Comparing general mobility and mobility by car

Luca Pappalardo*‡, Filippo Simini†, Salvatore Rinzivillo‡, Dino Pedreschi* and Fosca Giannotti‡
*KDD Lab, Department of Computer Science
University of Pisa, Italy
Email: {lpappalardo, pedre}@di.unipi.it
†Institute of Physics, Budapest University of Technology and Economics
Budapest, Hungary
Email: f.simini@neu.edu
‡KDD Lab, ISTI-CNR
Pisa, Italy
Email: {rinzivillo, fosca.giannotti}@isti.cnr.it

***Abstract*—In the last years, the emergence of big data led scientists from diverse disciplines toward the study of the laws underlying human mobility. Although these recent discoveries have shed light on very interesting and fascinating aspects about people movements, they are generally focused on global and general mobility patterns. For this reason, they do not necessarily capture phenomena related to specific types of mobility, such as mobility by car, by public transportations means, by foot and so on. In this work, we aim to compare general human mobility with mobility expressed by a specific conveyance, trying to address the following question: What are the differences between general mobility and mobility by car? To answer this question, we present the results of an analysis performed on a big mobile phone dataset and on a GPS dataset storing information about car travels in Italy.**

I. Introduction

In the last few years, the emergence of big data led scientists from diverse disciplines toward the study of human mobility, helping to discover and understand the hidden patterns in the trajectories people follow during their daily life. Such a social microscope showed that traditional mobility models adapted from the observation of particles or animals (such as Brownian motion and Lévy-flights) [1], [2] and recently from the observation of dollar bills [5], are not suitable to describe people's movements. Indeed, at a global scale humans are characterized by a huge heterogeneity, since a power law was observed in the distribution of the characteristic distance traveled by users [3], [4]. Despite the observed heterogeneity in people's movements, through the observation of past mobility history the whereabouts of most individuals can be predicted with a very high accuracy, higher than 80% [6], [7].

These recent discoveries have undoubtedly shed light on very interesting and fascinating aspects about human mobility. However, they are generally focused on global and *general* mobility patterns: through the analysis of GSM and other types of data describing travels with different transportation means, researchers revealed that our movements are not random, but follow their own laws. Since such laws are very general, they do not necessarily capture phenomena related to specific types of mobility. To clarify this point, let us consider movements by bike. Bikes are convenient and efficient transportation means to use within a city, but they are not suitable to cover very large distances. For this reason, while the accuracy in the predictability of bikers could not differ so much from the general pattern, the variability with respect to the characteristic traveled distance is presumably much lower, leading to a different mobility pattern.

The aim of this paper is to compare general mobility with mobility by car, trying to answer the following question: What are the differences between general mobility patterns and patterns of car travel? To address this question, we exploit a big mobile phone dataset collected by a European mobile phone carrier and a GPS dataset consisting of the detailed spatio-temporal trajectories of travels performed by cars in Italy. By exploiting these data, we show the difference between global mobility and mobility by car in two important aspects: the distribution of radius of gyration and the distribution of time spent in the visited locations.

The rest of the paper is organized as follows: Section II describes the dataset and highlights the main difference between GSM and GPS data. Section III briefly describes the individual mobility measures we used to unveil the patterns, while Section IV presents a comparison between GSM and GPS patterns. Finally, Section V concludes the paper, providing some conclusions.

II. Data description: GSM vs GPS

Mobile phones are nowadays very common technological devices offering a good proxy to capture individual trajectories. Indeed, each time a user makes a call the carrier records the tower that communicates with the phone, effectively pinpointing users' location. Unfortunately, this information is not terribly accurate because an individual could be anywhere within the tower's reception area, which can span tens of square kilometers. Furthermore, the location is usually recorded only when a person uses her phone, providing little information about the whereabouts between calls. Since call patterns are bursty [8], for most of the time the actual position of a user is unknown.

In the current study, we exploit a GSM dataset collected by a European mobile phone carrier for billing and operational purposes. It contains temporal (date and time) and spatial (the cell phone tower's coordinates) information of all calls and text messages sent by 3 million costumers[1]. Table I shows an

[1] to guarantee anonymity, each user is identified with an anonymized security key.

example of phone records. In order to select the most reliable users for our purpose, we restricted our period of observation to three months and applied some filters to the data. For each user, we discarded locations visited only once during the entire period of observation, and those with a number of calls less than 0.05% of the total[2]. Then, from the resulting dataset we deleted all users who visited only a single location, and those with a call frequency less than twelve calls per day on average during the period of observation. The filtering phase resulted in a final dataset of $67,000$ active users.

Timestamp	**Coordinates**	**Caller**	**Callee**	**Type**
2008/04/01 - 23:45:00	$(32.567, -2.642)$	A45J23	F45J23	SMS
2008/04/02 - 06:02:10	$(33.282, -2.221)$	K65232	V56YT4	Call
...	...	...	...	...

TABLE I. EXAMPLE OF PHONE RECORDS IN THE GSM DATASET.

Unlike GSM data, GPS traces provide high resolution location data, storing the geodetic coordinates with an average sampling rate of few seconds. Even though these features are ideal in making a refined statistical analysis of human mobility patterns, relatively few works in literature are based on GPS data, mainly due to the difficulty to obtain complete traces covering movements along the whole day. In this work, we have access to a GPS dataset that stores information of approximately 9.8 Million different car travels from $159,000$ cars tracked during one month (May 2011) in an area of 250km×250km in central Italy. The GPS traces are collected by Octo Telematics Italia Srl[3], a company that provides a data collection service for insurance companies. Since GPS data do not provide explicit information about visited locations, we assigned each origin and destination point of the travels to the corresponding Italian census cell, according to information provided by the Italian National Institute of Statistics[4]. After such aggregation, many users are found to have only one visited location. We discarded them and took into account only those users with the most frequent location (most likely their home or work) inside the region of Tuscany. These filtering operations produced a dataset of $46,121$ users, where a travel is described by a timestamp and a pair of coordinates corresponding to the centroids of the origin and destination cells of the travel (Table II).

Timestamp	**Origin**	**Destination**	**Car**
2011/05/12 - 08:31:20	$(32.567, -2.546)$	$(32.7, -2.511)$	F45J23
2011/05/24 - 17:53:08	$(32.1982, -2.333)$	$(33.123, -2.31)$	H2705L
...	...	...	...

TABLE II. EXAMPLE OF RECORDS IN THE GPS DATASET.

Table III summarizes a few properties of the two datasets described above. It is worth noting that GPS data provide us information about displacements performed by car only. For this reason, we have a partial knowledge about the whole mobility of individuals. Conversely, GSM data may provide information about travels made using all transportation means, although they are actually recorded only when a user calls before and after the trip.

Dataset	**Volume**	**Conveyance**	**Precision**
GPS	46,121 users	Cars	High
GSM	67,000 users	Many	Low

TABLE III. SUMMARY OF PROPERTIES OF OUR DATASETS.

III. MOBILITY MEASURES

In order to explore the statistical properties of the mobility patterns, for each user we computed several individual mobility measures.

The **center of mass** r_{cm} of a user represents the pivot of her individual mobility. Mathematically, it is a two-dimensional vector representing the weighted mean of the visited locations:

$$\vec{r}_{cm} = \frac{1}{W} \sum_{i=1}^{L} w_i \vec{r}_i \qquad (1)$$

where L is the total number of distinct towers/cells visited by the user/car; $\vec{r}_i$ is a two-dimensional vector representing the geographic coordinates of tower/cell i; w_i is the weight assigned to location i; and W the sum of the weights over all locations. Depending on the measure considered to evaluate the weight of a location, two different centers of mass can be defined. The *frequency-based* center of mass weights locations according to their visitation frequency, hence w_i is the number of calls/arrivals performed in location i. In the *time-based* center of mass we take w_i as the total time spent by the user in location i.

Another interesting measure of an individual's central position is the most frequent location L_1, i.e. the location where she can be located with the highest probability, which is most likely the home or work place. Such measure can be computed in a very straightforward way by simply taking the tower/cell from which the user performs the highest number of calls/arrivals.

The *radius of gyration* of a user [3], [4] is a mobility measure representing the characteristic distance traveled by each individual. It is a concept borrowed from physics, defined as the root mean square of the weighted sum of all locations' distances from the center of mass:

$$r_g = \sqrt{\frac{1}{W} \sum_{i=1}^{L} w_i (\vec{r}_i - \vec{r}_{cm})^2} \qquad (2)$$

where $\vec{r}_{cm}$ is the vector of coordinates representing the center of mass. We computed two types of radius of gyration: i) r_g with respect to the frequency-based center of mass, i.e. w_i is the total number of calls/arrivals in i; ii) r_g with respect to the time-based center of mass, i.e. w_i is the total time spent in i.

IV. RESULTS

In this section, we compare the patterns found on the two datasets, highlighting the main differences between general mobility and mobility by car.

The first aspect we investigated is the difference between the frequency-based radius of gyration and the time-based one. In other words, how does the choice of locations' weight influence the value of the radius of gyration? Figure 1 shows the scatter plots of frequency-r_g versus time-r_g, for GSM (left) and GPS (center) users. For a better understanding of

[2] this means that a location i is deleted if $n_i/N < 0.005$, where n_i is the number of calls performed in the tower i, and N the total number of calls performed by the user.

[3] http://www.octotelematics.it

[4] http://www.istat.it

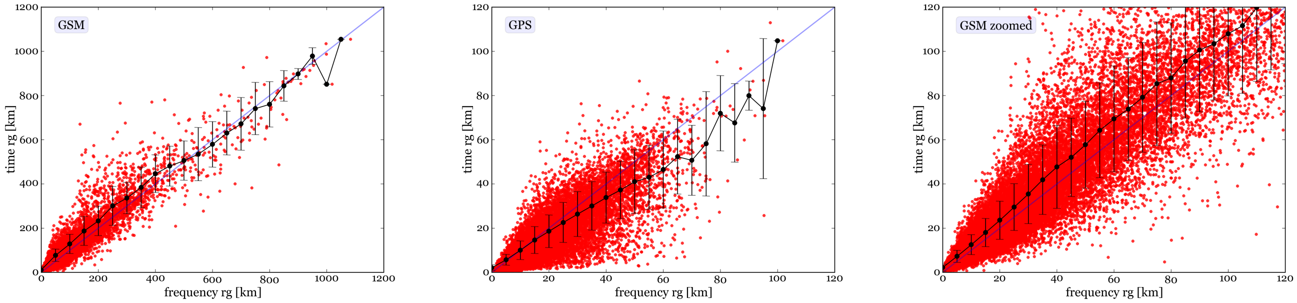

Fig. 1. Scatter plots of frequency-r_g versus time-r_g for GSM (left) and GPS (center) users. A zoomed version of the GSM scatter plot is proposed on the right. Error bars use bin size of 50 km (left) and 5 km (center and right).

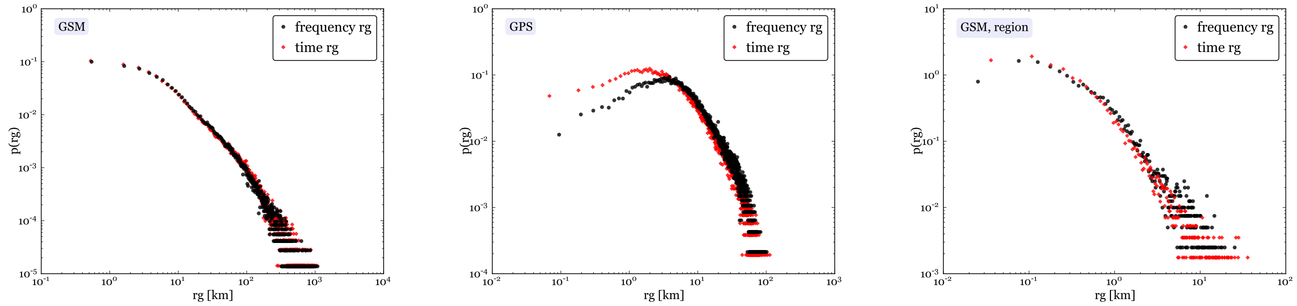

Fig. 2. Frequency- and time-based probability distributions of r_g for GSM (left) and GPS (center) users. On the right, a GSM plot for users with the most frequent location L_1 in a region with size comparable to Tuscany is provided (only locations inside the region are considered in the computation of the radii).

the underlying correlation, also error bars are drawn, with bin size of 50km (GSM) and 5km (GPS). At a first glance, in both cases the measures seem to be correlated, as expected. However, from a closer examination an interesting difference emerges: while in the GSM case the mean of the error bars tends to be biased toward the time-r_g, in GPS data the mean frequency-r_g tends to be higher. One possible interpretation of the phenomenon is that car visitation frequency of locations distant from the center of mass is higher, leading to a bigger characteristic traveled distance.

Figure 2 (left) shows the distributions of frequency- and time-based radii of gyration computed on the GSM dataset. There is no significant difference between the two curves, which practically coincide. Conversely, a sharp difference clearly emerges from the distributions of radii in the GPS case (Figure 2, right). Indeed, the time-based distribution is shifted towards shorter radii, and peaks at 2km instead of 5km. This aspect confirms the prominent role of frequency with respect to time observed in Figure 1 (center), suggesting that cars are usually parked for a long time in locations close to the center of mass (like home and work locations). This effect is absent in GSM data because people can continue their call activity even while being stationary. Another difference we can notice is that while the GSM curves decrease over the entire range, GPS radii show a growing value up to ≈2km. This is presumably due to the tendency of covering those small distances by foot, bike, or bus, resulting in a lower probability to find such travels in the GPS dataset.

To test at what extent the differences in the distribution of r_g are due by the geographic scale (GSM data refers to a whole country, while GPS to a single region), we computed the r_g distribution of those GSM users having the most frequent location in a region of the country of a size and population comparable to Tuscany. In the computation of the radius, only the locations within the region are taken into account. As Figure 2 (right) shows, the distribution of the radius does not change significantly, suggesting that the shape of the curves, and their slopes, are rather independent from the scale and are related to the portion of mobility they represent.

The time spent across the visited locations is another interesting mobility aspect it is worth investigating. Figure 3 (left) shows the GSM distribution of time spent for the five most frequent locations $L_1, \ldots, L_5$. As we can see, the time spent is clearly unbalanced in favor of the most important location L_1. This is reasonable, because the most frequent location usually corresponds to user's home or work place, which are the locations where an individual spends most of the time. The plot also suggests that time is proportional to frequency: the more a user visit a location, the more time she spends there. Such phenomenon is confirmed by Figure 4, where the correlation, though not perfectly linear, is evident.

The same pattern is also observed on GPS data, although the difference between L_1 and the other locations is less sharp (Figure 3, right). It is worth to note that in both plots, L_1 intersects the other curves approximately at the same points. This is very interesting because, independently from the geographical scale and from the portion of mobility considered, beyond a certain fraction of time is much more likely for a user to be located in the L_1 than all the other locations.

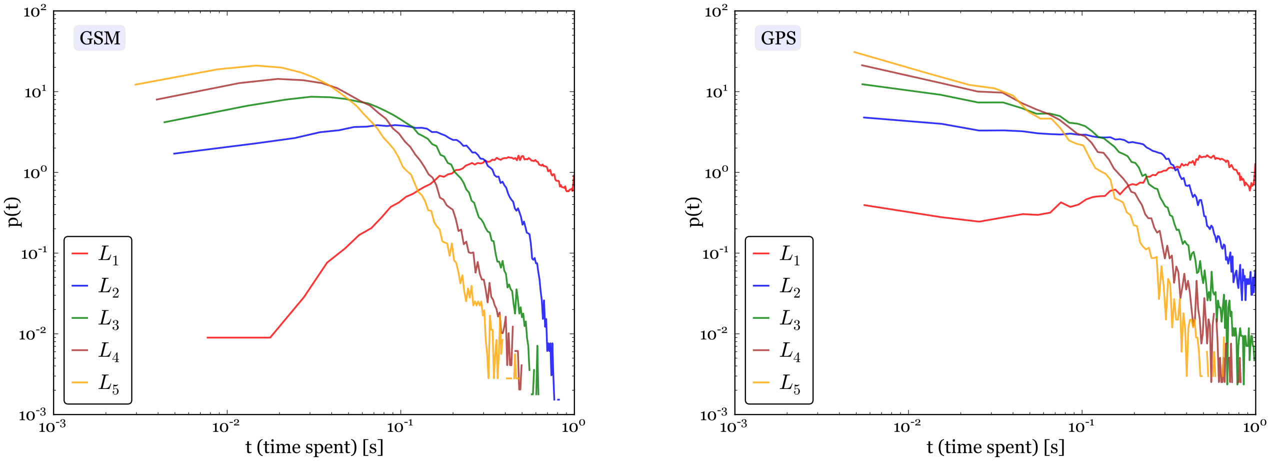

Fig. 3. Distribution of fraction of time spent in the five most frequent locations $L_1, \ldots, L_5$ for the GSM (left) and the GPS (right) datasets.

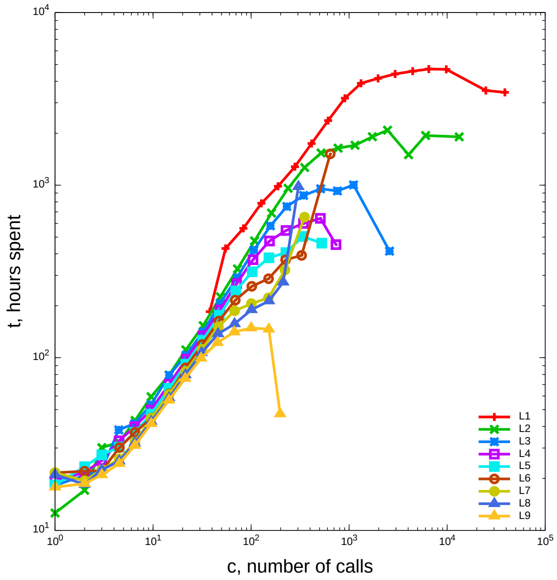

Fig. 4. Correlation between calls and time spent in the nine most frequent locations (GSM dataset).

V. Discussion and Conclusion

Our data-driven analysis showed the difference between general mobility and mobility by car regarding two main aspects: the distribution of radius of gyration and the distribution of time spent in the visited locations. We discovered that, regardless the geographic scale, the shape of the r_g distribution is different since mobility by car tends to poorly cover displacements within small distances. However, in both cases the distribution of time spent in visited locations present a clear dominance by the most frequent location L_1, with such dominance more pronounced in the GSM dataset. Moreover, the L_1 curve intersects the others approximately at the same points in both cases, hinting the presence of a general pattern underlying the phenomenon.

Acknowledgment

The authors wish to thank the company Octo Telematics Italia Srl for providing the car travels GPS data.
The research reported in this article has been partially supported by European FP7 project DATASIM (http://www.datasim-fp7.eu/).

References

[1] G. M. Viswanathan et al., Lévy flight search patterns of wandering albatrosses. Nature 381, 413-415 (1996).

[2] G. Ramos-Fernandez et al., Lévy walk patterns in the foraging movements of spider monkeys, Behavioral Ecology and Sociobiology 55, 25 (2003).

[3] M. C. González, C. A. Hidalgo, A.-L. Barabási, Understanding individual human mobility patterns, Nature 453, 779-782.

[4] L. Pappalardo, S. Rinzivillo, Z. Qu, D. Pedreschi, F. Giannotti, Understanding the patterns of car travel, European Physics Journal Special Topics 215, 61-73 (2013).

[5] D. Brockmann, L. Hufnagel, T. Geisel, The scaling laws of human travel, Nature 439, no. 7075, 462-465 (2006).

[6] N. Eagle, A.S. Pentland, Eigenbehaviors: identifying structure in routine. Behavioral Ecology and Sociobiology 63, 1057-1066 (2009).

[7] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility. Science 327, 1018-1021 (2010).

[8] A.-L. Barabási, The origin of bursts and heavy tails in humans dynamics, Nature 435, 207 (2005).