, and K. Janowicz, "On the Semantic Annotation of Places in Location-Based Social Networks," *Proc. 17th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '11),* pp. 520-528, 2011.
[31] L. Liao, D. Fox, and H. Kautz, "Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields," *Int'l J. Robotics Research,* vol. 26, no. 1, pp. 119-134, 2007.
[32] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An Empirical Study of Geographic User Activity Patterns in Foursquare," *Proc. Fifth Int'l AAAI Conf. Weblogs and Social Media (ICWSM '11),* pp. 570-573, 2011.
[33] C. Song, T. Koren, P. Wang, and A. Barabási, "Modelling the Scaling Properties of Human Mobility," *Nature Physics,* vol. 6, no. 10, pp. 818-823, 2010.
[34] E. Pas, "The Effect of Selected Sociodemographic Characteristics on Daily Travel-Activity Behavior," *Environment and Planning A,* vol. 16, no. 5, pp. 571-581, 1984.
[35] X. Lu and E. Pas, "Socio-Demographics, Activity Participation and Travel Behavior," *Transportation Research Part A: Policy and Practice,* vol. 33, no. 1, pp. 1-18, 1999.
[36] N. Eagle, A. Pentland, and D. Lazer, "Inferring Friendship Network Structure by Using Mobile Phone Data," *Proc. Nat'l Academy of Sciences of USA,* vol. 106, no. 36, pp. 15274-15278, 2009.
[37] A. Liaw and M. Wiener, "Classification and Regression by Randomforest," *R News,* vol. 2, no. 3, pp. 18-22, 2002.
[38] J. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. Do, O. Dousse, J. Eberle, and M. Miettinen, "The Mobile Data Challenge: Big Data for Mobile Computing Research," *Proc. Mobile Data Challenge (MDC) Workshop,* 2012.

Trinh Minh Tri Do received the PhD degree in computer science from Pierre and Marie Curie University, Paris, France, in 2010. He is a postdoctoral researcher at Idiap Research Institute, Switzerland. His research interests include machine learning, optimization, pattern recognition, and ubiquitous computing. His current work focuses on analyzing human and social behavior from large-scale mobile phone data.

Daniel Gatica-Perez received the PhD 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 the Swiss Federal Institute of Technology in Lausanne (EPFL). His research develops computational models, algorithms, and systems for sensing and analysis of human and social behavior from sensor data. He has served as an associate editor of the *IEEE Transactions on Multimedia*, *Image and Vision Computing*, *Machine Vision and Applications*, and the *Journal of Ambient Intelligence and Smart Environments*. He is a member of the IEEE.