A review of urban computing for mobile phone traces

Citation

As Published
http://dx.doi.org/10.1145/2505821.2505828

Publisher
Association for Computing Machinery

Version
Author's final manuscript

Accessed
Thu Dec 13 15:02:26 EST 2018

Citable Link
http://hdl.handle.net/1721.1/81784

Terms of Use
Creative Commons Attribution-Noncommercial-Share Alike 3.0

Detailed Terms
http://creativecommons.org/licenses/by-nc-sa/3.0/
A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities

Shan Jiang
Department of Urban Studies and Planning
Massachusetts Institute of Technology, Cambridge, USA
shanjiang@mit.edu

Gaston A. Fiore
Department of Aeronautics and Astronautics
Massachusetts Institute of Technology, Cambridge, USA
gafiore@mit.edu

Yingxiang Yang
Department of Civil and Environmental Engineering
Massachusetts Institute of Technology, Cambridge, USA
yyang@mit.edu

Joseph Ferreira, Jr.
Department of Urban Studies and Planning
Massachusetts Institute of Technology, Cambridge, USA
jfi@mit.edu

Emilio Frazzoli
Department of Aeronautics and Astronautics
Massachusetts Institute of Technology, Cambridge, USA
frazzoli@mit.edu

Marta C. González
Department of Civil and Environmental Engineering
Massachusetts Institute of Technology, Cambridge, USA
martag@mit.edu

1. INTRODUCTION

The emergent field of Urban Computing seeks to develop computational solutions that make cities more livable, more efficient, and better positioned for the centuries ahead [2]. An important aspect in this endeavor is to have good reliable estimates of when and how the millions of individuals that cohabit a metropolis use their facilities. These daily set of individual choices are very diverse and difficult to infer in urban populations. One reason for this difficulty lies in the stochasticity of the options for activity types, travel modes, routes, sequences, and trip purposes that an individual can make in a given city. However, despite some degree of change and spontaneity, human mobility is, in fact, characterized by a deep-rooted regularity that allows us to detect predictable trends of urban dynamics [43, 42, 19, 15, 4].

Increasing storage capacity and processor clouds make possible to capture petabytes of digital traces from individual activities worldwide. Internet usage, credit card transactions, GPS-equipped vehicles, subway smart cards, among others, save in the cloud our time-stamped coordinates every time we use them [28, 14, 38]. But few things are better sensors of our daily whereabouts than our mobile phones [27]. A mobile phone tracks our location every time we text, call or web browse; and even passively when it communicates to the cellular network access points. The recently improved ability to capture, store, and understand massive amounts of data is changing the methods for inferring human behavior [13]. As our data collection grows, so will the opportunity to find better methods to interpret and transmit the data in a world where people and machines are more interconnected.

The dynamics of a population’s daily movement is a complex system; still, there are several non-trivial features that have been measured independently of the specific details of the urban group. These features are called universal in analogy with reproducible phenomena that appear in the natural sciences. Basic and common mechanisms are responsible for the presence of each of those ubiquitous features in systems governed by human activity. One widespread example that uses these universal approaches includes gravity-like models governed by human activity. One widespread example that uses these universal approaches includes gravity-like models to estimate the aggregate statistic in mobility and migration among populations. It has been proved that a simple stochastic process can capture local mobility decisions that help us derive analytically commuting and mobility fluxes, requiring as an input-only information on the distribution of population and facilities [41] without details about individual demographics, socioeconomic or activity types. The effectiveness of this simplified model stems from the high correlation that exists among the aforementioned distribu-
tions and the resulting production and attraction of trips in diverse populations.

The findings of other kinds of essential characteristics in urban mobility serve as a powerful way to convert passive data into useful models that help city planning. In this work, we will present methods that capture generalizable patterns not in aggregated but in individual trips. Our goal is two-fold. First, we will review and illustrate some of the ubiquitous findings in human mobility, as captured by mobile phone data or travel surveys, introducing the methodology to treat the data. Second, we will present the current computational challenges involved in treating these data for inferring trip patterns and routes.

The first universal feature that we will explore is the presence of preferential returns to visited locations mixed with the exploration of new ones [42, 43]. The frequent return to previously visited locations is captured by the average increase in the number of visited places over time as a result of the exploration behavior to seek for new locations mixed with the tendency for revisiting locations [42, 19]. General findings for individual urban motion have to contain these two principles that govern human mobility. A particular challenge in this paper is how to reconcile previous findings observed at more aggregate spatial scales, such as use of subway stations [19] or mobile phone towers, with triangulated mobile phone data sets containing thousands of noisy coordinates per individual user [42, 29].

The organization of daily trips have revealed ubiquitous configurations that can be expressed as daily networks with nodes representing locations and directed edges representing trips. The same distribution of trip configurations has been found in different cities, and measured by both travel surveys and mobile phone data [39]. Individuals make daily trips to five or fewer locations using only 17 of the more than 1 million possible network configurations. The basic mechanism generating these networks is the circadian rhythm of our daily movement and a perturbation factor expressed in a hidden Markov model. This factor implies that once individuals are engaged in a single flexible activity that lasts at least 30 minutes, they are 10 times more likely to engage in an additional flexible activity that day, compared with those people who have not yet left a fixed activity, such as the workplace or home. The prevalence of the 17 trip configurations indicates that they represent "motifs", which are network patterns occurring with such frequency that the statistical probability of their random occurrence is negligible. The presence of motifs indicates a basic principle that can be used to infer activity models of daily trip chains. Here we show how to detect the stay points from triangulated mobile phone records that give rise to the daily motifs.

In the paper, we analyze triangulated mobile phone record sets for the Boston metropolitan area as a demonstration. In Sections 2 and 3, we present the data characteristics and the required computational methods to extract the two known universal features of individual trips (i.e., the explorations combined with preferential returns), as well as the daily mobility motifs compared with the travel survey data for Boston metropolitan area [32]. In Section 4, we discuss the methods and current challenges of how to combine the extracted urban mobility features with land use information to infer activities and types of destinations associated with these trips. In Section 5, we present the methods to match these trips to five or fewer locations using only 17 of the more than 1 million possible network configurations. The basic mechanism generating these networks is the circadian rhythm of our daily movement and a perturbation factor expressed in a hidden Markov model. This factor implies that once individuals are engaged in a single flexible activity that lasts at least 30 minutes, they are 10 times more likely to engage in an additional flexible activity that day, compared with those people who have not yet left a fixed activity, such as the workplace or home. The prevalence of the 17 trip configurations indicates that they represent "motifs", which are network patterns occurring with such frequency that the statistical probability of their random occurrence is negligible. The presence of motifs indicates a basic principle that can be used to infer activity models of daily trip chains. Here we show how to detect the stay points from triangulated mobile phone records that give rise to the daily motifs.

In the paper, we analyze triangulated mobile phone record sets for the Boston metropolitan area as a demonstration. In Sections 2 and 3, we present the data characteristics and the required computational methods to extract the two known universal features of individual trips (i.e., the explorations combined with preferential returns), as well as the daily mobility motifs compared with the travel survey data for Boston metropolitan area [32]. In Section 4, we discuss the methods and current challenges of how to combine the extracted urban mobility features with land use information to infer activities and types of destinations associated with these trips. In Section 5, we present the methods to match these trips to five or fewer locations using only 17 of the more than 1 million possible network configurations. The basic mechanism generating these networks is the circadian rhythm of our daily movement and a perturbation factor expressed in a hidden Markov model. This factor implies that once individuals are engaged in a single flexible activity that lasts at least 30 minutes, they are 10 times more likely to engage in an additional flexible activity that day, compared with those people who have not yet left a fixed activity, such as the workplace or home. The prevalence of the 17 trip configurations indicates that they represent "motifs", which are network patterns occurring with such frequency that the statistical probability of their random occurrence is negligible. The presence of motifs indicates a basic principle that can be used to infer activity models of daily trip chains. Here we show how to detect the stay points from triangulated mobile phone records that give rise to the daily motifs.

The next step is to identify stay-regions from stay points since different stay-points identified from one user's several different trajectories may refer to a same location, but these stay-points' coordinates are unlikely to be exactly the same. We use a grid-based clustering method to cluster stay-points to get stay-regions. As shown by Zheng et. al [47] the advantage of the grid-based clustering method over the k-means algorithm and the density-based OPTICS clustering algorithm is that it can constrain the output cluster sizes, which is desirable when we know that each location should have a bounded size and the accuracy of the records is within a certain range. In this study the maximum stay region size is set to d = 300m to approximate the area that might likely be traversed on foot as part of an urban activity. The procedure to perform grid-based clustering is to first divide the entire region into rectangular cells of size d/3. Next to map all the stay-points to each cell. Then iteratively merge the unlabeled cell with the maximum stay-points and its unlabeled neighbors to a new stay-region. Once a cell is assigned to a stay region, it is marked as labeled. For the detailed algorithm please refer to [47]. In Fig. 2.1, the three
stay-points are clustered to one stay-region \( r_1 \).

3. UNIVERSAL PATTERNS OF INDIVIDUAL MOBILITY

3.1 Exploration and Preferential Returns

Several ubiquitous characteristics of individual human trajectories have been found [4, 15, 42, 5], most of which are using tower level cell phone records. One important aspect is to measure the degrees of predictability of human mobility. Previous studies [43] found that, on average, 70% of the time the most visited location coincides with a user’s actual location at a given time of the day. The distributions of travel distance \( (P(r)) \), inter-activity time \( (P(t)) \), radius of gyration \( (r_g) \), location visiting frequency \( (f_k) \) and location exploration probability show that human trajectories present statistical regularities that can be related via scaling laws. The location visiting frequency usually conforms to Zipf’s law [50]. This implies a hierarchical ranking in our visitation patterns that relates to the exploration and preferential return to certain locations, which is an ubiquitous mechanism in human mobility [42]. The more locations a person has visited, the less likely s/he is going to visit a new location, or in other words, the more likely s/he is going to return to a previously visited location. This probability is proportional to the previous visiting frequency of that location.

With data at the finer granularity level available, we would like to test whether these scaling laws still hold on the stay locations extracted as described in the previous section.

We begin our exploration by calculating the users’ mobility regularity \( R(t) \), which is defined as the probability of finding the user in/ her/his most visited location at hourly interval in a week. Fig. 3.1(a) shows that, under finer granularity, there is a regularity for all users in a week. The average regularity drops from 70% (in tower level data) to 64% measured at the level of stay cells. As expected, the regularity is still higher during night and lower during the day. It’s also higher during weekends. The average numbers of distinct locations visited over time, \( S(t) \), follows the following trend:

\[
S(t) \sim t^\mu,
\]

where \( \mu = 0.6 \pm 0.02 \) for tower-level cell phone data [42].Fig. 3.1(b) shows how \( S(t) \) vs. \( t \) changes for user groups that visited a different number of locations during the two-month period. As can be expected, user groups that visited more locations in the two-month period have higher slopes. For group \( s: 80–100 \), \( \mu = 0.66 \) while for group \( s: 20–40 \), \( \mu = 0.41 \). For all the users \( \mu = 0.59 \), which agrees with previous findings.

Next, we measure the visiting frequency \( f_k \) of the \( k \)th most visited locations, which follows the shape: \( f_k \sim k^{-\xi} \), with \( \xi = 1.04 \) (Fig. 3.1(c)), which is slightly smaller than the one observed with higher granularity data, of \( \xi = 1.2 \). The multi-scale effects still deserve more thorough investigation and may be sensitive to the scale of stay regions.

Fig. 3.1(d) shows that if a user returns to a previously visited location, the probability \( \Pi \) to return to that location is proportional to that location’s previous visiting frequency \( f \). This evidence again supports the exploration and preferential return mechanism in human mobility.

3.2 Daily Motifs

Individual daily mobility is well described by activity chains that include the start time, the end time, and the location of each activity within a day. Activity chains are usually obtained from travel survey data, which is accurate but with low sampling rate (around 1% of total households in a metropolitan area) and usually records only one day of travel diaries per household [23]. Cell phone data has the opposite characteristics: not all stays in a day can be captured in the cell phone data, but it has much longer periods of sampling over larger fractions of the population—in this study it contains 20% of the population in Boston over a 2 month period. So a question arises: can the larger volume and longer periods of observation make up for the less accuracy in the cell phone data? We find that this is the case if the data is filtered in a proper way.

Larger volume of data enables us to filter out the noise and select only users with enough information for detecting daily trip chains. The sampling method requires: (1) To select only frequent cell phone users with enough records; (2) To remove pass-by points which are only used during travel; (3) To eliminate signal transitions between neighboring locations; (4) To detect individual trips only for days with at least 7 identifiable time-slot locations (a day is divided into 48 half hour slots); (5) To overcome the small number of night calls, the location which is visited most frequently during all nights between 12 am and 6 am of a single user, is assigned to be the user’s home location. For a detailed description of the filtering process please refer to our previous study [39].

For each sampled individual trajectory, we construct a travel network in which nodes represent the visited stay-regions and directed edges stand for trips between them. We count the statistically significant configurations in the data sets, which are called motifs, adopting the term from network science [31]. This notion is similar to the notion of activity chains. The difference is that here we distinguish locations by the coordinates of the stay-region rather than the functionality of the location such as home, work, etc. This formulation is suitable for passive observations of mobile phone data in which we have higher degrees of uncertainty in inferring trip purposes.

In a previous study [39], we measured that over 90% of the identified daily mobility networks can be described with only
17 different motifs. We test this finding here by extracting motifs for both cell phone users in our dataset and from the 2010/2011 Massachusetts travel survey [32], which contains 37,023 people’s travel diary over one day on a rolling basis. Fig. 3.2 shows the distributions of the 17 motifs which are similar for the two different data sources, and they also agree with the previous findings measured in Chicago and Paris. This result shows the validity of the proposed method for triangulated cell phone data, which presents a good alternative for analyzing daily human mobility patterns and complements expensive surveys.

Figure 3.2: Frequent daily motifs. The 17 most frequent motifs account for over 90% of the measured daily trips. The distributions of the 17 motifs extracted from cell phone data and the Massachusetts travel survey are similar, and also conform to previous findings in Paris and Chicago [39].

4. INFERRING INDIVIDUAL ACTIVITIES AND TRAVEL

To make the million users’ mobile phone traces useful for urban land use, community planning and transportation planning, it is crucial to answer one of the most important questions: “What are people doing in space and time?” [1, 6, 16, 30, 24, 25, 26]. This question includes inferring the spatiotemporal activities that people engage in, and their travel (e.g., trip chaining, and road usage, etc.) induced by the needs of pursuing activities [35].

In order to infer the types and patterns of activities of anonymous individuals, by learning their historical presence in space and time and characteristics of their destinations (e.g., land use, points-of-interest (POIs)), we need to address several challenges presented by the mobile phone records (for billing purposes) as opposed to by GPS data for which many algorithms and methods have been developed to study human behavior [47, 17]. First, mobile phone data are perceived with indefinite gaps in space and time, while GPS data are recorded with a high frequency such that they can be treated as continuous trajectories. Second, the locational accuracy of mobile phone data is lower than the pinpointed GPS traces (depending on the technologies [36]).

In this section, we present a class of algorithms that are tailored to address the distinct characteristics of the mobile phone data (triangulated at 200- to 300-meter accuracy level) such that we can use the filtered data to infer human activities and their travel in space and time. In contrast to the grid-based algorithm presented in Section 2, the one presented here is designed to exploit the maximum spatial accuracy possible. Comparing these two classes of data filtering methods is beyond the scope of this paper and will be presented elsewhere.

4.1 Extracting Stay, Pass-by and Potential Stay Areas

For the purpose of extracting individuals’ whereabouts from phone records, including their stationary stay locations (so as to infer their activity types) and their moving pass-by locations (so as to infer their travel path and road usage), we employ a method inspired by Hariharan and Toyama’s study [17]. We demonstrate this process of data filtering in Figure 4.1, and discuss details as follows.

Let sequence $D_i = (d_i(1), d_i(2), d_i(3), ..., d_i(n_i))$ be the observed data for a given anonymous user $i$, where $d_i(k) = (t(k), x(k), y(k))$ for $k = 1, ..., n_i$. $t(k)$, $x(k)$, and $y(k)$ are the time, longitude, and latitude of the $k$-th observation of user $i$. First, we extract points $d_i(k)$ that are spatially close (i.e. within roaming distance of 300 meters) to their subsequent observations, say, $d_i(k + 1), d_i(k + 2), ..., d_i(k + m)$. To reduce the “jumps” in the location sequence of the mobile phone data, we assume that $d_i(k), ..., d_i(k + m)$ are observed when user $i$ is at a specific location, i.e., the medoid of the set of locations $\{(x_i(k), y_i(k))\}_k \cup \{(x_i(k + m), y_i(k + m))\}$, which is denoted by $\text{Med}(\{(x_i(k), y_i(k))\}_k \cup \{(x_i(k + m), y_i(k + m))\})$.

This treatment respects the time order at first, to ignore noisy “jumps” in estimated location, but then disregards time ordering to apply the agglomerative clustering algorithm [17] to consolidate points that are close in space but may be far away in time. The points to be consolidated together form a cluster whose diameter is required to be no more than a certain threshold (set to be 500 meters). Again, we modify the observation locations to the corresponding medoids of the clusters (see Figures 4.1(a) and (b)). It turns out that by these treatments, we greatly reduce the noise in the location sequences of the mobile phone data (i.e., errors in signal triangulations).

Figure 4.1: Extracting Stay, Pass-by and Potential Stay Areas from the Phone Data for an Anonymous User in a 2-Month Period.

Second, we impose the time duration criterion on the clean data, and extract the stay locations whose durations exceed a certain threshold (set as 10 minutes) (see Figure 4.1(c)). In the presented example we extract 31 distinct stay locations from the 1,776 phone records in the 2-month period of the exhibited anonymous user. The rest of the points are called pass-by points, where we don’t observe any lengthy stays in these areas. Note that it is possible that the user might actually stay in some of these pass-by areas or areas that we don’t even observe. In these cases, information about time and location is totally or partially latent to us as we don’t observe it from the phone records. However, all the
stay locations frequently visited by the user ought to be extracted from the mobile phone data, if the observation period is long enough.

Third, we treat the distinct stay locations obtained from the last phase as the user’s destinations, and flag the pass-by points that collocate with any of the destinations as potential stays (see Figure 4.1(c) and 4.1(d), in which some pass-by’s are converted into potential stays).

In Figure 4.2, we demonstrate the difference in inferring the individual’s travel when the stay points, potential stay points, and pass-by points are gradually detected. We can see that the inference is more accurate when including pass-by points of the individual’s travel. Note that including potential stay points may also change the type of individual daily mobility motif patterns, because we may capture activities of less importance in terms of spent time.

One of the main challenges here is that none of the data sources (travel survey or phone data) pinpoint locations of the trip destinations, but areas surrounding these precise locations. For the travel survey data, locational information of trip destinations are usually the centroids of administrative zones (e.g., traffic analysis zones, postal zones, or census tract/block group/block) for privacy consideration and/or the ease of administrative efforts in conducting the travel survey. For example, in the 2010/2011 Massachusetts travel survey (which covers the Boston metropolitan area), the reported locations of trip destinations are the centroids of census blocks (whose areas range from a few square meters up to around ten square kilometers). However, land use within the same zone (e.g., census block) can be mixed, depending on the size and location of the zone. This adds uncertainty in predicting individuals’ activities conditional on land use information from the survey. Similar issues exist in the mobile phone data to certain extent, depending on the accuracy and resolution of the spatially triangulated locations of the anonymous phone users. Nonetheless, finding smart techniques to link the association rules of semantic-enriched land use and POI information of the diverse areas that individuals visit is an open challenge for estimating the activity types that individuals engage in [45, 46]. And this information is extremely valuable to put mobile phone data into the service of urban and transportation planning.

In Figure 4.3, we illustrate in one example the situation we must address when inferring activity types given an anonymous mobile phone user’s records on a Saturday.

4.2 Inferring Human Activities

The individual daily mobility motif analysis presented in Section 3.2 provides a framework to further analyze the types of activities (e.g., home, work, school, shopping, recreation, social, etc.) that individuals engage in at different destinations (i.e., nodes in the motifs, see Figure 4.2(d)). In traditional survey data, activity types/purposes are revealed by individuals who answer the travel diaries; whereas in the mobile phone data (for billing purposes), activity types at certain destinations are not revealed and latent. Due to the constraint of land use and availability of business establishments in certain economic sectors (i.e., points-of-interest), individuals’ activity types and destination characteristics are closely related in general. Table 1 shows probabilities of various types of human activities given land use types calculated from the 2010/2011 Massachusetts travel survey data and land use data for the Boston metropolitan area.

With the emerging availability of semantic-enriched digital geographic data on land use and POIs, we can cross reference the spatial information and characteristics of destinations that anonymous individuals visit, which allows us to build probabilistic models to infer activity types at different destinations in space and time. The probability that individual i commits activity a at time t conditional on her/his destination information and daily mobility motif is written as $f(a_{it} = a|t, d_i, m_{id})$, where $d_i$ is a vector containing various destination information such as the location, land-use type, population density, etc., and $m_{id}$ is the motif of individual $i$ in day $d$. This probability can be estimated from the survey data, where both the activity type and destination information are available, for example using multinomial logit model [3], applied to activity inference using mobile phone data.

One of the main challenges here is that none of the data sources (travel survey or phone data) pinpoint locations of the trip destinations, but areas surrounding these precise locations. For the travel survey data, locational information of trip destinations are usually the centroids of administrative zones (e.g., traffic analysis zones, postal zones, or census tract/block group/block) for privacy consideration and/or the ease of administrative efforts in conducting the travel survey. For example, in the 2010/2011 Massachusetts travel survey (which covers the Boston metropolitan area), the reported locations of trip destinations are the centroids of census blocks (whose areas range from a few square meters up to around ten square kilometers). However, land use within the same zone (e.g., census block) can be mixed, depending on the size and location of the zone. This adds uncertainty in predicting individuals’ activities conditional on land use information from the survey. Similar issues exist in the mobile phone data to certain extent, depending on the accuracy and resolution of the spatially triangulated locations of the anonymous phone users. Nonetheless, finding smart techniques to link the association rules of semantic-enriched land use and POI information of the diverse areas that individuals visit is an open challenge for estimating the activity types that individuals engage in [45, 46]. And this information is extremely valuable to put mobile phone data into the service of urban and transportation planning.

In Figure 4.3, we illustrate in one example the situation we must address when inferring activity types given an anonymous mobile phone user’s records on a Saturday.

![Figure 4.2: Comparison of Potential Travel Paths with and without Pass-by and Potential Stay Points for an Anonymous User (Same as in Figure 4.1) in One Day](image)

**Figure 4.2:** Comparison of Potential Travel Paths with and without Pass-by and Potential Stay Points for an Anonymous User (Same as in Figure 4.1) in One Day

**4.2 Inferring Human Activities**

The individual daily mobility motif analysis presented in Section 3.2 provides a framework to further analyze the types of activities (e.g., home, work, school, shopping, recreation, social, etc.) that individuals engage in at different destinations (i.e., nodes in the motifs, see Figure 4.2(d)). In traditional survey data, activity types/purposes are revealed by individuals who answer the travel diaries; whereas in the mobile phone data (for billing purposes), activity types at certain destinations are not revealed and latent. Due to the constraint of land use and availability of business establishments in certain economic sectors (i.e., points-of-interest), individuals’ activity types and destination characteristics are closely related in general. Table 1 shows probabilities of various types of human activities given land use types calculated from the 2010/2011 Massachusetts travel survey data and land use data for the Boston metropolitan area.

With the emerging availability of semantic-enriched digital geographic data on land use and POIs, we can cross reference the spatial information and characteristics of destinations that anonymous individuals visit, which allows us to build probabilistic models to infer activity types at different destinations in space and time. The probability that individual $i$ commits activity $a$ at time $t$ conditional on her/his destination information and daily mobility motif is written as $f(a_{it} = a|t, d_i, m_{id})$, where $d_i$ is a vector containing various destination information such as the location, land-use type, population density, etc., and $m_{id}$ is the motif of individual $i$ in day $d$. This probability can be estimated from the survey data, where both the activity type and destination information are available, for example using multinomial logit model [3], applied to activity inference using mobile phone data.

One of the main challenges here is that none of the data sources (travel survey or phone data) pinpoint locations of the trip destinations, but areas surrounding these precise locations. For the travel survey data, locational information of trip destinations are usually the centroids of administrative zones (e.g., traffic analysis zones, postal zones, or census tract/block group/block) for privacy consideration and/or the ease of administrative efforts in conducting the travel survey. For example, in the 2010/2011 Massachusetts travel survey (which covers the Boston metropolitan area), the reported locations of trip destinations are the centroids of census blocks (whose areas range from a few square meters up to around ten square kilometers). However, land use within the same zone (e.g., census block) can be mixed, depending on the size and location of the zone. This adds uncertainty in predicting individuals’ activities conditional on land use information from the survey. Similar issues exist in the mobile phone data to certain extent, depending on the accuracy and resolution of the spatially triangulated locations of the anonymous phone users. Nonetheless, finding smart techniques to link the association rules of semantic-enriched land use and POI information of the diverse areas that individuals visit is an open challenge for estimating the activity types that individuals engage in [45, 46]. And this information is extremely valuable to put mobile phone data into the service of urban and transportation planning.

In Figure 4.3, we illustrate in one example the situation we must address when inferring activity types given an anonymous mobile phone user’s records on a Saturday.

![Figure 4.3: Land Use of (Potential) Stay Areas of the Anonymous User in One Day (the Same as in Figure 4.2)](image)

**Figure 4.3:** Land Use of (Potential) Stay Areas of the Anonymous User in One Day (the Same as in Figure 4.2)

Figures 4.3(a) demonstrates the stay areas, pass-by areas and potential stay areas en route in a day extracted from an anonymous phone user’s record data. Figure 4.3(b), (c),
(d) zoom into the areas of stay and potential stay, and cross reference the land use information of the area. A great alternative to land use data are on-line POI information from user-generated platforms (such as Yahoo!, Yelp, etc.). Incorporating the business categories of on-line POIs to enrich land use information is also feasible [37]. From figure 4.3, we see that the anonymous user left destination A (presumably “home”) after 9:52am (phone usage detected for 22 minutes since 9:30am), and the land use types include residential, commercial, and urban public/institutions. Around 9:57am, s/he stopped for 13 minutes at destination B, where land uses include urban public/institutions, and commercial, and possible for a quick shopping/errand. Around 10:50am s/he showed up at place C, where the majority of land use is commercial and urban public/institutions. Her/his phone use at place C was prompt. But since place C is among her/his historical destinations, s/he could have potentially stayed at C for activities. Since 11:52am, s/he stayed for 116 minutes in destination D, where the majority of land use is residential (presumably s/he could be visiting friends/family members). S/he appeared again at destination D at around 14:59pm, and stayed for another 30 minutes. At some point in the afternoon/evening s/he returned to destination A, where s/he resided, at some point in the night.

5. MATCHING CELL PHONE DATA TO SPATIAL NETWORKS

In order to map previously discussed individuals’ daily motifs, activity sequences and locations to spatial networks, so as to address transportation management and planning issues (such as mode choice, traffic congestion, etc.), cell phone data have to be incorporated into the spatial networks within which the flow of people is restricted. In this section we first argue for the additional insights that can be gained by analyzing cell phone data within spatial networks. We then discuss a software architecture based on state-of-the-art open source projects that can be used for the geospatial analysis of phone data within spatial networks.

5.1 The Necessity of Data Analysis Within Spatial Networks

The network science and human mobility community has so far mostly analyzed cell phone datasets in the Euclidean space. The Euclidean space can be suitable for certain kinds of analyses, but makes it hard to link the human dynamics insights, which can be gained from analyzing cell phone data, to the road network systems in a city that give rise to these dynamics. A holistic study of human mobility and transportation systems will enable the development of technologies that render cities truly intelligent. In order to pursue this goal in the context of our work, it is necessary to analyze cell phone data within spatial networks [11, 18], in particular time-dependent spatial networks [8, 12, 33].

On the one hand, large-scale cell phone data matched to the road network can be used to uncover traffic patterns and road usage. On the other hand, since cell phone data contain information on a massive amount of individuals (usually in the millions) and can be used to infer home locations of the anonymous users, they can be used to infer sources of traffic together with destination information. For example, Wang et al. [44] used phone data at the cell tower level to find that major traffic flows in congested roads in both the Boston and San Francisco Bay Areas are generated by very few driver sources. These results are directly useful for urban transportation policy making to reduce traffic congestion by targeting specific driver sources. One potential disadvantage of the approach presented in [44] is that it used proprietary software (such as TransCAD, a specialized GIS software in transportation system modeling and planning) to conduct the spatial analysis, which may limit the reach of such studies to broader contexts as it would require the purchase of a software license. It would thus be desirable to conduct the analysis of cell phone data within spatial networks using open source software.

5.2 Incorporating Cell Phone Data to Spatial Networks

Cell phone datasets may have large sizes and consequently it becomes imperative to implement algorithms paying particular attention to speed and scalability. In order to achieve this goal, there is a need for a geospatial software foundation that is fast, reliable, and scalable, upon which the necessary algorithms are implemented. At the same time, the end user needs to be able to extend and modify the software to be able to add new functionality. Some alternatives within the open source geospatial software ecosystem seem particularly suitable to satisfy these requirements [40]. We have identified as suitable options PostgreSQL extended with PostGIS to store

<table>
<thead>
<tr>
<th>Table 1: Probability of Human Activities Conditional on Land Use Types in Boston Metropolitan Area</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Activity</strong></td>
</tr>
<tr>
<td>a</td>
</tr>
<tr>
<td>b</td>
</tr>
<tr>
<td>c</td>
</tr>
<tr>
<td>d</td>
</tr>
<tr>
<td>e</td>
</tr>
<tr>
<td>f</td>
</tr>
<tr>
<td>g</td>
</tr>
<tr>
<td>h</td>
</tr>
<tr>
<td>i</td>
</tr>
<tr>
<td>j</td>
</tr>
</tbody>
</table>

† Activity types: a. Working at home (for pay), and all other home activities; b. Work/job, all other activities at work, volunteer work/activities, and work business related; c. Attending class, and all other school activities; d. Changed type of transportation, drop off passenger from car, pick up passenger from car, traveling, service private vehicle (gas, oil lube, etc.), and loop trip; e. Routine shopping (groceries, clothing, convenience store, HH maintenance), and shopping for major purchases or specialty items (appliances, electronics, new vehicle, major household equipment, purchase of land use types: 1. Cropland, pasture, forest, woodland, open land, orchard, nursery, etc.; 2. Multi-family, high/medium/low/very low density residential; 3. Commercial; 4. Mining, industrial, powerline/utility; 5. Transitional; 6. Transportation; 7. Waste disposal, junkyard; 8. Urban public/ institutional; 9. Cemetery; 10. Participation/Spectator/Water-based recreation, golf course, and marina.

† Data Sources: 2010/2011 Massachusetts travel survey data (MassDOT), and the Massachusetts land use data (MassGIS).
and manipulate the cell phone data and the spatial network, and pgRouting to add geospatial routing capabilities to the database. Beyond the availability of numerous features that make it possible to conduct various analyses, these open source tools provide a great foundation upon which to implement extensions related to time dependency [34, 10, 9] and multimodality [7] for example.

PostgreSQL is a powerful, open source object-relational database system [22]. It is fully ACID compliant, has full support for foreign keys, joins, views, triggers, and stored procedures. It includes most SQL:2008 data types and also supports storage of binary large objects, including pictures, sound, or video. It has native programming interfaces for Python, which is our programming language of choice for algorithm implementation and analysis. It is highly scalable in terms of the quantity of data it can manage, with an unlimited maximum database size and a 32 TB maximum table size. The key component that makes PostgreSQL suitable for geospatial analysis is PostGIS.

PostGIS is a spatial database extender for PostgreSQL that adds support for geographic objects, allowing PostgreSQL to be used as a spatial database for geographic information systems (GIS) [21]. PostGIS adds support for geographic objects allowing location queries to be run in SQL. It adds extra types (geometry, geography, raster and others) as well as functions, operators, and index enhancements that apply to these spatial types. This PostgreSQL/PostGIS combination results in a fast, feature-rich, and robust spatial database management system. Navigation for road networks requires complex routing algorithms that support turn restrictions and ideally time-dependent attributes. Towards this end, geospatial routing can be done at the database level with pgRouting.

pgRouting is a library that extends PostgreSQL/PostGIS to support geospatial routing and adds routing functionality to the database [20]. It provides a variety of tools for shortest path search, including functions for Shortest Path Dijkstra, Shortest Path A-Star, Shortest Path Shooting-Star (routing with turn restrictions), Traveling Salesman Problem (TSP), and Driving Distance calculation. The key value of pgRouting is that it allows these high-level functions to run at the database level. Furthermore, the database routing approach has two main advantages that make it suitable for eventually incorporating time-dependency into the problem. First, any data changes in the database will be taken into account instantaneously by the routing engine. Second, the “cost” parameter can be dynamically calculated through SQL and its value come from multiple fields or tables.

PostgreSQL/PostGIS are very useful not only to perform routing through a road network, but also in concert with the python interface capabilities—to manage the raw data (after some initial data filtering and processing) and to handle the rest of the algorithms for converting the raw data into manageable individuals’ daily traces, activities, and trip destinations, etc. More specifically, the raw cell phone data and the corresponding local road network can be stored in a PostgreSQL/PostGIS database. pgRouting can then be used to perform geospatial routing and enable the execution of analyses related to road usage as in [44] but based on an open source software architecture. The different algorithms to process the data as have been presented in the paper can be implemented in Python and use the open-source Python module PyGreSQL to interface to the PostgreSQL database. As the data gets refined, the smaller subsets can be stored in new tables and data processing can proceed from these subsets.

Figure 5.1 gives a preview of the advantages of analyzing cell phone data in a road network. In this figure we can see the paths along stay, potential stay, and pass-by points for the same person as in Figures 4.1 and 4.2 for three different days. These paths give an approximate indication of the mobility of a person throughout time while taking into account the dynamic constraints of the spatial networks (in this case the road network) to which the human dynamics are subjected. Notice how the home location of the individual as well as nearby locations repeat spatially among the three subsequent days.

6. CONCLUSION

As mobile phone devices become ubiquitous sensors in people’s daily life, accompanied by the fast advancing ICT technologies that make available fine-grained location information for millions of anonymous phone users, great opportunities exist for us to fully understand patterns and mechanisms of human mobility, activities, and their relationship with the urban environment. These opportunities open the door to an new era, when researchers from interdisciplinary fields can re-examine ways that human interact with the environment and mechanisms that drive these dynamics. While these are fundamental questions to answer for sustainable urban future, we still need to address several challenges in fully embracing the opportunities.

In this paper, we review, extend, and illustrate the “Big Data” processing that can translate voluminous urban sensing (such as cellphone traces) into parsimonious trip chains, activity sequences, and travel paths that are more suitable for use by land use and transportation planners. We utilize three classes of methods to exploit such high-resolution mobile phone data (at 200- to 300- meter accuracy level) for different purposes in urban computing applications, using Boston metropolitan area as a demonstration. First, we analyze human mobility and derive universal features revealing the predictability and scaling laws of human dynamics. By using method that extracts individuals’ phone signals into small grid cells, we find that the measured users’ mobility regularity $R(t)$ drops to 64% at the represented grid cell level compared to 70% measured at the cell tower level. Meanwhile, the probability distribution measuring an individual’s return to her/his previously visited location for all phone users in our sample also supports previous findings on the exploration and preferential return patterns in
human mobility. We then extract individual daily mobility networks, called motifs in the study, from both the phone data and travel survey data. By comparing the 17 most frequent motifs (covering 90% of the total daily travel), we find similar distributions obtained from phone and survey data.

Second, we introduce a class of algorithms to extract custom-tailored point-based fine-grain areas (at the 300-meter accuracy level) representing each anonymous individual’s stay, pass-by, and potential stay locations. We adjust parameters of the spatiotemporal constraints, according to data accuracy and prior knowledge of human behavior, to define these three categories of individual’s presence in space and time.

Our method is particularly useful to examine surrounding semantic-enriched land use types or points-of-interest information in the areas of (potential) stays so as to model the activity types engaged by individuals in their daily life. We propose a framework to infer human activity types by incorporating the extractable individual daily motifs, land use information of an individual’s (potential) stays, and time. Existing techniques (such as multinomial logit models) can be easily applied under such a framework, and will be tested in our future research. We also demonstrate that including pass-by points, and potential stay points will be very useful in inferring individuals’ travel path and road usage, which is directly linked to our third method discussed in the paper.

Finally, we present a software architecture based on state-of-the-art open source projects that can be used for managing and analyzing mobile phone data in geospatial databases. This software architecture becomes the foundation upon which we develop new algorithms to study human mobility as extracted from cell phone data within the spatial networks that constrain human dynamics. The method presented here will be very useful especially since we study human mobility from an agent-centric perspective. It also represents a flexible framework that makes it possible to consider in future work the multimodal aspect of the transportation systems in the city and go beyond the road network. In particular, the development of algorithms, which can match geographic coordinates out of cell phone data to rail, subway, or road networks and thus enable us to infer potential travel modes, is of particular interest and an avenue for future research.

With the three methods brought together in this paper, we demonstrate and evaluate state-of-the-art algorithms that can be adopted to take full advantage of triangulated mobile phone data with high spatial resolutions for urban computing and modeling applications. To summarize, such applications include extracting individual’s daily motifs (or trip chains), semantic-enriched land use information of activity destinations, activity types, and human flows in the spatial networks. As cities are growing in unprecedented speed in human history, and become more diverse as economic structures and changes over the world, these urban applications are fundamental in solving urban problems (i.e., growth and land use management, traffic congestions, efficiency and equity of public resource allocations, etc.). We showcase how an interdisciplinary approach in urban computing can be more useful and relevant for sustainable urban futures in the “Big Data” era.

7. ACKNOWLEDGMENTS

This research was funded in part by the BMW Group, the Austrian Institute of Technology, the MIT School of Engineering, the MIT Department of Urban Studies and Planning, and by the Singapore National Research Foundation (NRF) through the Singapore-MIT Alliance for Research and Technology (SMART) Center for Future Mobility (FM).

8. REFERENCES


