How many places do you visit a day?

Michela Papandrea[‡], Matteo Zignani[†]
Sabrina Gaito[†], Silvia Giordano[‡], Gian Paolo Rossi[†]

[‡]NetLab, ISIN-DTI, SUPSI, Manno, Switzerland
[michela.papandrea,silvia.giordano]@supsi.ch

[†]Dipartimento di Informatica, Universitá degli Studi di Milano, Italy
[matteo.zignani,sabrina.gaito,gianpaolo.rossi]@unimi.it

ABSTRACT

People mobility enormously augmented in the last decades. However, despite the increased possibilities of fast reaching far places, the places that a person commonly visits remain limited in number. The number of visited places of each person is regulated by some laws that are statistically similar among individuals. In our previous work, we firstly argued that a person visit most frequently always few places, and we confirmed that by some initial experiments. Here, in addition to further validating this result, we build a more sophisticate view of the places visited by the people. Namely, on top of our previous work, which identifies the class of Mostly Visited Points of Interest, we define two next classes: the Occasionally and the Exceptionally Visited Points of Interest classes. We argue and validate on real data, that also the occasional places are very limited in number, while the exceptional ones can grow at will, and by the analysis of the classes of visited points we can distinguish the type of users mobility. This paper firstly demonstrates this property in large experimental scenario, and put the basis for new understanding of people places in several areas as localization, social interactions and human mobility modelling.

Keywords

human mobility, visited places, mobility traces, evaluation, simulation

1. INTRODUCTION

In this paper, we introduce a classification of the points of interest visited by people and this allows us to define a general profile of people, characterized by the number of locations and time spent there per class. We first elaborate on existing works on human behaviors and especially on their analysis in terms of the social world around people, arguing that similar properties are present in terms of human mobility and visited points. Then, we present our experimental approach in Section 3, starting from the Microsoft traces [11] and presenting the huge pre-processing work. We dive into the obtained results, their meaning and statistical evaluation in Section 5. We use the results to separate the visited points into 3 classes (*Mostly, Occasionally and Exceptionally Visited Points of interest - MVPs* [6], *OVPs, EVPs*) and show that, on average, people frequent just few points, but they are frequented for more than 50% of the time. Also the OVPs are low in number, and they are visited for about 10%, while the remaining points, the ones in the EVP class are visited for a very short amount of time.

2. BACKGROUND AND MOTIVATION

Recent faster transportation methods have made people mobility very common for both businesses and daily life. In addition, advances in communications technology, data analysis and smart infrastructure are enabling to streamline the transportation strategies, simplifying connections and shortening the commuting times. These two aspects together resulted in a high mobility degree for many people, both for their business or as a lifestyle. However, despite the higher mobility degree, we argue that the MVPs, the places that a person visit more frequently and thus were a person can be caught with higher probability, are still a limited number. Mid 90s, Strogatz and Watts [10] were modeling the famous 'six-degree' property of Milgram, giving birth to the small world phenomena era: the average path length for social networks of people was established to be six. In 1992, Dunbar measured the correlation between neocortical volume and typical social group size [2]. He showed that, because of the limit imposed by neocortical processing capacity, people can have stable interpersonal relationships with only a limited number of individuals. Thus, the Dunbar's number is the measure of the humans' social network size, and is between 100 and 200 individuals [3]. In addition to the neuro-scientific limits, we can also individuate some physical constraints, as our time and interests are finite and therefore we cannot have (strict) social interactions with the whole world. Both results concur to give a surprising view of how our social world is "small" (connected with small number of hops) and cannot go over certain limits (we have limited numbers of strong connections). We argue that, similarly, *our physical world is small* and cannot go over certain limits [7]: we can commute, with small number of hops, between very far places, but the number of points that we frequently reach is limited. Intuitively, the

PerMobi 2013 San Diego, US
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

978-1-4244-9529-0/13/$31.00 ©2013 IEEE

fact that we can commute everywhere, with small number of hops is clear, but the fact that our mostly visited points are few is not so evident, especially if we consider the evolution of our society toward a very dynamic lifestyle. Thus, regardless the increased attractiveness of a place or the possibility to reach places more quickly, people will keep on moving around their limited number of MVPs (Giordano-Papandrea class of Points of Interest [6]) for most of their time. In this work, we present some initial analysis on a real dataset. We show that our intuition is validated by the empirical results, and also that we spend more than 50% of our time in those MVPs. This indicates that, willy-nilly, those points are the ones that better represent and characterize our life. We show that we can also go further, using the time spent in each class to distinguish different types of human mobility profiles: the stay-at-home users, with a very low time spent in the EVPs, and the globetrotter users, which, as opposite, present a very high time within EVPs. Our result could impact on several areas as: localization [9], where it can be predicted that people are in MVPs with a probability higher than 0.7; social interactions [5], as people tend to meet more frequently people with some MVPs in common; human mobility modelling [12], as mobility can be described in terms of movement among MVPs.

3. DATASET

In this paper we use a very large GPS dataset recording the movement of 178 people in a period of over 4 years (from April 2007 to October 2011). It was collected in GeoLife project and released by Microsoft Research Asia [11]. People participating to the experiment are students, government staff and employees from Microsoft and several other companies equipped with GPS loggers or GPS-phones. Overall the dataset provides 17,621 trajectories with a total distance of 1,251,654 kilometers and a total duration of 48,203 hours. With respect to other datasets with mobility data collected in a limited area or in a particular context, Geolife dataset offers a high heterogeneity. As a matter of fact, it contains a broad range of users' outdoor movements, including both everyday routines imposed by working activities and free time activities.

Besides having been conducted over a long period of time and involving a high number of users, this dataset is interesting also for its temporal and spatial fine granularity, as 91% of the GPS trajectory are recorded in a dense representation, every 1∼5 seconds or every 5∼10 meters per point. This allows us to precisely capture Points of Interest (PoIs) associated to the different activities an user undertakes.

If on the one hand the dataset is very rich, on the other side it exhibits a high level of fragmentation, especially with regard to features as the effective duration of the trajectories, the data collection period and the number of trajectories per user. Indicatively, more than half of the trajectories span less than one hour, while about 60% of users collected data for less than a month. Furthermore the dataset covers a large area of the earth from Europe to USA to Asia, although it does not represent a problem as the major part of the data is located in Eastern Asia, in an area corresponding to the region around Beijing. We limit our analysis to GPS data collected in this area, as our main goal is to characterize the most visited PoIs.

3.1 Dataset pre-processing

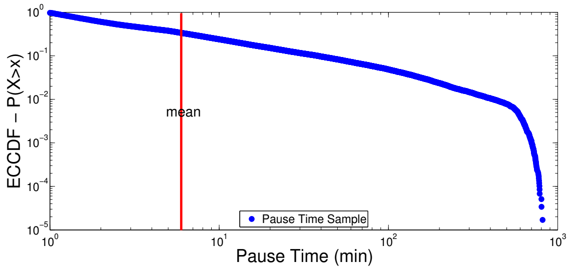


Figure 1: ECCDF of the aggregated pause times in the stay-locations.

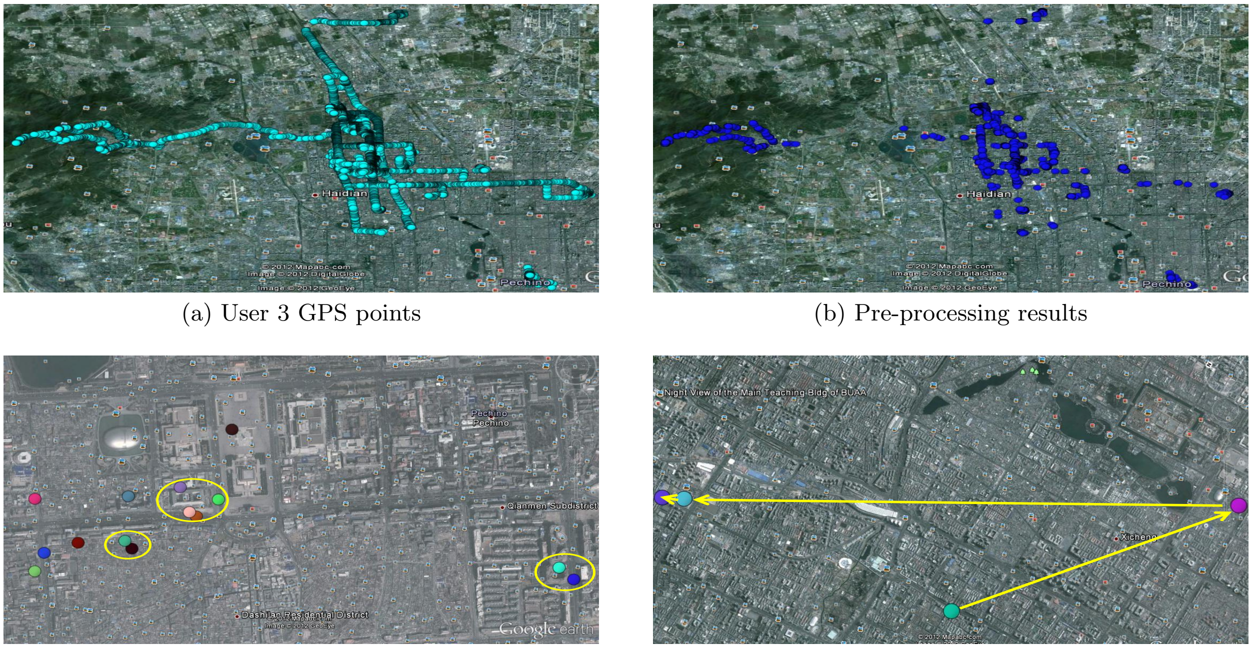
Although GeoLife represents the most reliable dataset publicly available, it was not collected to extract PoIs and thus we need to pre-process trajectories in order to find the most meaningful ones to our goal. The need of a pre-processing phase is dictated by the dataset bias which flavors movement, while we are interested in people still in their PoIs. In particular we aim at densifying trajectory points corresponding to the pause phase by a filling heuristic, while removing points belonging to users' movements.

Indoor filling Mobility data collected by GPS devices present gaps because GPS signals are often disrupted inside buildings. This represents a big problem, especially if one is interested in detecting the PoIs of a user. In fact, in many cases, buildings or other indoor locations represent the most of the PoIs visited by a person during the day. To overcome the problem given by missing records [8], so to avoid an underestimation of the number of PoIs, we apply the following simple rule. When the ending and beginning GPS points of a gap are within a distance of 35 meter and the gap duration is greater than 5 min, the user is taken as residing at the same location during that time. This rule also supplies for the situation where the individual enters a building, or where the individual turns off the GPS devices in an indoor place. Practically, we add as many GPS points equal to the entry point as the duration in sec of the gap. After the trajectory reconstruction phase, we noticed a big increment of points, anyway limited by the threshold imposed on the gap duration.

Movement phase reduction We apply a filter aimed at leaving out data which describe the movements among the PoIs a user visits, thus reducing the number of points to analyze. This way we consider the periods in which a user stays still in a place, assuming that users manifest their interests by spending an amount of their time in such places. In order to extract the pause periods and their related GPS points from the whole individual trace, we apply the heuristic proposed in [14, 13], where a similar but smaller dataset has been analyzed. If two points p_i and $p_i + 1$, with timestamps indicated by $t(p_\cdot)$, do not satisfy

$$\frac{\|p_{i+1} - p_i\|}{t(p_{i+1}) - t(p_i)} \leq \Delta \quad (1)$$

then we delete p_{i+1} from the original trace, since it belongs to the movement phase. Analyzing walking mobility data, we set the threshold to the very low value of $\Delta = 1.3m/s$, according to the fact that we observe that human walking speed is about 4-5 km/h (1.1-1.4 m/s). It seems a reasonable value as generally, in a location, people do not reach the maximum speed. This way, we capture points where a person is still or is moving very slowly inside a small area. The


Haidian
Pechino
Qianmen Subdistrict
Dashilan Residential District
Google earth
Night View of the Main Teaching Bldg of BUAA
Xicheng

(a) User 3 GPS points

(b) Pre-processing results

(c) Sub-PoIs issue

(d) Compact representation

Figure 2: PoIs extraction applied to the user 3's trajectories. In 2(a) we plot all the recorded points (raw data). In 2(b) we show the points resulting from the application of the pre-processing phase. In 2(c) the sub-PoIs that have to be grouped in the real PoI (yellow circle). In 2(d) a compact representation of user 3's mobility during a single day.

result of the speed filtering process is a sequence of points that forms the trajectory $S = ((p_1, t_1), ..., (p_n, t_n))$, where t_i is a timestamp and $p_i \in \mathbb{R}^2$, on which we apply the PoIs extraction methodology proposed in Section 4.

Users' selection The point reduction has also effects on the number of users and the number of days, per user, from which we can extract places of interest. The reduction is mainly due to the fact that GeoLife dataset has been built for the transportation prediction task, and, as a consequence, it flavors movements.

To overcome these limitations, we classify the users considering two properties: the period (in hours) a single day trace spans and the number of days the single user traces cover. In particular, for each user, we only consider the daily traces that record more than h hours. On these tracks we count the number of users that have more than d days of data. In particular, for all the users of the dataset, we filter out all the days of sampling (data collected within the 24 hours, going form 00:00 AM until 11:59 PM) which have $h \leq 3$ hour of sampling. All the remaining days are considered *relevant days*. After this first processing, we filter out all the user which collected less than 20 *relevant days* of data ($d = 20$): the resulting number of users is 21, over the total number of 178 users. We apply these values for the users filter parameters, in order to optimize the trade-off between the importance of having a large number of users, to be able to generalize our analysis; and the need to deal with sampled data which does not only correspond to trajectories. For example, only by increasing of one hour the threshold h we obtained a number of users that is not enough to our goal (10). Though the dataset used is a collection of trajectories, hence only a reduced subset of collected data fulfill our requirements, we are able, also with this small dataset, to obtain initial but powerful results. Besides, note that the resulting dataset almost completely spans the original GeoLife period.

4. POIS EXTRACTION METHODOLOGY

GPS datasets, such the one we are analyzing, present many difficulties as concerns the PoI extraction task with respect to mobility data inferred from geo-coded or geo-tagged social networks [1] (e.g. Foursquare, Facebook Places,. . .). In our context we do not have any information about the interest expressed by the user, but we must rely only on the periods where a user is still.

Assuming a constant sampling rate, as in our case, the pause periods and the places visited by users translate in an higher concentration of recorded points. This way the PoIs extraction corresponds to the unsupervised task of density-based clustering. In particular, we are extending the methodology proposed in [14], adopting a two-level density-based clustering combined with a thresholding mechanism based on pause in the regions extracted from the first clustering phase.

4.1 Finding PoIs

All the points of a trajectory belong to the pause phase and are the starting points to extract the PoIs. To reach this goal, we first find the possible regions of interest via a clustering algorithm and then we detect the real PoIs considering the pause time feature.

Formally, we capture the possible regions by introducing the concept of stay-location L.

Definition 1. *Let S be a trajectory and $L = \{L_1, \ldots, L_k\}$ a partition of $\{p_1, \ldots, p_n\}$ s. t. for each $Li \in L$, L_i is* *maximal w.r.t. the property that for each $p_u, p_v \in L_i$ exists a sequence $(p_u = p_w, ..., p_{w+j} = p_v)$ of points in L_i, s.t. $\|p_{w+k} - p_{w+k+1}\| \leq \delta, k = 0, ..., j-1$ for a fixed δ. A stay-location is an element of L.*

Informally, a stay-location is an area where a person stops, independently of how long he stays there. Let us consider individual traces in order to extract stay-locations and analyze their properties. To find stay-locations we apply the density-based clustering algorithm DBSCAN [4]. As DBSCAN parameters we use $\delta = 10$ mt and $\epsilon = 2$ neighbors (δ represents the maximum distance such that two points are considered neighbors, while ϵ is the minimum number of neighbors that a node must have to be considered in a cluster).

We observe that in daily movements, there are many stay-locations where an individual stays for a short amount of time. These stay-locations are meaningless as they represent small pauses in the movement towards the real destinations that we call Points of Interest.

Definition 2. *Let S be a trajectory and $L_i \in L$ a stay-location. L_i is a Point of Interest (PoI) if in S there exists a subsequence $((p_i, t_i), \ldots, (p_{i+k}, t_{i+k}))$ such that $p_{i+j} \in L_i$ for $j = 0, \ldots, k$ and $t_{i+k} - t_i \geq \phi$.*

In the following analysis of the dataset we set the threshold $\phi = 5$ min, which corresponds to the mean of the pause distribution in stay-locations, shown in Figure 1. We must underline that we do not consider the sum of the pause times in a stay-location; rather, we consider the single values. The thresholding results in the meaningful PoIs, although we observe situations, such that presented in Figure 2(c), where we have many sub-PoIs of the same general PoI. To overcome this empass we run a second passage of DBSCAN with a larger ϵ on the centroids of the sub-PoIs detecting the real points of interest. Thus we obtain two important effects: we drastically reduce the number of stay-locations and can infer which are the main destinations, the PoIs.

Aside from finding PoIs, the above methodology has the capability to express human mobility as a compact trace that summarizes the transitions between PoIs and the users' pause time in them as shown in Figure 2(d). Adopting this compact representation in the following section we can analyze some properties of the human mobility and of the PoIs human beings visit during their daily movements.

4.2 Relevance

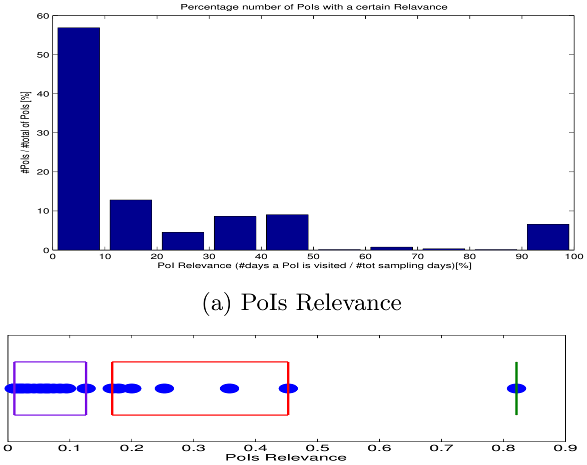
We classify the user's visited PoIs according to their relevance, and then we derive some characterizations of the user's mobility within each locations' class of relevance.

The *relevance* of a certain location L_i has been calculated on the mobility history of each user, and it is defined as:

$$relevance(L_i) = \frac{d_{\text{visit}}(L_i)}{d_{\text{total}}} \quad (2)$$

where $d_{\text{visit}}(L_i)$ is the number of days a location L_i has been visited (one or more times per day) by the user and d_{total} is the total number of sampling days, collected by the user. The relevance of a certain location is, according to the formula, the percentage of days the user visit this location, over the total number of days of sampling.

According to relevance values, we show that the PoIs associated to each user can be grouped in 3 classes:



(a) PoIs Relevance

(b) Three classes of relevance in a sampled user

Figure 3: Aggregated and single user classes of PoIs based on the relevance.

- **Mostly Visited PoIs (MVP)**: locations most frequently visited by the user. We can easily infer their semantic meaning, and associate them to home location, work place, gym.
- **Occasionally Visited PoIs (OVP)**: locations of interest for the user, but visited just occasionally.
- **Exceptionally Visited PoIs (EVP)**: PoIs unlikely visited more than very few times.

The evaluation of the PoIs' relevance allows a straightforward identification of these three classes (more details in the next paragraph). In figure 3(a) we show a cumulative characterization of the PoIs identified for all the users of the dataset (here we use the complete dataset, without applying any filter). In the x-axis we identify 10 bins of relevance spanning form 0% to 100%, where each of them has a width of 10%. For each class of relevance we show in the figure the percentage of the average (calculated over all the users) number of PoIs belonging to the corresponding class. From the figure, it is easily visible that, on average, 57% of the PoIs visited by a user are within the EVP relevance class: this means that more than half of the PoIs seen by each user, are exceptionally visited PoIs, that the user hardly visits for multiple times. 6.7% of PoIs can be classified within the MVP class, which gives an idea of the limited number of locations which are visited by each user almost daily. The identification of the upper and lower bounds for each of the three classes is strictly related to every single user; in fact it depends on the user's mobility style.

4.3 Finding class of relevance

As it has been highlighted by the above discussion, relevance class bounds could change among the subjects. As a consequence, class bounds cannot be fixed *a priori* but claim at an automatically detection algorithm able to adapt to the single user mobility pattern. In particular, we adopt an unsupervised approach which groups the PoIs of a single user according to the values of PoI relevance and maximize their separability. The clustering algorithm we choose is the k-means with $k = 3$ which corresponds to the number of PoI classes. To avoid the problem related to the initial choice of the centroids, we run 10 replicas of k-means with different initial seeds and choose the partition that minimizes the within-cluster sums of point-to-centroid distances, thus maximizing the separability. In Figure 3(b), we show the result of the k-means clustering on a sampled user. The EVP class (purple box) covers the range from 0.01 to 0.12, the OVP (red box) spans the range from 0.16 to 0.46 and the MVP class (green line) contains only one PoI with relevance 0.82.

5. RESULTS

In this paragraph we present the experimental results of our analysis performed over the data after performing the pre-processing and the analysis of the PoIs and related classes of relevance, described in the previous sections of the paper.

For each filtered user, we apply the k-mean algorithm (as explained in paragraph 4.3) to classify the related PoIs in three main classes of relevance (4.2) and over these classes we study three main features: *(i)* the number of PoIs which reside within each class of relevance, *(ii)* the percentage of time spent in each class and *(iii)* the average time of the visits to the PoIs of the classes.

In Figure 4 we represent the number of PoIs associated to each class of relevance, per user. In the upper plot we can notice the large difference in the number of EVPs, with respect to the PoIs belonging to the other two classes of relevance (OVPs and MVPs - which can bareley be seen): this is an evidence of the fact that the user always visits new locations, but only few of them are visited regularly. In the lower plot, we zoom on the classes OVPs and MVPs: the number of OVPs is limited and its average value is 4.19; also for the MVPs the number per user is limited, and its average value is 1.76. As expected, each user has a very small number of preferred locations (MVPs) which are visited daily (e.g., home, work place), and a higher but still limited number of location of interest (OVPs) which are visited with a lower frequency but regularly (e.g., gym, favourite restaurant, parent's house).

Figure 5 shows the average visit time to the PoIs, according to their class of relevance. From the figure we notice that for all users, the average visiting time to EVPs is very limited and on average lower than one hour. The average visiting time for OVPs and EVPs depends to the mobility style of the user: some users tend to spend long time in their MVPs, other users instead, use to have very long visits to the OVPs. We will talk about the classification of the user's behaviour below in this paragraph. However, considering the PoIs classification, the MVPs and OVPs can be considered equally relevant for the user, even if the MVPs are visited more frequently and more periodically than the OVPs. The EVPs are instead locations not really important to the user, and where (according to the figure) it spends on average a shorter interval of time.

In Figure 6 we represent a cumulative measure of the percentage of the total time each user spends visiting PoIs belonging to the three different classes of relevance. According to this figure, a user tends to spend more than half of the total time in the MVPs and the rest of the time is almost equally distributed between the EVPs and the OVPs.

The interesting aspect of this feature is further exploited in Figure 7, where we show the *percentage of the visit time per class*, for some users. While always showing the three classes pattern, the behaviour of the two users radically differs. The user 69 has a very stay-at-home behaviour: it spends close to the 81% of the time in the MVPs, and less than 9% in the EVPs.

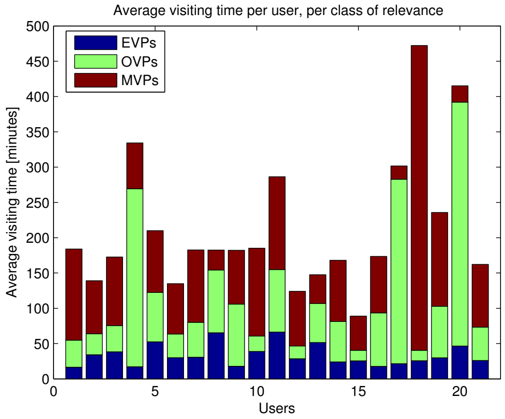


Figure 5: Average visiting time per class of relevance

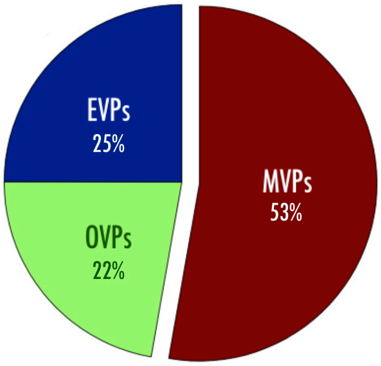


Figure 6: Percentage of the visiting time, per class of relevance (Cumulative)

As opposite, user 25 is a globetrotter: the percentage of time spent in the MVPs is below 10% (rounded to 10% in the figure), and the user spends most of the time in the EVPs (close to 73%), even if the average time spent in each EVP is still significantly smaller than the average time spent in each MVP. This opens for new research approaches to human mobility based on visited places distribution, and is matter of ongoing work.

6. CONCLUSION

People visited places are regulated by statistical laws that tell us that each person visits very few places very frequently and very few places occasionally. We have experimentally validated this property on a large dataset and derived 3 classes, the MVPs, OVPs and EVPs, that reflect those laws. We further extracted a relation between the time spent within each class and the type of human mobility of the user. Our future work includes further validation on ad hoc traces, as well as further elaboration of the human mobility classification, considering different timing features (e.g., circadian rhythm and different seasons).

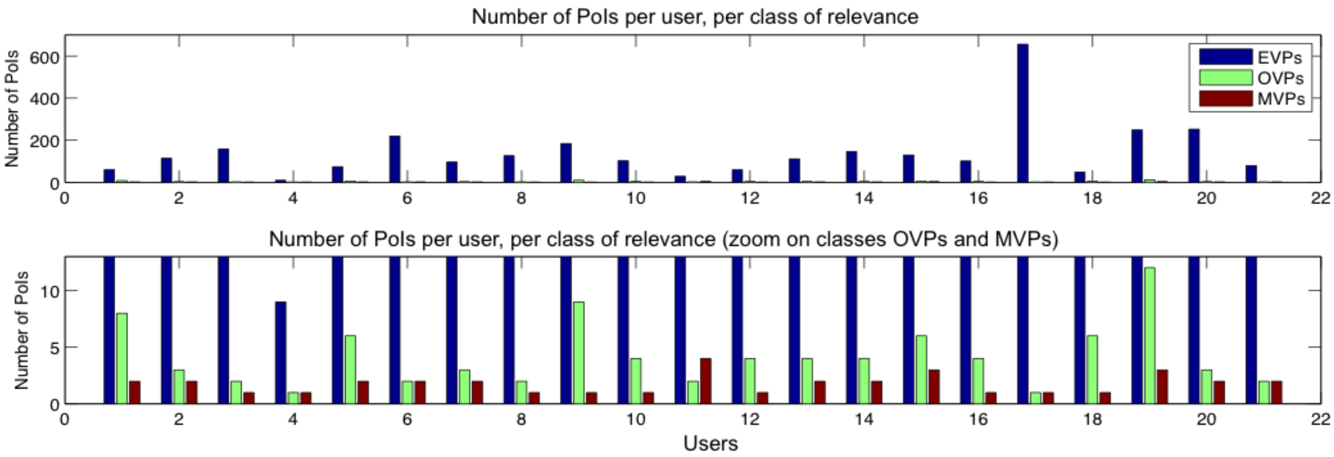


Figure 4: Number of PoIs per class of relevance

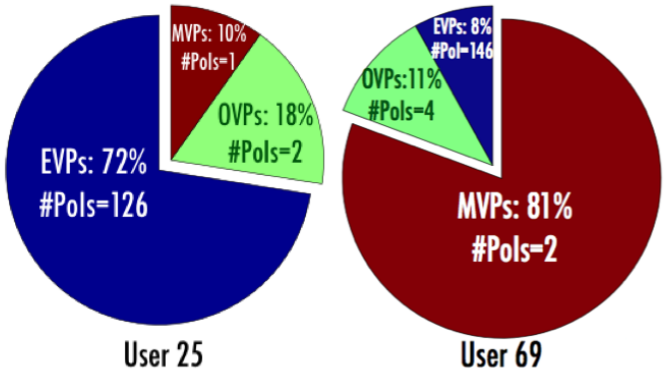


Figure 7: Different types of user mobility: the globetrotter (user 25) and the stay-at-home (user 69) behaviour derived by the time and number distribution of PoIs.

Acknowledgment

This work is partially supported by the EU project SCAMPI and by the Italian Ministry for Instruction, University and Research under the PRIN PEOPLENET (2009BZM837) Project.

7. REFERENCES

[1] G.B. Colombo, M.J. Chorley, M.J. Williams, S.M. Allen, and R.M. Whitaker. You are where you eat: Foursquare checkins as indicators of human mobility and behaviour. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, 2012.

[2] R.I.M. Dunbar. Neocortex size as a constraint on group size in primates. In *Journal of Human Evolution*, volume 22, pages 469–493, June 1992.

[3] R.I.M. Dunbar. The social brain hypothesis. In *Evolutionary Anthropology*, volume 6, pages 178–190, 1998.

[4] Martin Ester, Hans P. Kriegel, Jorg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, 1996.

[5] Anna Förster, Kamini Garg, Hoang Anh Nguyen, and Silvia Giordano. On context awareness and social distance in human mobility traces. In *Proceedings of the third ACM international workshop on Mobile Opportunistic Networks*, MobiOpp '12, 2012.

[6] Silvia Giordano and Michela Papandrea. Four places where i can be. Technical report, ISIN DTI SUPSI, June 2012.

[7] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[8] Miao Lin, Wen-Jing Hsu, and Zhuo Qi Lee. Predictability of individuals' mobility with high-resolution positioning data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, 2012.

[9] M. Papandrea and S. Giordano. Enhanced localization solution. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, 2012.

[10] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. In *Nature*, volume 393, pages 440–442, 1998.

[11] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, 2009.

[12] M. Zignani. Geo-comm: A geo-community based mobility model. In *Wireless On-demand Network Systems and Services (WONS), 2012 9th Annual Conference on*, 2012.

[13] M. Zignani and S. Gaito. Extracting human mobility patterns from gps-based traces. In *Wireless Days (WD), 2010 IFIP*, pages 1–5. IEEE, 2010.

[14] M. Zignani, S. Gaito, and G. Rossi. Extracting human mobility and social behavior from location-aware traces. *Wireless Communications and Mobile Computing*, 2012.