



Understanding Human Movement Semantics: A Point of Interest Based Approach

Ionut Trestian, Kévin Huguenin, Ling Su, Aleksandar Kuzmanovic

**RESEARCH
REPORT**

N° 7716

August 2011

Project-Team ASAP



Understanding Human Movement Semantics: A Point of Interest Based Approach

Ionut Trestian*, Kévin Huguenin†, Ling Su*, Aleksandar
Kuzmanovic*

Project-Team ASAP

Research Report n° 7716 — August 2011 — 21 pages

Abstract: The recent availability of human mobility traces has driven a new wave of research – on human movement – with straightforward applications in wireless/cellular network algorithmic problems. However, all of the studies isolate movement from the environment that surrounds people, *i.e.* the points of interest that they visit.

In this paper we revisit the human mobility problem with new assumptions. We believe that human movement is not independent of the surrounding locations; most of the time people travel with specific goals in mind, visit specific points of interest, and frequently revisit favorite places. Points of interest are also differently spread. We study the correlation between people’s trajectories and the differently spread points of interest nearby.

More specifically, by analyzing GPS mobility traces of a large number of users located across two distinct geographical locations, we find that: *(i)* users do not particularly visit only locations that are close to them but the functional aspect of the location matters as well, *(ii)* although users in different parts of the globe exhibit different time-of-day behavior, we also find that there is a striking correlation in the frequency of visits to the basic points-of-interest categories that we define.

Key-words: Human movement; Mobility traces; GPS; Points of interest

* Department of Electrical Engineering and Computer Science, Northwestern University

† School of Computer Science, McGill University

**RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE**

Campus universitaire de Beaulieu
35042 Rennes Cedex

Mieux comprendre la sémantique des mouvements humains: Une approche par points d'intérêt

Résumé : La disponibilité récente de traces de mobilité humaine a engendré une vague de recherche sur les modèles de mobilité ayant des applications directes dans les réseaux cellulaires/sans fil. Cependant, les études menées jusqu'à présent ignorent l'environnement dans lequel se trouvent les personnes en mouvement : les points d'intérêt que ces derniers visitent et qui, par conséquent, guident leurs déplacements.

Cet article revisite le problème de la modélisation des déplacements humains en se basant sur de nouvelles hypothèses. Considérer que les déplacements humains ne sont pas indépendants des lieux environnants semble être une hypothèse raisonnable : la plupart du temps, les gens se déplacent avec un but précis en tête, visitent des lieux précis et se rendent régulièrement dans leurs endroits favoris. Ces points d'intérêt sont répartis de manière non uniforme. A partir de ces observations, on étudie les interactions entre les déplacements humains et la répartition des points d'intérêt.

Plus précisément, en analysant des traces de mobilité GPS de personnes évoluant dans deux zones géographiques distinctes du globe, il apparaît que : les gens ne basent pas le choix des lieux qu'ils visitent uniquement sur la proximité mais également sur des aspects fonctionnels de ces derniers ; si l'activité et la mobilité de gens originaires de régions distinctes présentent des motifs temporels différents, les fréquences auxquelles ils visitent certaines catégories de points d'intérêt ont des similarités frappantes.

Mots-clés : Mouvement humain ; Traces de mobilité ; GPS ; Points d'intérêt

1 Introduction

Human mobility has been studied extensively in the past several years. This is mainly due to the implications that the results could have in fields such as urban planning, disease prevention, mobile advertising, and mobile infrastructure placement. As such, researchers have tried to understand statistical properties and have also proposed mathematical models to capture particular aspects of human movement, for example by relating to the movements of banknotes or subatomic particles¹. But people are not banknotes and they are not subatomic particles. When they move they are driven by concrete goals such as buying groceries, dining at a restaurant, or watching a movie. Although models (*e.g.* Lévy flights) manage to capture important aspects of human movement (such as the distribution of trip sizes), they fail to capture its *semantics*.

Furthermore, we believe that most human movement is indeed driven by certain purposes. Specific and important characteristics of human movement, which should be captured in the aforementioned models, are in fact linked to two important factors. The first factor consists of *people's daily routines* (such as visiting a coffee shop in the morning, going to work, and/or shopping or going to a restaurant in the evening). A second factor is the actual *location* of the different *points of interest* that people visit. For example, in a given region there could be many restaurants scattered over the whole area, but only several stadiums. Restaurants attract people on a daily basis, whereas stadiums attract a large number of people only on certain occasions. The combination of these two factors (daily routines, and point-of-interest placement) could therefore explain the different characteristics of human movement, which were only observed in previous studies.

Several applications can benefit from the understanding of the above characteristics. With the current advent of mobile devices such as smartphones and tablets, understanding human interactions with points of interest can have an impact on mobile infrastructure placement, transfer scheduling algorithms, mobile advertising, mobile social networking (to name just a few). Furthermore, understanding how people interact with the points-of-interest around them could help urban planners.

In this paper we answer the following research questions. What are the actual semantics of human movement? What are the locations that people visit, with what frequency, and on what time scales? What is the influence of the distribution of these locations on human movement? We address the above questions by extensively studying two movement traces from two different geographic regions. The traces are from users carrying GPS devices inside mobile phones (155 users from China and 4,429 users from the United States). We enhance the above traces with categorized points of interest from GPS devices. Furthermore, because of the geo-diverse nature of our traces we find cultural differences with respect to people and the points of interest near them. For example we found that the Chinese users have a smaller movement span compared to their American counterparts (probably because people in the United States usually live far from their work places).

To the best of our knowledge, we are the first to systematically study the relationship between human mobility and points of interest on such a scale. This is our main contribution. One of our key findings is strong evidence relating user affinities, on different time scales, toward certain location types such as restaurants, shops, and stadiums (some locations, *e.g.* stadiums, are not located particularly close to users).

To understand the relationship that exists between human movement and the surrounding points-of-interest, we apply an energy model inspired from physics that we build over a modified kd-tree data structure. We further use this methodology to analyze three aspects. First, we

¹We are referring to the Brownian Motion Mobility Model [9], and the well known human movement study based on the one dollar bill [7].

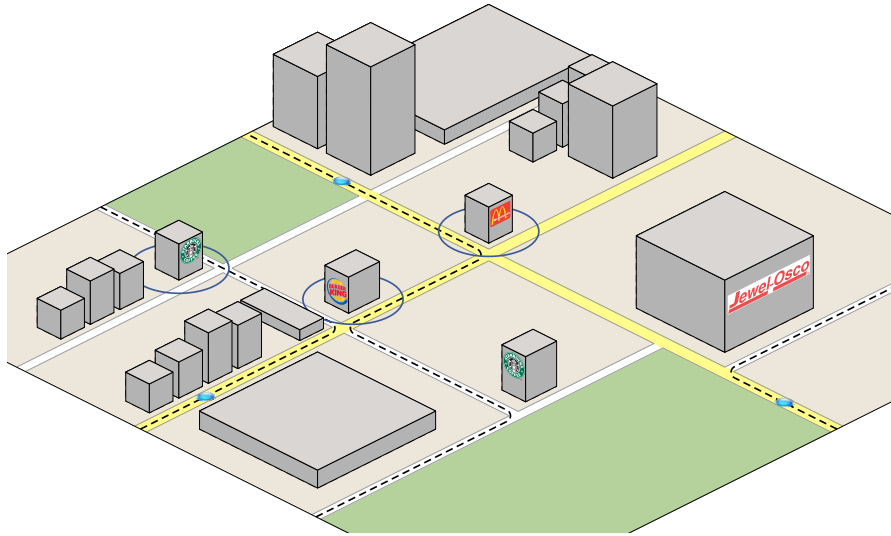


Figure 1: Overview of mobility and point-of-interest datasets.

perform user profiling in order to identify broad user interests in location types. We find that most users are interested in dining at restaurants, shopping, attending sports events, whereas they are not very interested in visiting libraries or hospitals. Second, we analyze how frequently users visit certain points of interest. We find that certain points-of-interest types attract visitors on a daily basis (e.g. restaurants), whereas others (e.g. stadiums) attract a larger number of visitors but only on occasion. Third, we analyze time-of-day effects on the frequency of users visiting points of interest. We find, for example, that the Chinese users try to optimize their time. One example of such behavior is users run errands during their lunch break.

We carried our analysis further by analyzing user interest in specific popular businesses such as Starbucks, McDonalds, and Burger King. We find that such businesses manage to attract a large number of users. Because some of the analyzed businesses carry WiFi access points, our findings could be useful to mobile providers in forming partnerships with businesses for content off-loading purposes.

This paper is organized as follows. In §2, we describe the trajectory and point-of-interest datasets that we use in our paper. At first, we analyze independently both user trajectories and the distribution of points of interest. As such, in §3.1 we perform a preliminary analysis of user trajectories, and in §3.2 an analysis on the placement and spread of different types of points of interest. We further analyze the influence of points of interest on people's trajectories. In §4 we present the interaction methodology that we use in the paper. In §5 we present the results of our study. We discuss the implications of our research in §6. We acknowledge related work in §7, and conclude our paper in §8.

2 Datasets

In order to correlate user movement with daily activities, we require two types of data: (i) *user mobility data* (obtained from users carrying GPS tracking devices in mobile phones, dashed lines in Fig. 1), and (ii) *point of interest data*—*i.e.* locations of *e.g.* restaurants (McDonalds in Fig. 1), churches, shops (Jewel-Osco in Fig. 1), parks, stadiums, hospitals.

2.1 Mobility Datasets

There are two mobility datasets containing user trajectories that we use in this paper. In an effort to make our study more generic we use datasets that were collected in different regions of the globe.

- The first dataset we use is a GPS trajectory dataset collected by Microsoft Research Asia in their GeoLife project [13]. It contains the data of 155 users that were monitored over a period of over two years. This dataset records a broad range of outdoor movements, including not only daily routines such as going home and going to work but also entertainment and sports activities such as shopping, sightseeing, dining, hiking, and cycling [35,36]. We call this the Chinese dataset.
- The second dataset we use is a dataset collected in the United States from the backbone of a mobile provider network. It contains trajectories of 4,429 users that fielded, for different purposes, applications that report their GPS locations back to the mobile provider. This is actually a service offered by the mobile provider in order to provide location-based services to subscribing users. This dataset was collected over a period of a week. We will refer to this dataset as the American dataset. Although much movement occurs in the Chicago area, there are many American users who live in or travel to other parts of the United States.

Table 1: Characteristics of mobility datasets.

Dataset	United States	China
Timespan	1 week	2 years
Subjects	4,429	155
Samples	708,079	23,209,033
Mass center	36.880,-85.932	39.309,113.016

Table 1 summarizes the datasets characteristics. Both the datasets contain the same type of information. This means that for each user we have an anonymized identifier, and timestamped latitude and longitude tuples. Such information suffices for the purpose of our analysis.

2.2 Point-of-Interest Datasets

The above datasets are enhanced by extracting datasets containing the GPS coordinates of points of interest from China and the United States.

There are several ways to proceed with regards to this. One way would be, based on the trajectories extracted in the previous section, to crawl Google Maps (or similar services) and extract close by-interesting locations. However this process would be very cumbersome and slow mainly because of the large number of location samples that can be noted in Table 1.

A second way would be to use the point-of-interest data that users share publicly on forums and specialized websites [24,25]. For example, users might share favorite restaurants or the locations of speed-trap radars. We find this data to be non-comprehensive as the websites contain mostly locations that spark the interest of only a few users.

The third way (and the method we have adopted in this paper) is to use the points of interest from GPS devices. In fact, the very notion of *point of interest* that we employ in this paper

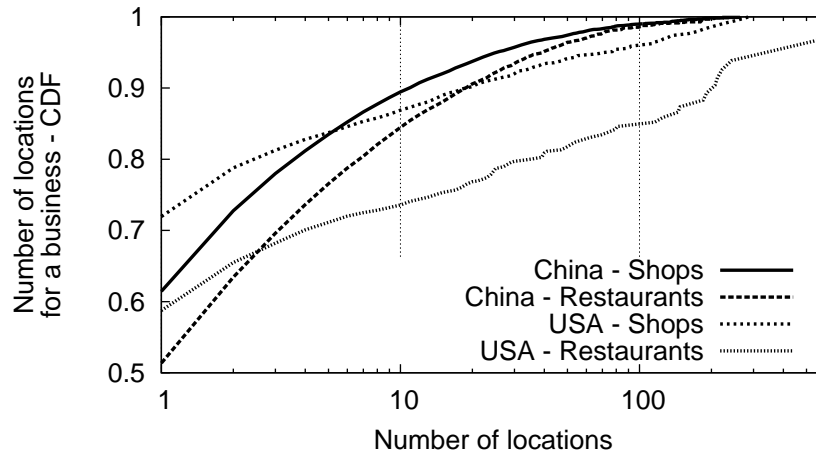


Figure 2: Distribution of the number of distinct locations businesses have across China and Illinois.

is part of the terminology used in GPS devices. Most importantly, GPS devices categorize points-of-interest into about 40-60 classes such as restaurants, shops, airports, and theaters. We therefore have the latitudes, and longitudes of the locations. Statistics for some of the categories we extracted can be seen in Table 2. We show the points of interest for China and for Illinois, Indiana, and Michigan. Note that for each point of interest there might be several entries corresponding to, for example, multiple entrances of the same building. In the table below we summarize only unique occurrences of the same point of interest.

Table 2: Point-of-interest statistics.

Dataset	Illinois	Indiana	Michigan	China
Restaurants	19,400	9,231	14,671	116,095
Shops	56,768	25,454	46,199	267,541
Govt. Offices	4,645	2,890	5,185	137,837
Hospitals, Clinics	748	466	963	54,545
Libraries	692	372	585	40,361
Stadiums	71	24	61	17,868
Population [mi]	12.91	6.42	9.96	1,331
Size [1000 sq. mi]	57.9	36.4	96.7	3,705

Note that, in the United States, there is a higher number of private businesses relative to the population (*e.g.* restaurants and shops). Whereas in China, the number of government offices is much higher (relative to private businesses). A few other insights can be extracted from Fig. 2. It depicts the distribution of the number of locations that businesses have. If one compares shops and restaurants across Illinois (other states exhibit similar characteristics), and China we can see that although Illinois has a lot more businesses with a single location, it also leads in the number of businesses with a lot of locations (basically large corporations). Although China leads the medium number of locations range (medium sized businesses), it lacks some of the larger corporations popular across the United States.

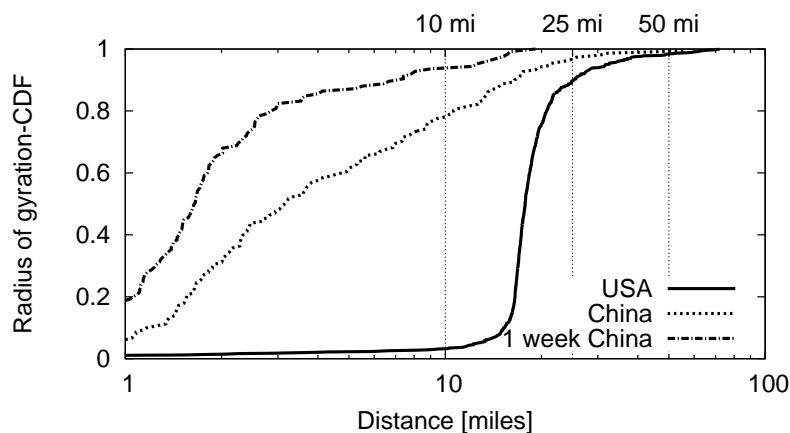


Figure 3: Radius of gyration across a week.

In this paper we restrict ourselves to just a few types of points of interest. Throughout this paper we use the following categories: restaurants, shops, companies, government offices, hospitals or clinics, libraries, and stadiums. We select restaurants, shops, and stadiums for obvious reasons; they attract a large number of users and they hold commercial interest. Also, these popular locations often carry WiFi access points or additional mobile infrastructure to accommodate large crowds. We select the rest based on the fact that the activities they represent are usually linked to people's daily routines. It is important to understand how these locations are spread and we analyze this in Section 3.2. Note also that, for the sake of presentation, some of the categories have been omitted in certain graphs.

3 Preliminary Analysis

In this section we perform independent analysis on user trajectories and on point-of-interest placements. Not all people exhibit the same behavior and although similar patterns can be observed there are particularities in movement spans. Points-of-interest also have a specific placement and a specific distribution, depending on the type.

3.1 Trajectory Analysis

A preliminary trajectory analysis is important, because we require a metric that will help us categorize users. Some people travel over large distances because of their daily job commutes. Other people optimize time and space so that they do not move very much; they live close to work, they shop and dine, respectively, at local stores and restaurants, *etc.* Such an analysis requires a metric that captures the physical space where users carry out daily activities.

When dealing with trajectories in space, a useful analysis metric is the *radius of gyration*. When dealing with certain physical objects, it tells us about the actual size that the object takes in three-dimensional space. For user trajectories, we consider the distance of all GPS points to the center of mass of the trajectory and, as such, it captures the actual movement span that people have. Basically in two dimensions the radius of gyration characterizes the linear size taken

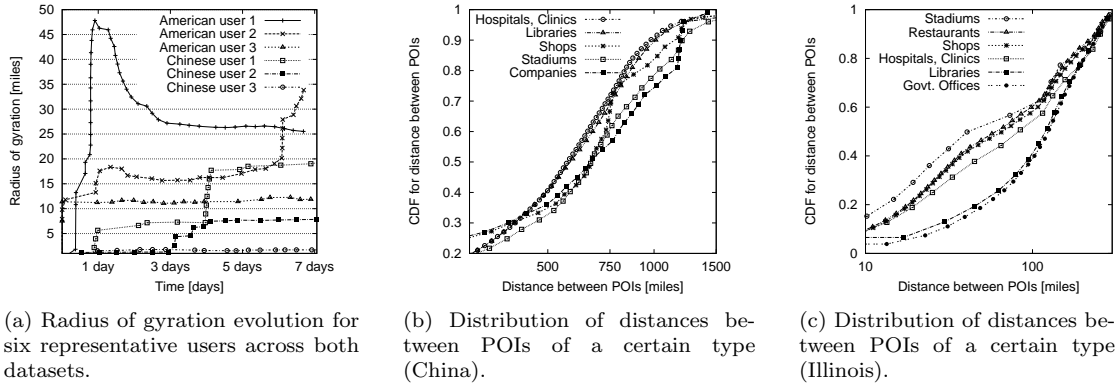


Figure 4: Dataset statistics.

by a certain user’s trajectory up to time t and is defined as:

$$r(t) = \sqrt{\frac{1}{n(t)} \sum_{0 \leq i \leq n(t)} d(\mathbf{x}_i, \mathbf{x}_{\text{cm}}(t))^2} \quad (1)$$

The function $n(t)$ returns the number of trajectory samples until time t . The point \mathbf{x}_i is the i -th location sample for the user, and the point \mathbf{x}_{cm} is the center of mass of the trajectory until time t . Naturally, we use the Haversine distance² as the distance function $d()$ in the above evaluation.

Fig. 3 presents the results of the radius of gyration evaluation for the two datasets. We make the evaluation over the total duration of the traces. Given that the American trace is only one week long, we present the radius of gyration of the Chinese users evaluated over the duration of one week. The results show that the American users travel over larger distances than their Chinese counterparts. This is likely because a many American users live in the suburbs, far from where they work. Also note that the distribution of the radius of gyration values for the American users carries a lower variance than for the Chinese users. To summarize the results, the Chinese users travel less but they exhibit more diverse behavior.

Furthermore, Fig. 4a displays the radius of gyration for six representative users (three from each dataset). The steps in the figures correspond to movement that takes the user significantly out of range from the area occupied so far (captured using the center of mass). This is in accordance to what previous studies observed: People usually move inside one area where they go about their daily activities, occasionally taking larger trips that take them out of this space [14].

To characterize user movement, other studies have used the radius of gyration in conjunction with the *moment of inertia*. This brings the trajectory of each user to a common frame of reference where common movement patterns can be observed. For the purposes of our analysis, considering only the radius of gyration for individual users suffices.

3.2 Point of Interest analysis

In this section we examine the distribution of different point-of-interest types, because it has an obvious effect on people’s trip lengths. We take as metric the distribution of all distances between all points of interest of a certain type. Given that for some point-of-interest types the

²This is the geodesic or great-circle distance between two points that takes the earth’s sphericity into account.

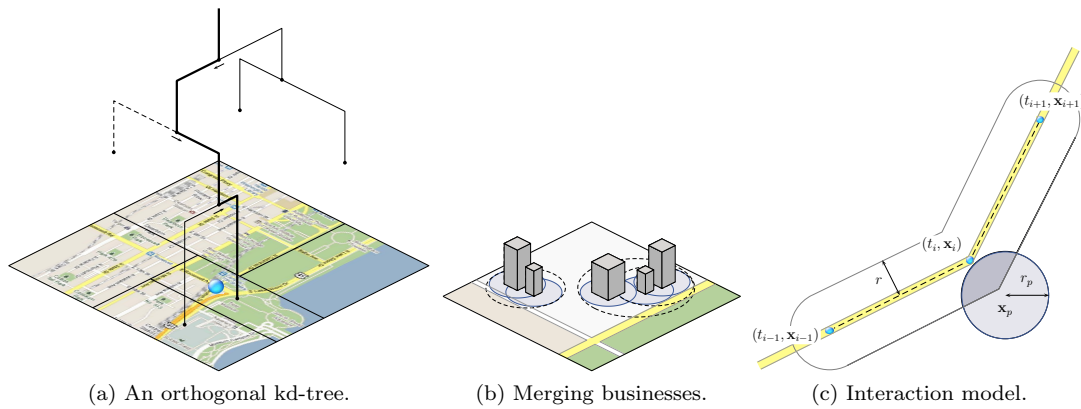


Figure 5: Methodology of our analysis.

number of distances that make up the distribution can be particularly large (in the trillions) we sample the distributions by randomly computing a significant amount of values. Figs. 4b and 4c contain the results for China and Illinois, respectively.

Fig. 4b displays the distribution of distances between the points of interest inside certain different categories. Some insight from the figure is as follows. Companies and shops are clustered together in several centers; this is because of the large number of companies and shops in the distribution are at small distances from one another; and also the jumps are close to 750 and 1250 miles, which correspond to the distances between large city centers (the distance between Shanghai and Beijing, for example, is around 750 miles). Stadiums are also clustered in certain regions; keep in mind that stadiums are not a necessity but more of a commodity. Necessarily they do not exist in certain remote regions, which would explain the appearance of the graph. Also note that hospitals, clinics and libraries all have smoother graphs, which corresponds to the fact that, because they are public services, they need to be almost evenly spread across the whole country.

The same kind of information can be drawn from Fig. 4c corresponding to the state of Illinois where many of the users from the American trace are located. Certain points of interest are spread over all the state of Illinois (*e.g.* libraries, clinics, government offices). Other points of interest (*e.g.* restaurants, shops, stadiums) have a different distribution: more of them are clustered in the largest centers, such as Chicago, Springfield, and Rockford. In fact, the jumps in some of the graphs account for the distances between Chicago, Rockford, and Springfield. This is expected as private businesses are clustered in the most populous areas.

Note that by computing the variance of the distributions we show above, we obtain similar conclusions (although not as expressive). The graphs that are smoother have a larger variance than those corresponding to points of interest that are clustered together. Most points of interest corresponding to private businesses are clustered in the larger populated areas (where our users are located), whereas the public points of interest are more uniformly spread across the whole country/state.

4 Methodology

The analysis of mobility traces and the actual applications that suggest interesting points of interest around a user's position (be they on websites or mobile phones) require an efficient

way to find the closest points of interest to a given location. Basically, a request is a point in two-dimensional space and the k closest points of interest among all (or a subset of them when a certain type of point of interest is desired) must be returned. Obviously, this is a classic k nearest-neighbors (kNN) search in a two-dimensional space with respect to a distance measure. The measure is the Harvesine distance in our case.

To achieve a fast k nearest-neighbors search, the set of points of interest needs to be structured based on latitude and longitude. It is known that r-trees and kd-trees allow for fast lookups, *i.e.* logarithmic on average for multiple-dimension data as our two-dimensional coordinates. Both are trees where nodes correspond to nested regions and leaf nodes contain pointers to the points of interest in the region. Each level of a r-tree splits a region into several possibly overlapping rectangular regions that form a minimum bounding box of the points it contains. Kd-trees partition a region in exactly two sub-regions along a hyperplane, *i.e.* a plane in three-dimensional spaces and a line in two-dimensional spaces. Orthogonal kd-trees split regions only along the axis. Both r-trees and kd-trees exist in an exact and approximate version and give similar performances.

In our analysis, we use orthogonal kd-trees (see Fig. 5a) for their simplicity and their performance and opt for exact k nearest-neighbors search (the approximate version only reduces the worst case complexity). Other efficient solutions exist, *e.g.* ANN [2] or FLANN [11] (set of kd-trees), but they are optimized for high-dimension datasets—such as image descriptors that face the curse of dimensionality—and bring only small improvements for two-dimensional points, compared to the overhead incurred in programming. Interestingly enough, the aforementioned techniques work for any symmetric distance function respecting the triangular inequality, *e.g.* traffic-aware travel time.

The k nearest-neighbors of a location (shown with a circle in the figure) are determined in two stages when using a kd-tree (see Fig. 5a). First, a depth-first search is performed to reach the leaf corresponding to the region containing the desired location (bold line). At each internal node, the searched location is compared to the corresponding splitting hyperplane thus determining which of the children must be therefore recursively explored (\leftarrow or \rightarrow). The k nearest-neighbors are selected among the points of interest in the same region as the location. Secondly, the recursion is unwound to refine the k nearest-neighbors by looking for closer points of interest in other regions. At each internal node, if the distance between the location and the splitting hyperplane is smaller than the distance to the k -th closest neighbor, (so far) the node is recursively explored (plain lines), otherwise, it is ignored (dashed lines).

Our analysis requires a proper formalization of the notion of interactions between users and possibly grouped points of interest. The mobility trace of a user is given by a list of couples (t_i, \mathbf{x}_i) , called *snapshots*, ordered by increasing t_i . Such a couple indicates that the corresponding user was at position \mathbf{x}_i at time t_i . The velocity v_i of a user is assumed to be constant between two successive snapshots. The trajectory of a user is therefore a series of connected segments, namely a polygonal chain:

$$\mathbf{x}(t) = \mathbf{x}_i + \frac{(t - t_i)}{(t_{i+1} - t_i)}(\mathbf{x}_{i+1} - \mathbf{x}_i) \text{ for } t_i \leq t \leq t_{i+1} , \quad (2)$$

and the velocity is $v_i = d(\mathbf{x}_{i+1}, \mathbf{x}_i)/(t_{i+1} - t_i)$. A point of interest p is given by its coordinates \mathbf{x}_p on the map together with meta information such as its name and type (*e.g.* McDonalds/Restaurant).

The first stage of the formalization consists in turning one-dimensional trajectories and positions into two-dimensional areas in such a way that an interaction can be thought of as intersecting areas over a time interval. User positions and points of interest are therefore replaced by discs of radius r and r' respectively. A linear trajectory thus becomes a rectangle (or a tube in

three dimensions).

To capture the fact that businesses are often grouped, *e.g.* in shopping malls and the fact that larger points of interest (*e.g.* stadiums) have multiple close-together coordinate entries in our datasets, we use a region-based segmentation approach, inspired by image processing, defining meta points of interest for points of interest of the same type. The segmentation step is performed directly on the kd-tree. A point of interest or meta point of interest p is represented as a disc. The radius r_p is fixed and equal to r' for points of interest but it can be larger than r' for meta points of interest. When two (meta) points of interest are close enough, *i.e.* the intersection between their respective discs is greater than a given threshold, they are merged into a single meta point of interest represented by the *minimum enclosing disc*, *i.e.* the disc enclosing both discs with the minimum radius (see Fig. 5b). Given two (meta) points of interest at positions \mathbf{x}_{p_1} and \mathbf{x}_{p_2} with radii r_{p_1} and r_{p_2} respectively, the position and radius of the amalgamated meta point of interest are:

$$\mathbf{x}_p = \left(\frac{d + r_{p_1} - r_{p_2}}{2d} \right) \mathbf{x}_{p_1} + \left(\frac{d + r_{p_2} - r_{p_1}}{2d} \right) \mathbf{x}_{p_2}, \quad (3)$$

and $r_p = (d + r_{p_1} + r_{p_2})/2$ where $d = d(\mathbf{x}_{p_1}, \mathbf{x}_{p_2})$ is the distance between the two points of interest. Consider the three right-most points of interest from the example depicted in Fig. 5b. The radius of the two points of interest on the right have a large intersection. The corresponding points of interest are therefore merged and the discs are replaced by a larger one (dashed), tightly enclosing them. The second iteration of the segmentation algorithm merges the newly created meta points of interest with the left-most point of interest yielding a new meta point of interest with an even larger disc (dashed).

We define a model of interaction inspired from physics: given two successive snapshots (t_i, \mathbf{x}_i) and $(t_{i+1}, \mathbf{x}_{i+1})$ of a user and a point of interest p at position \mathbf{x}_p , the *power* of the interaction at time t is defined as a function $p()$ of the distance between the user's position $\mathbf{x}(t)$ (as defined in Eq. (2)) and the position \mathbf{x}_p of the point of interest. The *energy* E of the interaction during the time interval $[t_i, t_{i+1}]$ is obtained through integration:

$$E((t_i, \mathbf{x}_i), (t_{i+1}, \mathbf{x}_{i+1}), \mathbf{x}_p) = \int_{t_i}^{t_{i+1}} p(d(\mathbf{x}(t), \mathbf{x}_p)) dt. \quad (4)$$

Although we impose the power function to depend only on the distance between the respective positions of the user and the point of interest, the choice of its expression is left open. We further constrain its support to positions where the disc around the user and the disc around the point of interest intersect: $p(d(\mathbf{x}(t), \mathbf{x}_p)) = 0$ if $d(\mathbf{x}(t), \mathbf{x}_p) > r + r_p$. This constraint allows for efficient computation of the energy generated by a trajectory by pruning, using the kd-tree, all the points of interest too far from a position. In our analysis we used the area of intersection as a power function:

$$p(d) = -\frac{1}{2} \sqrt{(-d + r + r_p)(d + r + r_p)(d^2 - \delta^2)} + r^2 \arccos\left(\frac{d^2 + r^2 - r_p^2}{2dr}\right) + r_p^2 \arccos\left(\frac{d^2 + r_p^2 - r^2}{2dr_p}\right), \quad (5)$$

where $\delta = r - r_p$. If one of the two discs is included in the other, *i.e.* $d < |r - r_p|$, the expression becomes $p(d) = \pi \min(r, r_p)^2$.

The energy received by a point of interest can be obtained by summing the energy of its interactions with all the segments of all users' trajectories. Such an energy-based approach captures both the spatial and temporal aspects of interactions. In case of an interaction with a meta point of interest, the energy of the interaction is evenly split among the points of interest composing the meta point of interest.

5 Evaluation

In this section, by using the methodology we defined above, we evaluate user interaction with the points of interest. We proceed as follows. First we look at user interests by profiling them based on the types of places they visit. Second, we look at the frequency of visits, a measure captured by the power model defined above. Third, we look at time-of-day effects to find out when the users visit certain types of points of interest.

5.1 User Interest in Locations

In this section we perform *user profiling* to identify broad user interests in the types of locations they visit. To this aim, we match user proximity to points of interest in a conservative manner. If the user trajectory has been very close to a certain point or meta point of interest (r and r' are both chosen to be *five meters*³), we say that the user has shown an interest in that specific type of point of interest. We analyze this by using our above methodology on the entirety of a user's trajectory.

The results of this study are presented in Fig. 6 for both Chinese users and American users. One can see that a majority of the people in both datasets visit companies. The users in the Chinese dataset are described as being professionals so this result is not unexpected: companies employ people who are expected to be at their workplace. The users in both datasets visit restaurants and shops. An interesting fact is that libraries are not visited very much but this can be explained by the fact that people tend to read less and less [27]. The same result holds for government offices. People deal less with authorities in person, performing most of the administrative tasks online instead. The results also show that most users in our datasets do not visit hospitals.

Further, we correlate the results obtained in this section with what we present in Table 2 and Figs. 4b and 4c. We notice that the locations of hospitals, clinics and libraries are more spread than stadiums for example, but there are more users that go to stadiums than to the other types of locations. Considering our previous analysis, there are less stadiums and they are not as uniformly distributed as libraries for example.

Most of the users tend to engage in regular activities such as dining in a restaurant or shopping, yet they also occasionally go to stadiums even though it means traveling more.

Some other facts related to users visiting other types of points of interest (not shown in the figures) is that Chinese users tend to engage more in cultural activities than the American users. A larger proportion of the Chinese users are seen visiting theaters and cultural centers than the American users. Also, a small percent of the American users (4%) are observed to engage in religious activities (*e.g.* going to church). The American users also seem to be more engaged in car related activities. Approximately twice the number of American users can be seen at gas stations, car dealerships, and car repair or maintenance places.

Also, it is important to mention that when considering the radius of gyration measure that we analyzed in Section 3.1 a larger radius of gyration corresponds to a more diverse user profile. Basically, users that have a larger radius of gyration manifest interests in more point of interest types than users with a lower radius of gyration (result not shown).

³Studies of several GPS commercial device receivers puts the accuracy of GPS within a few meters 95% of the time [16]

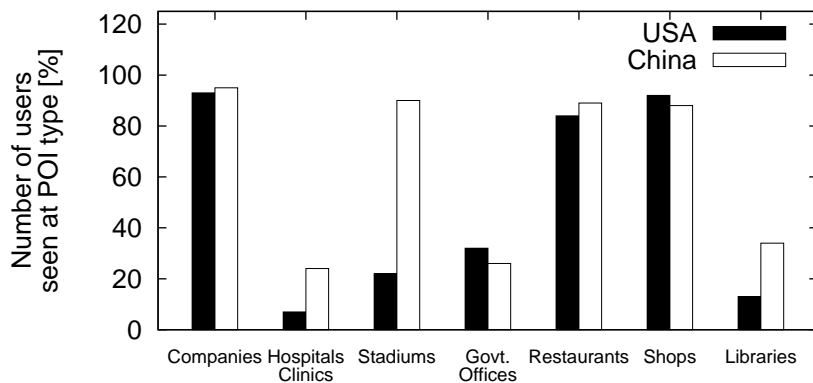


Figure 6: User profiling based on users visiting specific points of interest.

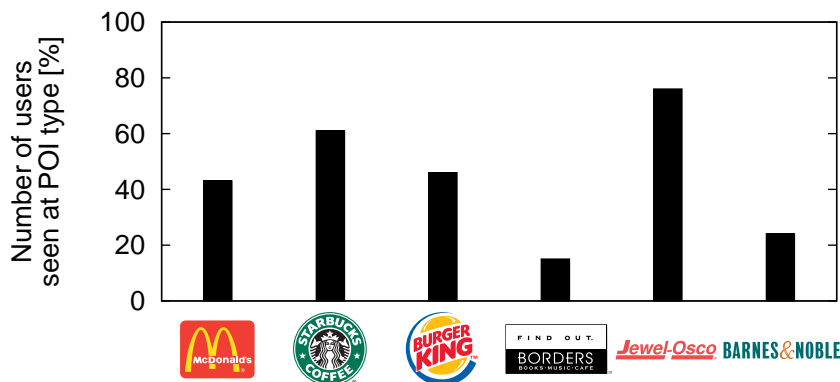


Figure 7: United States – User affiliation with specific businesses.

5.2 Going Deeper: Business Case Studies

In this section we go beyond the analysis we performed above and we analyze specific businesses and the users that visit these locations. We use the American dataset and we select several businesses for our analysis. They are shown in Fig. 7. From the restaurant category, we select the fast foods McDonalds and Burger King. We select Starbucks, Borders, Jewel-Osco, and Barnes and Noble from the shop category. Starbucks is a coffee shop. Borders and Barnes and Noble are book stores. Jewel-Osco is a popular grocery store in the Midwest. Note that all of the analyzed locations (except the Jewel-Osco grocery shops) carry publicly available WiFi access points that are potential content off-loading targets for mobile users.

Fig. 7 contains the number of users in percents seen at the specific locations. For example, a majority of the users are seen shopping at Jewel-Osco grocery shops. Starbucks is more frequently encountered within the area than Jewel-Osco grocery shops and a large number of users are seen there as well. One can note that none of the book stores are popular in terms of the number of users that visit them (Borders recently filed for bankruptcy [6], and several analysts expect Barnes and Noble to follow suit [5]).

Furthermore, because we are dealing with actual business entities and not broad categories, one could imagine scenarios where mobile providers form partnerships with specific businesses

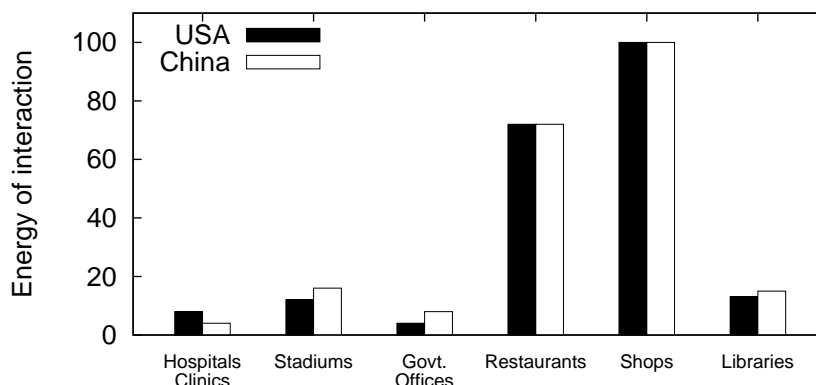


Figure 8: Energy of interaction for users interacting with the specific point-of-interest types.

and have their clients off-load content through publicly available WiFi access points. A way to formalize this problem is through a set covering problem (a well-known NP hard problem). For example each of the above businesses can jointly cover a certain amount of users from all the users. Therefore, we might want to find the minimum number of businesses that a mobile provider can form partnerships with such that we cover all or a majority of the users. By carrying out this analysis on our data and restricting ourselves to businesses that carry WiFi access points, we find that McDonalds, Burger King, and Starbucks jointly cover up to 94% of the mobile users.

Further analysis would be required to identify other interesting aspects with regard to the temporal interaction between users and specific businesses. However, as this is not the main focus of our paper, we leave such analysis for future work.

5.3 Frequency of Visits

In this section we answer the question about which point-of-interest types are more popular. In general, people do not interact with different types of points of interest in the same way. In particular, people visit certain places frequently depending on their function: groceries shopping, dining at a restaurant, watching a game in a stadium, *etc.* We analyze this by relating to the *energy* measure that we introduced in Section 4. We therefore try to capture the total amount of energy that specific points of interest gather during the trace interval from users visiting them.

Fig. 8 shows the results. The histograms are scaled to the highest energy value observed for each dataset, in both cases shops. We purposely do not show values for companies because values for interactions that users experience with the companies they work at dwarf other values.

A few findings from the figure are as follows. The most important is that the results are surprisingly consistent across the two datasets even though they were collected in different parts of the globe. Further, restaurants and shops lead the energy-of-interaction race in both figures. They hold the most energy out of all the analyzed point-of-interest types. This is not surprising as they are part of people's daily activities. In addition, although many more users are seen at stadiums for the duration of the trace, users do not go to stadiums frequently. In fact libraries and stadiums compare in terms of energy of interaction most probably because the users that *do* visit libraries do so more frequently than the users that visit stadiums.

Next, we take a user perspective in order to capture the user activity related to the points of interest. First we contrast the results shown in this section with the results shown in Fig. 9. The figure shows the average distance over all American users from the center of mass for each user

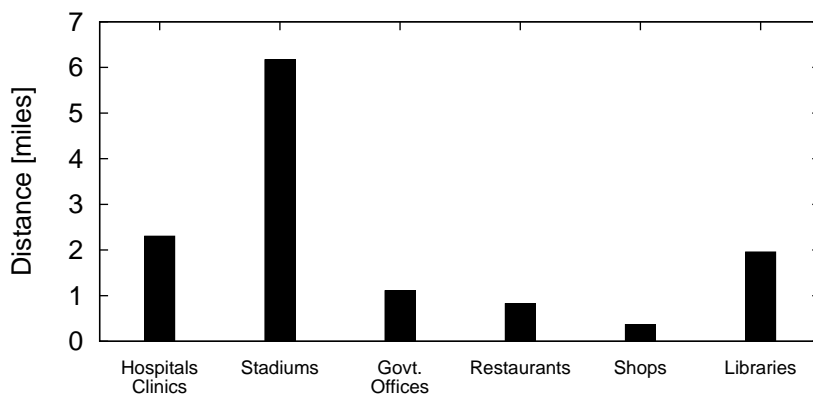


Figure 9: Average distance from the center of mass of each user to the closest point of interest of specific types for American users.

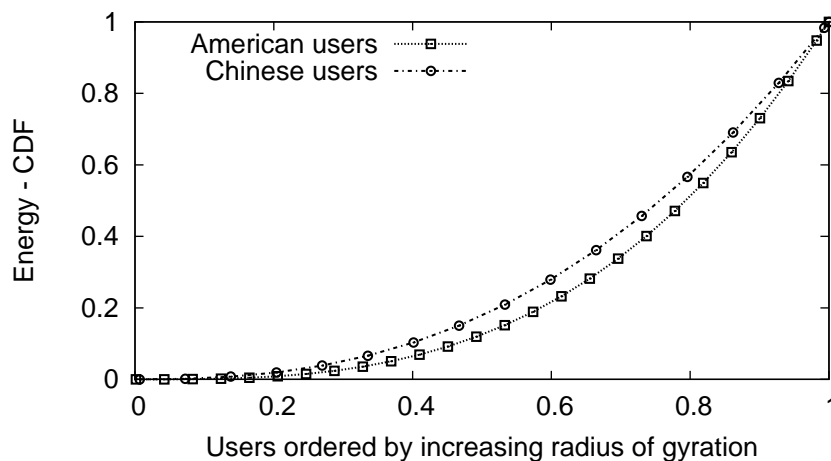


Figure 10: Energy of interaction for users ranked by increasing radius of gyration.

up to the closest point of interest of that specific type. Several categories of points of interest exhibit a direct correlation between the energy they have and the proximity to users. However this is not always the case: Libraries and stadiums are at comparable or greater distances to users than hospitals and clinics, whereas they exhibit higher energy.

Here, we once again refer to the radius-of-gyration measure. We compute the total energy that users gather by interacting with points of interest. Fig. 10 shows the results. The x axis represents all users ranked by increasing radius of gyration values (computed until the end of the trace). A user farther up the x axis will always have a radius of gyration larger or equal than a user who is lower on the x axis. The figure depicts a cdf over the total amount of energy gathered by users. Note that users who have higher radius-of-gyration values interact more with the points of interest around them and thus have higher energy values (the curve is steeper at the end). As shown, this result holds for both American and Chinese users. This confirms our key hypothesis that people move for specific purposes, even if it means traveling over larger distances.

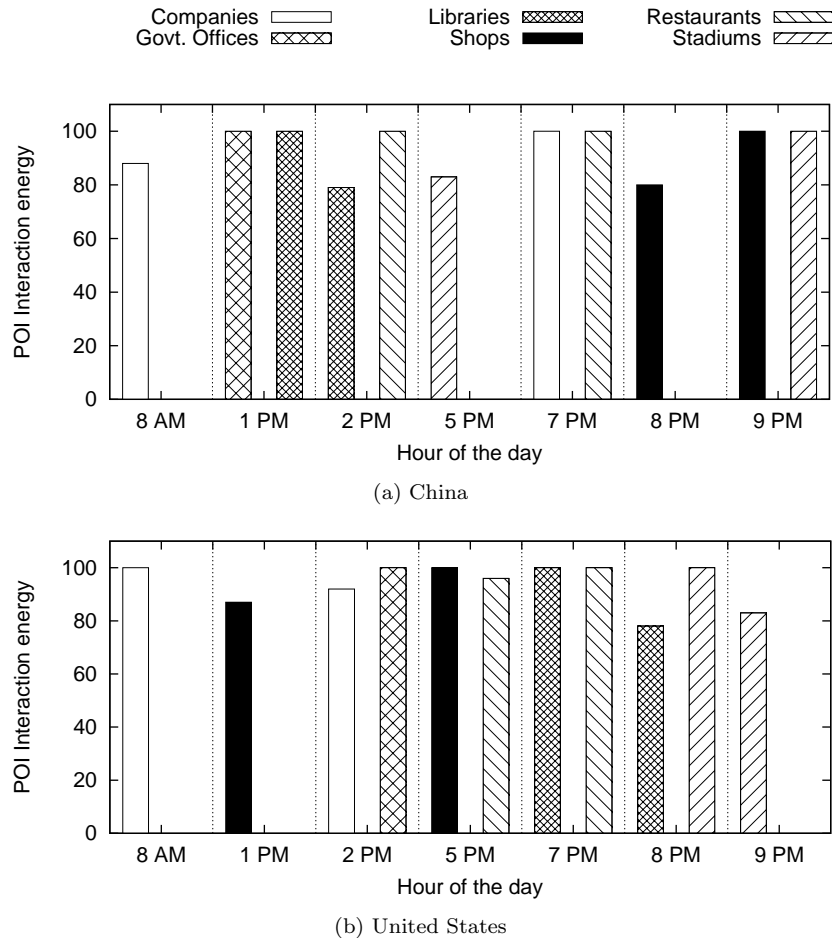


Figure 11: Hour-of-day effects for users interacting with points of interest.

5.4 Time-of-Day Effects

In this section, we analyze the time-of-day effects that user interaction with points of interest might display. As such, we compute the interaction energy across different point-of-interest types for each hour. We only show the times of the day when the energy is the highest or the second highest. In Figs. 11a, and 11b we present the results. For each of the point-of-interest types we have a bar 100% full for the time bin where the energy is the highest for the specific point-of-interest type shown. In cases where the second to highest energy value is larger than 70% we show it as well. For example restaurant visits peak at 2:PM and 7:PM for the Chinese users (Fig. 11a) and at 7:PM for the American users (Fig. 11b). A second bar for restaurants stands at 6 PM and has a value of 96% of the highest energy for the American users (Fig. 11b).

The results in the figures can be explained as follows. Restaurant visits peak during early evening (after work hours) for both the Chinese and the American users. For Chinese users we have a second peak at 2:PM corresponding to lunch hours. The American users also experience a peak at around noon (not shown) that is lower than the peaks we show for 6:PM and 7:PM. The difference in lunch hours can be explained by the different cultures involved.

User presence at companies peaks in the hours of the morning when people arrive at work (8

AM for the Chinese users and 9 AM for the United States users), and during the late hours of the afternoon when people leave (7:PM for the Chinese users and 5:PM for the American users). An interesting phenomenon can be seen in Fig. 11a for the Chinese users. During lunch hours (1:PM and 2:PM) some users engage in other activities such as visiting libraries or government offices. The peaks displayed for both libraries and government offices are significant. They are considerably larger than similar activities for the same points of interest in other time bins (3 times as much). This could be explained by the fact that users optimize their time and group some of the activities and do them during their lunch break.

Both Chinese and United States users engage in shopping during the evening hours (8:PM and 9:PM for the Chinese users, and 6:PM for the American users). Similarly, we note that sports related activities, such as going to stadiums, occur during the evening hours again (9:PM for the American users and 8 and 9:PM for the Chinese users).

We note a tendency of American users to engage in certain activities during late afternoon, most probably after leaving work. This can involve dealing with the authorities, shopping, or dining at a restaurant. This is somewhat in contrast to some Chinese users who group some activities during their lunch break.

6 Discussion

In this section we point out related areas that benefit from research presented in this paper: (i) delay-tolerant networking and content off-loading, and (ii) mobile social networking and mobile advertising.

Delay-tolerant networking and content off-loading. Most research on delay-tolerant networking considers random interactions between human carried mobile devices. Only recently there have been studies on delaying content uploading until better connectivity is available.

Not surprisingly, the trade-off between content delay and energy savings has been recognized in several other recent projects [3, 18, 34] that propose off-loading 3G data to WiFi networks or other mobile infrastructure. Such off-loading or scheduling approaches would definitely benefit from understanding human movement and particularly the relationship that users have with the businesses around them. Businesses often carry their own WiFi access points that could be useful as off-loading targets (*e.g.* Starbucks). Businesses could host new infrastructure for a mobile provider.

Recent works [1, 22] have proposed using the fact that people always take the same routes (the predictability of human movement), linked together with external signs (*e.g.* the presence of certain Bluetooth devices) that could often indicate better connectivity options.

Our work has the potential to take this research one step further by enhancing GPS-equipped mobile phones with the location knowledge about points of interest and businesses that carry public WiFi access points.

Mobile social networking and mobile advertising. Social networking websites have recently become very popular. Users try to keep in contact with each other all across the world. They try to follow the lives of their friends, the locations they visit, the things they buy, *etc.* There are different reasons users post information about their lives or follow their friends. Some people want to get help from friends in their daily lives and some other people just want to “show off” [21]. Although it is an integral part of the social networking experience, human movement has begun to fascinate social networking users. Some mobile social networking applications, including the increasingly popular social gaming systems Foursquare [12] and SCVNGR [28], are built entirely around users visiting places of interest. Some of these systems (*e.g.* SCVNGR) partner with existing retailers and offer badges or rewards such as gift cards or gaming items for

doing everyday activities such as shopping at Walmart or buying coffee from Starbucks.

Such applications (and others that might emerge) involve users actively moving to obtain the desired result and are integrated with newer technologies that allow users to use their mobile phones as credit cards. For example, [20] carries the potential of becoming the next trend in mobile social networking [29,30].

Our study could benefit all of the above. Understanding human interaction with points of interest could help mobile social networking applications better target their campaigns when dealing with specific retailers. Mobile advertising could also be enhanced by targeting users that manifest an interest in a specific businesses or point-of-interest types.

7 Related Work

Motivated by the emergence of Internet connected hand-held devices, mobility models for human movement have recently received increasing attention in the networking research community [9]. For example, the Brownian Motion (*i.e.* random walks with Gaussian steps), Lévy Flights (*i.e.* random walks with possibly truncated heavy-tailed steps), Random Way Point (*i.e.* linear motion between randomly chosen points) and Fractal Way Point (*i.e.* random points forming a fractal). This is mainly because of their applications in delay-tolerant networking and vehicular ad-hoc networks [15] (*e.g.* routing [10] and data off-loading [3,4,18,26]), mobile applications (*e.g.* location-aware data storage [32], and social gaming [12]), *etc.* The so-called digital footprints obtained from popular GPS-enabled mobile devices sparked the idea of these models and were used to validate them.

Even though some of the aforementioned models are able to capture spatial and temporal patterns of social behaviors, they do not take into account the particular semantics incorporated in human movement. Although the mobility of a user may effectively be a polygonal line as in the random way-point model, the destinations are not chosen (totally) at random but are determined by the semantics of each point of interest and the users' interest at that specific time. This is precisely what we investigate in this work, based on the notion of points of interest.

In [14], it is shown that human movement shows strong temporal and spatial patterns with a high probability of returning to a few locations. These locations can be of interest for a single person or a few people (*e.g.* home or office), or for a larger number of users (*e.g.* restaurants). The type of locations of interest to few was studied in [19] based on an analysis about the way people name locations. And those of interest to many, effectively captured by the notion of points of interest, are at the core of our work. Our work confirms the findings from [14] and further details the semantics of these locations and the frequency at which they are visited. In [33] (our own work), we show that the applications accessed by mobile users are correlated with their current location and their mobility patterns. In [7], a large-scale (*i.e.* the whole United States) mobility trace of banknotes is analyzed by highlighting the limits of Lévy Flights to model human travel and therefore a two-parameter continuous-time random walk model is proposed instead.

Another noticeable work in the analysis of human mobility is the SLAW model [17] which unifies the most significant previously proposed models in order to generate realistic mobility traces for human walks. Although we acknowledge the usefulness of human mobility models, we are more concerned with their lack of expressiveness. As such, our work can lead to better, semantically enhanced mobility models.

Closest to our work, [36] studies the interactions between users and interesting locations (similar to the notion of point of interest). The users' travel experience and location interest are jointly inferred using a new model, outperforming traditional rank-by-count and rank-by-frequency methods. The study is limited, however, as they only use a limited number of scenic

locations mainly of interest to tourists' and therefore they do not target points of interest at large. Assuming that human mobility is driven by specific goals, points of interest are used in [8] to predict human mobility. Our work backs up this assumption and conveys additional information on, for instance, the effect of the time of the day on interactions with specific businesses. In [31] a similar approach, based not only on points of interest but also on spatial constraints (roads, obstacles, *etc.*), is used to model outdoor human mobility. In [23] and [37], activity patterns are extracted from mobility traces by considering the points of interest surrounding users at different hours of the day. It is shown that the activity patterns of users working in the same area exhibit strong correlation.

8 Conclusions

In this paper we have proposed a new approach to analyzing human mobility. Noting the lack of expressiveness incorporated in existing studies, we have performed the first-of-its-kind joint analysis of human mobility correlated with the surrounding environment, *i.e.* the points of interest that they visit. Our key contribution is the demonstration of user affinity (with different visiting frequency depending on point-of-interest type) towards specific points of interest and specific businesses. We believe our results show noteworthy promise for further research in this area, clearing the way for future advances in understanding basic human behavior and impacting problems related to mobile transfer scheduling algorithms, mobile infrastructure placement, mobile social networking, mobile advertising, *etc.*

Summary. From the user perspective, we find the following: *(i)* Users with a more diverse user profile and a larger number of interactions with points of interest exhibit a larger movement span characterized by a larger radius of gyration. *(ii)* There is strong evidence related to users who optimize their time by, for example, overlapping certain activities with their breaks. *(iii)* Although users in different parts of the globe exhibit different time-of-day behavior, we find that there is a striking correlation in the frequency of visits to the point-of-interest types that we have analyzed.

From the point of interest perspective, our findings are the following: *(i)* User presence at points of interest displays a strong time-of-day effect, certain points of interest being more popular in the afternoon and others are more popular in the evening. *(ii)* Point-of-interest proximity to users, together with functional aspects (the point-of-interest type) are the two factors that influence popularity.

Acknowledgements

Kévin Huguenin was partially funded by a scholarship offered by the University of Rennes I and an “explorateur” grant offered by INRIA.

References

- [1] G. Ananthanarayanan and I. Stoica. Blue-Fi: Enhancing Wi-Fi Performance Using Bluetooth Signals. In *MobiSys'09*.
- [2] ANN: A Library for Approximate Nearest Neighbor Searching. <http://www.cs.umd.edu/~mount/ANN/>.

- [3] A. Balasubramanian, R. Mahajan, and A. Venkataramani. Augmenting Mobile 3G Using WiFi. In *MobiSys'10*.
- [4] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani. Energy Consumption in Mobile Phones: A Measurement Study and Implications for Network Applications. In *IMC'09*.
- [5] Borders' Bankruptcy Shakes Industry. http://www.nytimes.com/2011/02/17/business/media/17borders.html?_r=1&partner=yahoofinance.
- [6] Borders Files for Bankruptcy, to Close 200 Stores. http://www.msnbc.msn.com/id/41536256/ns/business-consumer_news/t/borders-files-bankruptcy-close-stores/.
- [7] D. Brockmann, L. Hufnagel, and T. Geisel. The Scaling Laws of Human Travel. *Nature*, 439:462–465, 2006.
- [8] F. Calabrese, G. Di Lorenzo, and C. Ratti. Human Mobility Prediction Based on Individual and Collective Geographical Preferences. In *ITSC'10*.
- [9] T. Camp, J. Boleng, and V. Davies. A Survey of Mobility Models for Ad Hoc Network Research. *Wireless Communications and Mobile Computing*, 2:483–502, 2002.
- [10] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of Human Mobility on Opportunistic Forwarding Algorithms. *IEEE Transactions on Mobile Computing*, 6:606–620, 2007.
- [11] FLANN: Fast Library for Approximate Nearest Neighbors. <http://www.cs.ubc.ca/~mariusm/index.php/FLANN/FLANN>.
- [12] Foursquare. <http://foursquare.com/>.
- [13] GeoLife GPS Trajectories. <http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/>.
- [14] M. Gonzalez, C. Hidalgo, and A. Barabasi. Understanding Individual Human Mobility Patterns. *Nature*, 453:779–782, 2008.
- [15] J. Harri, F. Filali, and C. Bonnet. Mobility Models for Vehicular Ad Hoc Networks: A Survey and Taxonomy. *IEEE Communications Surveys Tutorials*, 11:19–41, 2009.
- [16] S. Kindra, W. Thomas, and W. Keith. Comparing GPS Receivers: A Field Study. *URISA Journal*, 18:19–23, 2006.
- [17] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong. SLAW: A Mobility Model for Human Walks. In *INFOCOM'09*.
- [18] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong. Mobile Data Offloading: How Much Can WiFi Deliver? In *CoNext'10*.
- [19] J. Lin, G. Xiang, J. I. Hong, and N. Sadeh. Modeling People's Place Naming Preferences in Location Sharing. In *UbiComp'10*.
- [20] Mobile Phone Payments. http://money.cnn.com/galleries/2011/technology/1101/gallery.mobile_payments/index.html.

- [21] Mobile Social Networking Set for Growth. <http://www.emarketer.com/Article.aspx?R=1006514>.
- [22] A. J. Nicholson and B. D. Noble. BreadCrumbs: Forecasting Mobile Connectivity. In *MobiCom'08*.
- [23] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data. In *HBU'10*.
- [24] POI Data. <http://poi.gps-data-team.com/>.
- [25] POI and Map Data. <http://downloads.cloudmade.com/>.
- [26] M.-R. Ra, J. Paek, A. B. Sharma, R. Govindan, M. H. Krieger, and M. J. Neely. Energy-Delay Tradeoffs in Smartphone Applications. In *MobiSys'10*.
- [27] Study: Americans Reading Less Than They Used to. <http://www.npr.org/templates/story/story.php?storyId=16739654>.
- [28] SCVNGR. <http://www.scvngr.com/>.
- [29] SCVNGR Goes Global, Unleashes Zombie Horde on Businesses. <http://personalmoneystore.com/moneyblog/2010/11/02/scvngr-global-zombies-google-places/>.
- [30] SCVNGR Unleashes Zombie Horde Through Social Check-in Feature. <http://www.wired.com/magazine/2010/10/scvngr-unleashes-zombie-horde-through-social-check-in-feature/>.
- [31] I. Stepanov, P. J. Marron, and K. Rothermel. Mobility Modeling of Outdoor Scenarios for MANETs. In *ANSS'05*.
- [32] P. Stuedi, I. Mohamed, and D. Terry. WhereStore: Location-Based Data Storage for Mobile Devices Interacting with the Cloud. In *MCS'10*.
- [33] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network. In *IMC'09*.
- [34] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Taming User-Generated Content in Mobile Networks Via Drop Zones. In *INFOCOM'11*.
- [35] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding Mobility Based on GPS Data. In *UbiComp'08*.
- [36] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining Interesting Locations and Travel Sequences from GPS Trajectories. In *WWW'09*.
- [37] M. Zignani and S. Gaito. Extracting Human Mobility Patterns from GPS-Based Traces. In *WD'10*.



**RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE**

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Volveau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399