Soc. Netw. Anal. Min. (2016) 6:59
DOI 10.1007/s13278-016-0369-2

ORIGINAL ARTICLE

Unveiling mobility complexity through complex network analysis

Riccardo Guidotti[1] · Anna Monreale[1] · Salvatore Rinzivillo[2] · Dino Pedreschi[1] · Fosca Giannotti[2]

Received: 30 November 2015 / Revised: 24 June 2016 / Accepted: 20 July 2016 / Published online: 12 August 2016

Abstract The availability of massive digital traces of individuals is offering a series of novel insights on the understanding of patterns characterizing human mobility. Many studies try to semantically enrich mobility data with annotations about human activities. However, these approaches either focus on places with high frequencies (e.g., home and work), or relay on background knowledge (e.g., public available points of interest). In this paper, we depart from the concept of frequency and we focus on a high level representation of mobility using network analytics. The visits of each driver to each systematic destination are modeled as links in a bipartite network where a set of nodes represents drivers and the other set represents places. We extract such network from two real datasets of human mobility based, respectively, on GPS and GSM data. We introduce the concept of *mobility complexity* of drivers and places as a ranking analysis over the nodes of these networks. In addition, by means of community discovery analysis, we differentiate subgroups of drivers and places according both to their homogeneity and to their mobility complexity.

Keywords Mobility network · Ranking · Communities

1 Introduction

One of the most fascinating challenges of our time is to understand the complexity of the global interconnected society and possibly to predict human behavior. A great part of human behavior is observable through individual movements, registered in many different layers: mobile phone network, GPS devices, social media applications, road sensors, credit card transactions, etc. Movement is the "hardware" of our daily life. We move to perform any activity: we have to move to bring children at school, to buy a new electronic device, to meet with colleagues at work, etc. If we understand the patterns of human movement, we can also comprehend the mechanics of human behavior.

On the basis of this assumption, in the last years, we have witnessed many studies exploring movements data to understand different aspects related to the mobility of individuals, such as the density of traffic (Giannotti et al. 2011), the identification of systematic movements (Trasarti et al. 2011), the identification of groups of drivers following common routes (Monreale et al. 2009) and many others. On one hand, the movement is an objective phenomenon that can be observed, measured, and recorded easily with the modern ICT services. On the other hand, the intended activity of each movement is not always easy to sense and register. A common approach to better understand movement behavior consists into the study of the motivations that push an individual to move toward a given

✉ Riccardo Guidotti
riccardo.guidotti@di.unipi.it

Anna Monreale
anna.monreale@di.unipi.it

Salvatore Rinzivillo
salvatore.rinzivillo@isti.cnr.it

Dino Pedreschi
dino.pedreschi@di.unipi.it

Fosca Giannotti
fosca.giannotti@isti.cnr.it

[1] KDDLab, University of Pisa, Largo B. Pontecorvo, 3, 56127 Pisa, Italy

[2] KDDLab, ISTI-CNR, Via G. Moruzzi,1, 56124 Pisa, Italy

destination. There are proposals in the literature to semantically enrich movement data on the basis of movement dynamics and properties. For example, Jiang et al. (2012) tries to estimate home/work locations of an individual by analyzing the frequency she visits a particular place; Lafferty et al. (2001) observe a sequence of movements to derive the sequence of activities performed; Rinzivillo et al. (2014) extract a series of individual mobility network to learn structured patterns of visits to places; and Furletti et al. (2013) exploit the background knowledge of the points of interest (POIs) available in a territory to derive the activities of persons stopping nearby.

In this paper, we propose an approach that can be considered as an intermediate step between the movement dynamics exploration and the semantic enrichment of movements. We start from the analysis of individual movements to understand the relevance of each destination. However, we are not interested in the specific activity a person is performing on her destination, rather we focus on the "relevance" that a specific destination has for the person.

A well-known proverb says that "Home is where the Hearth is," meaning that the *home* for an individual is not just a mere geographical place, but it represents a complex mixture of sensations, perceptions, and feelings linked to that place. It goes without saying that this kind of definition is strongly tied to a personal and subjective vision of that place. From the analytical point of view, it is difficult to *measure* this perception. The approaches based on semantic enrichment are focused either on places of general interest (like restaurants, shopping center) or on individual-based destinations (like home or work). Our proposal tries to fill this gap by starting from an individual ranking of personal places to generalize to collective relevance of destinations.

Concretely, we propose an approach based on complex network analytics methods to model the relevance of a place p according to the persons visiting p. The basic intuition is based on the concept of *complexity of individual mobility*: a person d is complex if she visits many different complex places. In a similar way, a place p has a high relevance, i.e., it is complex, if it is visited by many complex visitors. This interwined relation among users and places is modeled by means of a bipartite graph, called *Drivers–Places network*. Starting from this model, we propose two analytical processes based on ranking measures and community discovery. In the first process, we try to understand both the mobility complexity of people moving in a territory and the mobility complexity of places for the collectivity. Therefore, the analysis is focused on the mobility behavior of drivers with respect to some specific places, which are considered important for both their individual mobility and the collective mobility, and on the mobility in the interesting places with respect to the drivers who visit them. In the second analytical process, based on application of community discovery algorithms, we characterize the groups of similar drivers and places with respect to mobility complexity.

We experiment our analytical methodology in real case studies considering both *GSM* and *GPS* datasets of trajectories. Our finding is that drivers and places complexity in terms of mobility can be characterized according to the similarity of the movements that lead a certain user in a certain location. Then, by doing a deeper analysis with GPS data, we show how certain communities are characterized by their topological structure and by their mobility. Finally, as additional point, studying ranking measures we demonstrate that the method we use to calculate the mobility complexity scores is a particular case of HITS (Kleinberg et al. 1999), one of the most famous link analysis algorithms.

The rest of this paper is organized as follows. Section 2 discusses papers related with our work. In Sect. 3, we introduce some basic concepts useful to understand our analytical methodology. Section 4 illustrates the process of bipartite network Driver–Place construction, while Sect. 5 explains in detail the idea of mobility complexity. In Sects. 6 and 7, we present the experimental results obtained in the two case studies using real-life GPS and GSM data. Finally, Sect. 8 contains conclusions and describes future works.

2 Related work

In this section, we discuss some papers of the literature which are related to our work. First, we summarize some works which analyze mobility locations using a complex network approach. Then, we discuss other works related to link analysis methods and the analysis of bipartite networks in economic scenarios.

The mobility history of a driver may enable many services such as location recommendation or sales promotion. In Zheng and Xie (2010), by taking into account users travel experience and the subsequent locations visited, the authors learn the location correlation from GPS trajectories useful to construct a personalized location recommendation system. Also our approach extracts a correlation between drivers and places, and among the drivers themselves and places themselves.

In Brilhante et al. (2012), the authors analyze the urban mobility trying to feature the places in a city according to how people move among them. The authors build a network of points of interests by connecting places by the individual trajectories passing through them. From such network, they compute communities finding groups places highly connected by the mobility of the individuals. The main difference with our approach is that we try to characterize the relevance of the places with respect to the drivers and vice-versa extracting from the movements data their importance without the need of external data sources.

Mobility networks can be also employed to prevent the spread of diseases. In Eubank et al. (2004) from movements of individuals between specific locations, the physical contact patterns are modeled by dynamic bipartite graphs. The study found that this network is strongly connected with a well-defined scale for the degree distribution and that the locations graph is scale-free.

In Hossmann et al. (2011), the authors represent the mobility scenario by a weighted contact graph, where a tie strength represents how long and often a pair of nodes is in contact. This enables the mobility analysis by complex network and graph theory. Similarly to us, they found that mobility is strongly modular by using community detection. However, their finding is that communities are not homogeneous entities, while we will show that there exist both homogeneous and heterogeneous communities.

An interesting analysis on mobility data presented in Pappalardo et al. (2015) discover two distinct classes of individuals: *returners*, whose mobility is produced by the commuting between home location and work location, and *explorers*, whose mobility is generated by travels performed toward locations different from home and work and far from them. This work shows that returners and explorers play a distinct quantifiable role in spreading phenomena and that there exists a correlation between their mobility patterns and social interactions.

A completely different type of mobility is discussed in Kaluza et al. (2010) where it is built the network of ports by using the itineraries of cargo ships. This network has a heavy-tailed distribution for the connectivity of ports and for the loads transported on the links with systematic differences between ship types. Also in our work, we delineate some characteristics given by certain mobility patterns.

Complex networks are a powerful model to study and describe realities with different components. In Hidalgo and Hausmann (2009), the authors present a simple method to infer the relative number of inputs available in a country from trade data connecting countries to the products they export. They show that countries approach over the long run a level of income that is determined by the diversity of inputs available in the country, as approximated by the measures introduced. The same authors in Hidalgo and Hausmann (2010) develop a method to characterize the structure of bipartite networks called *Method of Reflections* (MOR), and they apply it to trade data to illustrate how it can be used to extract relevant information about the availability of capabilities in a country. They interpret the variables produced by MOR as indicators of economic complexity.

Furthermore, other authors faced the same macro-economical study with a slightly different approach. Caldarelli et al. (2011, 2012) analyzed the bipartite network of countries–products from United Nations data on country production. The authors define the country–country and product–product projected networks and introduce a novel method of filtering information based on elements' similarity. As a result, they find that country clustering reveals unexpected socio-geographic links among the most competing countries.

Other works use a bipartite graph to observe micro-economical relationships. In Pennacchioli et al. (2013), the authors inspect the market basket transactions observed over a large population for long time, offering a detailed picture of customers' shopping activity. They use the system of all customer–product connections and MOR to better understand the hidden knowledge governing the interplay between human desires and needs on one hand, and the offered goods and products on the other hand. They create a framework to exploit the characteristics of the customer–product matrix and test it on a transaction database storing purchases in supermarkets.

3 Preliminaries

In this section, we introduce notions and procedures from the state of art of mobility data mining that are employed in our approach to extract the places used for the construction of the *Driver–Place network*.

3.1 Systematic movements: mobility profiles

Movements are performed by *users* or *drivers* in specific areas and time instants, and each movement is composed by a sequence of spatio-temporal points. We call *trajectory* the movements of a driver described by a sequence of spatio-temporal points. The set of the trajectories traveled by a driver makes the driver's *individual history*. Given a driver i, we call *individual history* the set of *trajectories* $H_i = \{m_1, \ldots, m_n\}$.

The profiling procedure proposed in Trasarti et al. (2011) allows us to extract the systematic movements of a driver i. Applying this procedure, the trajectories can be grouped using a density-based clustering equipped with a *distance function* defining the concept of trajectory similarity. The result is a partitioning of the original dataset $\mathcal{C} = \{C_1 \ldots C_k\}$ where $C_c \subset H_i \; \forall C_c \in \mathcal{C}$. The *clusters* with few trajectories and the one containing noise are filtered out. *Representative trajectories* called *routines* r_c are extracted from each remained cluster. This set of routines is called *mobility profile* $S_i = \{r_1 \ldots r_k\}$ of driver *i*. The parameters required by the procedure in Trasarti et al. (2011) are: (1) *min size* representing the minimum size for a cluster of trajectories and (2) ε_r representing the threshold distance to consider two trajectories belonging to the same cluster.

The *mobility profile* describes an abstraction in space and time of the systematic movements of the drivers completely ignoring exceptional movements. Thus, the systematic behavior of each driver can be modeled with her mobility profile, and the daily mobility of each driver is characterized by her routines. Figure 1 depicts an example of mobility profile extraction.

3.2 Systematic places: mobility POIs

The routines extracted following the procedure (Trasarti et al. 2011) necessarily begin and end somewhere. The systematic profiled drivers have a mobility that gravitates around these locations. Thus, it results that these places are surely very important for them. We employed the procedure proposed in Guidotti et al. (2014) to identify these places called *individual POIs*.

Given the mobility profile S_i of the driver *i*, then, the *individual POIs* of *i* are the set I_i such that $I_i = \{p | \exists r \in S_i . p = start(r) \ \vee \ p = end(r)\}$, where $start(.)$ and $end(.)$ are two functions that given a routine return the start and end point, respectively. We indicate with $\mathcal{IP}$ the union of all the individual POIs. Note that, these POIs are not just "places frequently visited by someone" like restaurants, bar, museums, but they are places relevant in people everyday systematic life. Therefore, they are not only typical attraction points, but also important places for the individual, such as home or work, which are not available in the typical public sources.

Since individual POIs are spatial points represented by GPS coordinates, it is unlikely to observe two points with identical coordinates. Consequently, in order to discover places visited by more than one driver, we need to group close individual POIs in $\mathcal{IP}$ that should be part of the same collective POIs (Guidotti et al. 2014). To this aim, following Guidotti et al. (2014), we compute a density-based clustering on the individual POIs $\mathcal{IP}$ and then, we turn each valid cluster and each noise point into a buffered convex hull area representing a *collective POI*. In other words, we increase the area covered by the clustered points with a spatial buffer that together with the density-based clustering allows us to describe a collective POI by an area and not by GPS coordinates. Indeed, if we consider the extreme case where a cluster contains one single individual POI, without a buffering, we obtain the area covered by the coordinates of the POI.

We denote by $\mathcal{CP}$ the set of collective POIs. The input parameters of this procedure are (1) ε representing the threshold distance to consider two individual POIs belonging to the same collective POI and (2) ε' that is the distance of the buffer. Note that, two different POIs *p* and *q* could be overlapped because of the buffering phase. Anyway, keeping $\varepsilon' < \varepsilon$ ensures that the center of *p* is not included in *q*, otherwise the clustering algorithm would have put them in the same cluster since they would have been distant no more than ε.

The clusters returned can also be composed of noise points because each noise point represents an individual POI supported by at least a routine and thus, it is relevant for at least one driver. In the following, for the sake of simplicity, we call a *collective POI* simply *POI*. In other words, we can think to a POI as a geographical area with a certain extension that is visited frequently by at least one driver. Figure 2a–f illustrates how to extract POIs.

Alternative clustering methods to extract the POIs are Guidotti et al. (2015), Ashbrook and Starner (2003), Cao et al. (2010), Pappalardo et al. (2013), Zheng et al. (2010), and Zhou et al. (2004). However, Guidotti et al. (2014) is preferred since only systematic individual, and collective POIs are extracted automatically discarding the noise.

4 Driver–place network

The problem we face consists in understanding both the mobility of people moving in a territory and the mobility of places which are interesting for the collectivity. Our goal is twofold: we want to analyze (1) the mobility behavior of drivers with respect to specific places which are worth for both their individual mobility and for the collective mobility and (2) the mobility of these valuable places with respect to the drivers who visit them.

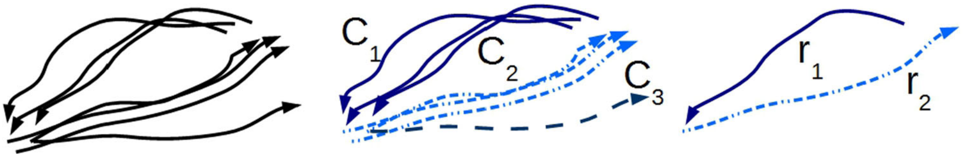


Fig. 1 The *individual history* (*left*), the clusters identified by the grouping function (*center*), and the extracted *individual routines* (*right*) forming the *individual mobility profile*

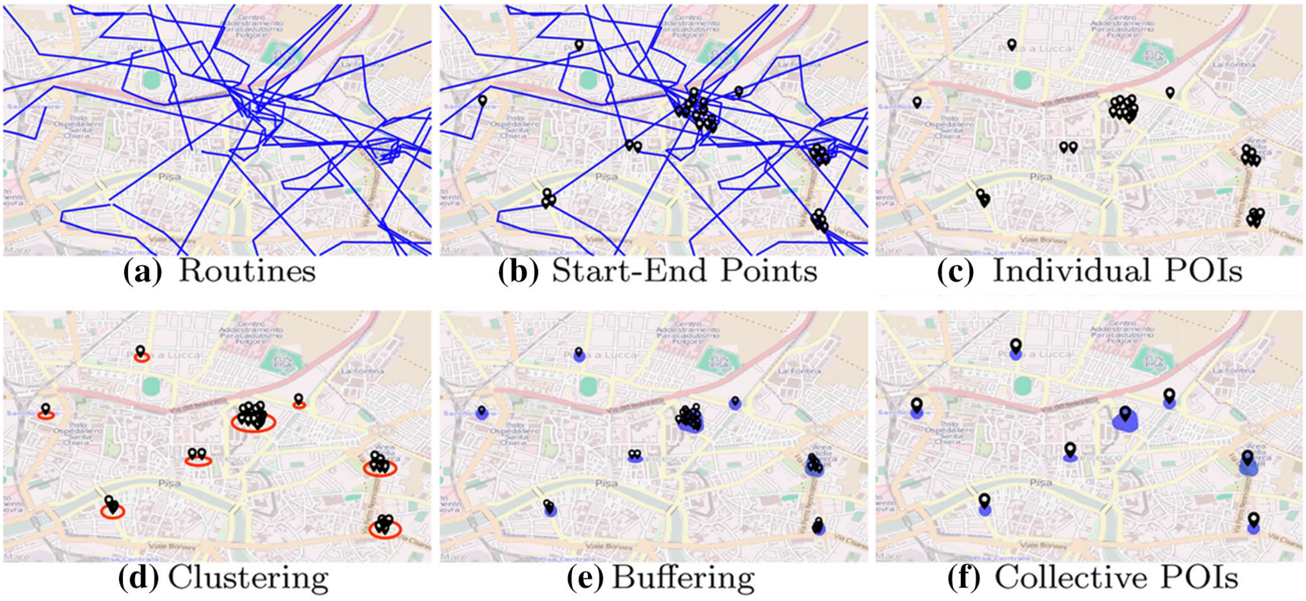


Fig. 2 Sequences of steps to perform the POIs extraction: **a** individual mobility routines, **b** start and end points extraction, **c** individual POIs separation, **d** density-based clustering, **e** buffering phase, **f** the collective POIs

The methodology we propose to address this problem is based on two main steps: (a) the construction of a mobility data-driven network that describes the relationship between *places* and *drivers* and (b) the mobility complexity analysis based on the information modeled by this network.

The mobility data-driven network must capture the information on which places are visited by a specific driver and which drivers visited a specific interesting place. For this reason, we propose to model the relationship between *places* and *drivers* with a bipartite network, named *Drivers–Places network*:

Definition 1 (*Drivers–Places Network*) The *Drivers–Places network* $G = (D, P, E)$ is a bipartite network such that D is the set of drivers, P is the set of places, $D \cap P = \emptyset$, and E is the set of edges $e = (i, j, w)$ where w_{ij} is the number of times driver $i \in P$ stopped in place $j \in D$.

An example of Drivers–Places network is reported in Fig. 3. A Drivers–Places network is composed of two disjoint sets of nodes, i.e., drivers D and places P, such that each link connecting a D-node to a P-node means that driver i visited place j. Moreover, on each edge, we have the information of how many times driver i stopped in place j.

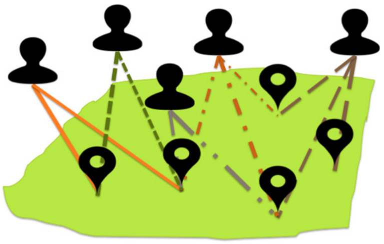


Fig. 3 Example of Drivers–Places network. Every driver is linked to the places visited and, every place is linked to its visitors. The thicker the line the higher the number of visits

Given a Drivers–Places network G, we can represent its adjacency matrix $M^{|D| \times |P|}$ as a rectangular matrix. Indeed, since there are only links between the two partitions (D and P), we do not need to represent the massive number of zeroes given by the links between nodes of the same partition. In M, the rows represent the D-nodes (drivers), while the columns represent the P-nodes (places), thus M_{ij}=1 means that driver i visited place j.

The above bipartite network can be built starting from any dataset of trajectories describing the human mobility and from a set of places which are considered interesting. The crucial point in the network construction is the identification of *interesting places* that compose the set of nodes P. As highlighted above, our goal is to consider places which are interesting both for the individuals and for the collective mobility. For example, we can use as places the set of POIs coming from online static datasets collected by specific websites (Brilhante et al. 2012). In our approach, we consider the POIs extracted directly from the driver movements by applying the method proposed in Guidotti et al. (2014). This gives the not negligible advantage to consider places capturing properties of everyday human mobility both individual and collective.

We illustrate in Algorithm 1 the workflow of the procedure adopted to construct the Drivers–Places network.

Given the drivers mobility $\mathcal{H}$ and the required parameters, we extract the mobility profiles and we derive from the mobility profiles $\mathcal{S}$ the eligible drivers D (i.e., those having a systematic behavior) and the POIs P (*lines 1–11*). Then, considering the driver movements and the POIs (*lines 13–14*), if driver i visited place j, i.e., it exists at least a trajectory of driver i starting or ending in j (*line 15*), then an edge is added to the Drivers–Places network G (*line 17*). Edges are weighted by counting the number of visits w_{ij}, i.e., the number of times driver i visited place j (*line 16*). Moreover, since we want to consider only relevant links we need a mechanism to evaluate how meaningful is the mobility of each driver i for each visited place j, i.e., we want to identify which journeys are significant. We exploit the concept of *lift*, typically applied to association rules (Agrawal et al. 1993), to evaluate how meaningful is the mobility of each driver i toward each visited place j (*line 22*). The output of Algorithm 1 is the bipartite network $G = (D, P, E)$ where D is the set of drivers, P the set of relevant places (i.e., the collective POIs) and E the set of meaningful links (according to the lift filter).

```
Algorithm 1: $buildDriverPlaceNetwork(\mathcal{H}, \theta, ms, \varepsilon, \varepsilon')$
   Input  : $\mathcal{H} = \{H_i, \ldots H_n\}$ - drivers mobility history
            $\theta, ms$ - parameters for mobility profile exaction
            $\varepsilon, \varepsilon'$ - parameters for POIs exaction
   Output: $G = (D, P, E)$ - driver place network
 1 $\mathcal{S} \leftarrow \emptyset$;
 2 $\mathcal{IP} \leftarrow \emptyset$;
 3 for $H_i \in \mathcal{H}$ do
 4     $S_i \leftarrow extractMobilityProfile(H_i, \theta, ms)$;
 5     $I_i \leftarrow extractIndividualPOI(S_i)$;
 6     $\mathcal{IP} \leftarrow \mathcal{IP} \cup \{I_i\}$;
 7     $\mathcal{S} \leftarrow \mathcal{S} \cup \{H_i\}$;
 8 end
 9 $D \leftarrow \{i \mid H_i \in \mathcal{S}\}$;                 /* retrive the drivers' indexes   */
10 $\mathcal{CP} \leftarrow extractCollectivPOI(\mathcal{IP}, \varepsilon, \varepsilon')$;
11 $P \leftarrow \{j \mid POI_j \in \mathcal{CP}\}$;                 /* retrive the POIs' indexes   */
12 $E \leftarrow \emptyset$;                                   /* build the edges   */
13 for $i \in D$ do
14     for $j \in P$ do
15         if $\exists m \in H_i \mid start(m) \in POI_j \wedge end(m) \in POI_j$ then
16             $w_{ij} \leftarrow countVisits(H_i, POI_j)$;
17             $E \leftarrow E \cup \{(i, j, w_{ij})\}$;
18         end
19     end
20 end
21 $G' \leftarrow (D, P, E)$;
22 $G \leftarrow liftFilter(F)$;
23 return $G$;
```

We briefly summarize in the following how the lift coefficient is evaluated in order to remove meaningless edges. We define the total number of visits as $W = \sum_{(i,j) \in E} w_{ij}$, the total number of travels in a certain location done by a driver i as $l_i = \sum_{j \in P} w_{ij}$, and the total number of stops in a place j as $s_j = \sum_{i \in D} w_{ij}$. Given a driver i and a place j, let $\frac{w_{ij}}{W}$ be the relative number of visits done by driver i to place j, $\frac{l_i}{W}$ the relative number of visits done by driver i to all places, and $\frac{s_j}{W}$ the relative number of visits received by place j to all drivers. Then, the lift coefficient of i and j is defined as

$$\mathit{lift}(i,j) = \frac{\frac{w_{ij}}{W}}{\frac{l_i}{W} * \frac{s_j}{W}} = \frac{w_{ij} * W}{l_i * s_j}$$

The lift coefficient takes values from 0 (when $w_{ij} = 0$, i.e., i has never visited j) to $+\infty$. When $\mathit{lift}(i,j) = 1$, it means that $\frac{w_{ij}}{W}$ makes the connection between i and j relevant. Therefore, $\mathit{lift}(i,j) < 1$ means that the event "i visited j" is not significant. The value of 1 for the lift indicator is a reasonable threshold to discern the meaningfulness of the number of visits: if it is strictly higher, then, the mobility is meaningful and the corresponding link is valid, otherwise the mobility is not meaningful. In the following, with the name Drivers–Places network, we refer to the bipartite graph formed by only meaningful links (i.e., $\mathit{lift}(i,j) \geq 1$). In the experiments, we will consider Drivers–Places network from which meaningless links are filtered out.

Finally, it is worth to recall that according to the procedures followed [i.e., (Trasarti et al. 2011; Guidotti et al.

2014)], the trajectories considered, i.e., starting or ending in a POI, are mainly the trajectories belonging to the mobility profiles of the users, i.e., *systematic* trajectories. However, also *occasional* movements ending in every collective or individual POIs are taken into account. Consider for example two friends A and B, and A visited B in the observation period, then also this trajectory will be added to G since A moved from a systematic POI for her (i.e., A's home) to another one that is systematic for the friend B (i.e., B's home).

5 Mobility complexity

A Drivers–Places network describes a detailed picture of the mobility between drivers and places in a certain area. Our goal is to identify a method for discovering users and places that in the Drivers–Places network are characterized by a complex mobility. Intuitively, a user with a complex mobility is a person visiting many *different complex places*, while a complex place is a location visited by many users with a *high mobility complexity*. In other words, the complexity of a place depends on the complexity of people visiting it and viceversa. This means that the definition of mobility complexity requires a recursive evaluation of the phenomenon. Note that, our proposal is to consider that the mobility complexity of an individual does not depend only on the diversity of visited locations, but we require to consider also the complexity of visited locations. Our experiments on real data show that our choice to have a recursive definition of mobility complexity is reasonable (see Sect. 6.4).

In order to clarify the concept of user/place mobility complexity consider the following example. Suppose that Alice's individual POIs are her home, the supermarket where she works and a mall. Now, consider Bob having as individual POIs his home, the farm where he works, his parents' home, a jazz pub. The mobility complexity of Bob is lower than Alice's complexity even if his diversity of visited places is higher. This happens because all Bob's POIs are not complex, while Alice has 2 over 3 complex places.

To understand the hidden knowledge governing the interplay between the most visited places on one hand, and who are the most interesting visitors, and to identify complex users and places with respect to their mobility, we propose to exploit *link analysis*, a data-analysis technique used to evaluate relationships, i.e., connections, between nodes. Among the widely adopted algorithms, there are *PageRank* (Page et al. 1999) and *HITS* (Kleinberg et al. 1999). Since PageRank makes use of a damping factor, it is not suitable for our analysis because we do not want to model random jump between D-nodes and P-nodes and vice-versa. Therefore, HITS would seem more suitable for our analysis.

However, in Hidalgo and Hausmann (2009) and Caldarelli et al. (2011) is presented an ad-hoc link analysis method for bipartite network, called *Method of Reflection* (MOR). Like HITS, it iteratively calculates the value of the previous-level properties of a node's neighbors. MOR is presented both in Hidalgo and Hausmann (2009) and Caldarelli et al. (2011) with slight but significant differences. In this paper, we consider the method proposed in Caldarelli et al. (2011) since it was proven that converge with all the parameter settings.

Consider a *bipartite network* $G = (D, P, E)$ described by the adjacency matrix $M^{|D|\times|P|}$. Let d and p be two ranking vectors to indicate how much a D-node is linked to the most linked P-nodes and how much a P-node is linked to the most linked D-nodes, respectively. Thus, it is expected that the most linked D-nodes connected to nodes with high p_j score have an high value of d_i, while the most linked P-nodes connected to nodes with high d_i score have an high value of p_j. This corresponds to a flow among nodes of the bipartite graph where the rank of a D-node enhances the rank of the P-node to which is connected and vice-versa. Starting from $i \in D$, the unbiased probability of transition from i to any of its linked P-nodes is the inverse of its degree $d_i^{(0)} = \frac{1}{k_i}$, where k_i is the degree of node i. Similarly, the unbiased probability of transition from a P-node j to any of its linked D-nodes is the inverse of its degree $p_j^{(0)} = \frac{1}{k_j}$. Let n be the iteration index, MOR is defined as:

$$d_i^{(n)} = \sum_{j=1}^{|V|} \frac{1}{k_j} M_{ij} p_j^{(n-1)} \forall i \quad p_j^{(n)} = \sum_{i=1}^{|U|} \frac{1}{k_i} M_{ij} d_i^{(n-1)} \forall j \tag{1}$$

These rules can be rewritten as a matrix-vector multiplication

$$d = \bar{M} p \quad p = \bar{M}^T d \tag{2}$$

where $\bar{M}$ is the *weighted adjacency matrix*. From these rules we have

$$d^{(n)} = \bar{M}\bar{M}^T d^{(n-1)} \quad p^{(n)} = \bar{M}^T \bar{M} p^{(n-1)} \tag{3}$$

$$d^{(n)} = \mathcal{D} d^{(n-1)} \quad p^{(n)} = \mathcal{P} p^{(n-1)} \tag{4}$$

where $\mathcal{D}^{(|U|\times|U|)} = \bar{M}\bar{M}^T$ and $\mathcal{P}^{(|V|\times|V|)} = \bar{M}^T\bar{M}$ are related to $x^{(n)} = Ax^{(n-1)}$ that is, MOR is solvable using the *power iteration method* (Lanczos 1950). This fact leads automatically to the proof of convergence.

Using MOR we can interpret the variables produced as indicators of *mobility complexity*. In practice, mapping the definition of MOR on the Drivers–Places network we obtain a mutual reinforcing definition of mobility complexity: a driver with an high mobility complexity visits places with an average high mobility complexity; a place with an high mobility complexity is visited by drivers with an average high mobility complexity. In Appendix, we formally proved that MOR is a particular case of HITS. Thus, can be used both HITS or MOR to characterize the structure of the network and to evaluate nodes ranking for our Driver–Places network. However, we decided to use MOR because useless scores are not calculated (see Appendix) and because of the similarity between our application and those on the networks presented in Caldarelli et al. (2011) and Hidalgo and Hausmann (2010). In the following, we will use d and p to indicate driver and place mobility complexity, respectively.

6 Case study on GPS data

To discover the latent knowledge in the relationship between drivers and places, we applied the methodology described above on datasets of trajectories. First of all, we briefly report some consideration about the dataset used and the mobility profile extraction. Then, we describe the study performed to extract reliable Places as POIs and what they represent on the analyzed area. Moreover, we analyze the GPS Drivers–Places network to understand how much the graph represents the overall mobility and how mobility complexity values are distributed among drivers and places. We also illustrate what arises applying community detection to the projected graphs of the bipartite network.

6.1 Mobility dataset

As proxy of human mobility, we used a GPS dataset collected for insurance purposes by *Octo Telematics S.p.A.*.[1] containing 9.8 million car travels performed by about 160,000 vehicles active in Tuscany in May 2011. In particular, we focused our study on Pisa and Florence provinces. In the following, we analyze the GPS Drivers–Places networks and what mobility complexity analysis applied to them can reveal. In this context, for the construction of the Drivers–Places network, we studied the systematic movements by exploiting the procedures for mobility profile and mobility POIs extraction described in Sects. 3.1 and 3.2, respectively. We used the procedure presented in 4 to extract the POIs.

Figure 4 (*left*) depicts a sample of the considered trajectories. The mobility dataset is geographically too various to be used for our purposes. Indeed, a basic issue is that mobility is not the same in every geographical area: every area is characterized by its own type of mobility with certain properties depending on the surface, the topology and the number of inhabitants. To consider this fact, we geographically filtered the dataset in provinces using as borders the administrative ones and for each province we selected all the trajectories passing through it. In this paper, we present the results obtained for Pisa and Florence provinces which are characterized by two different kinds of mobility.

In order to obtain reasonable routines, we performed some test to retrieve the best parameters to extract reliable mobility profiles. The distance function used in the clustering step is *Route Relative Synch* described in Trasarti et al. (2011). The clustering algorithm used is the density-based algorithm Optics (Ankerst et al. 1999). We studied Optics parameters on a subset of 1000 users in Pisa province. We varied $\varepsilon_r \in [0.1, 0.3]$ with step 0.01, Fig. 5a. The bigger ε_r is, the more different trajectories are allowed to be clustered together. Intuitively, this parameter represents the percentage of dissimilarity between two trajectories in a cluster, thus 0.1 means that we admit in the same cluster trajectories having a degree of similarity at least equal to 90 % while 0.3 means having a degree of similarity at least equal to 70 %. The choice of the above range of values is due to the fact that for our goal we want routines generated by trajectories with a degree of similarity lower than 70 %, are unreasonable. Moreover, we cannot set ε_r=0 (i.e., 100 % of similarity) because it is a too much strong requirement to find groups of similar trajectories that probably will lead to a no routine.

The parameter *min size*, i.e., the minimum number of trajectories that must be in a cluster considered valid, was varied in [4, 12], Fig. 5b. The aspects we considered to tune the values are: (1) the dataset coverage, (2) the profile distribution per user, and (3) the profile stability. From these distribution we use fixed a value for parameters in order to minimize the variance of observed indicators. Anyway, in each plot after the middle values, the curves change more rapidly than before them. We choose ε_r equal to 0.2 since it expresses 80 % of similarity between two trajectories and, a reliable value for *min size* is 8 since a routine is a movement repeated a sufficient number of times during a month. Figure 4 (*right*) depicts a sample of profiles extracted in Pisa modeling the users' systematic movements. Figure 5c shows the number of routines per users in Pisa province where each user has one or two routines on average, which, should correspond to the commute to and from work. Indeed, we can see that the average number of routines per profile is 2, which is probably due to the home-work-home pattern. Figure 5d shows the temporal distribution of the trajectories and routines. Here, we observe how the profile set has a working-like trend, highlighting the three peeks during early morning, lunchtime, and late afternoon.

[1] http://www.octotelematics.com/it.

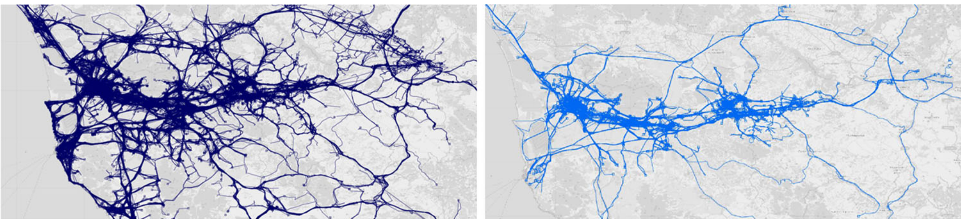
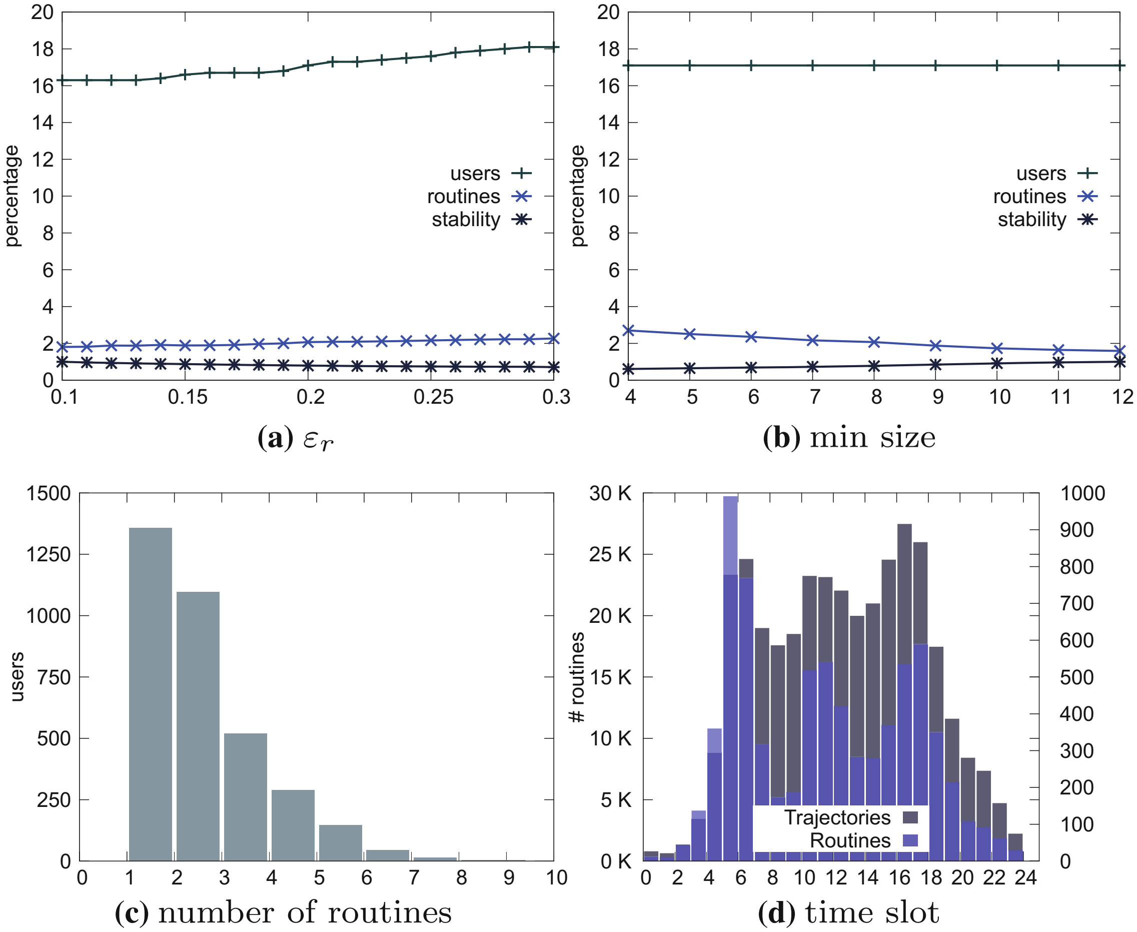


Fig. 4 (*Left*) A sample of the considered trajectories in Pisa province. (*Right*) Mobility profiles extracted in Pisa province



(a) ε_r

(b) min size

(c) number of routines

(d) time slot

Fig. 5 Parameters evaluation for mobility profile extraction. **a**, **b**, we observe the variation in percentage of the number of users, number of routines and number of mobility profiles remaining stable by varying ε_r and min size respectively. **c**, **d**, We show the distribution of the number of users per size of the mobility profile, i.e., number of routines, and the number of routines per time slot using $\varepsilon_r = 0.2$ and $minsize = 8$

6.2 Mobility POIs analysis

Now, we analyze the process of POIs extraction in term of parameters setting and results. The POIs are used as places of the Drivers–Places networks. In the extraction of POIs, we need to consider two issues: (a) a great number of POIs must be visited by at least two users otherwise they would not be a meaningful individual information in a global scenario, (b) the POI shape cannot degenerate, i.e., they cannot be too big, nor too long, nor tubular. Only two parameters must be set in the POIs building process: ε and ε'. However, we studied only ε since ε' depends on ε.

We tested the POIs construction using the routines of 1000 profiled users in Pisa province with $\varepsilon \in [20, 100]$. In this case, ε in Optics represents the maximum distance (in meters) between two individual POIs to consider them close. We recall that every place is important for someone because it is generated by a routine. We observed the number of POIs extracted and the average number of users in a POI [Fig. 6 (*left*)], the maximum area and diameter of built POIs [Fig. 6 (*right*)]. Observing the plots a reasonable value for ε appears to be 50 m. Consequently, we set $\varepsilon' = 45$ to have a remarkable buffer even for single POIs. In fact, this combination of parameters leads to a consistent number of POIs which are visited on average by at least two users. For each province, we obtain a POI distribution per profiled user telling us that the biggest subset of profiled users stop from 1 to 5 POIs. The average number of profiled users per POI ranges from 2 to 4 meaning that a place is on average always visited by at least two users. This is due to the fact that there are many places (probably users homes) which are visited only by a single user, while other social POIs like hospitals and shopping centers visited by many users. Due to the home-work-home pattern, the majority of the users visits at least two places. Moreover, both for Pisa and Florence, we note that the number of POIs is correlated neither with the number of routines nor with the surface, while it is quite correlated with the number of inhabitants and users.

6.3 GPS drivers–places network analysis

In this section, we analyze the GPS Drivers–Places network highlighting the topological characteristics of Pisa and Florence bipartite networks. According to the type of dataset used, we observe two different types of models.

Every network is made by few components, but in any case, the giant component is composed of the majority of nodes. Moreover, the Drivers–POIs networks are quite sparse considering the large number of nodes both for drivers and POIs and the fact that there are some POIs which are related with the life of few individuals and thus, they are not visited by many drivers.

We observed that the lift coefficient does not affect significantly the number of edges deleted. We have a reduction of 0.07 % edges for Pisa and 0.16 % for Florence, which means that the links generated by extracting the networks from systematic mobility data are already considerably meaningful. At any rate, using the lift coefficient, we ensure to remove irrelevant edges. Statistics in Table 1 show that the projected networks have a low level of density.

Log–log degree distributions for Pisa and Florence networks showed in Fig. 7 highlight that in both cities there are few drivers and POIs with a high degree: the value decreases following a long tailed power low distribution. This means there are few places visited by many people and many places visited by few drivers (probably one or two). The driver degree distribution is more uniform, and especially in Florence there are many drivers with a similar degree that is quite high. The average degree for drivers goes from 10 to 20, while the average degree for POIs goes from 15 to 35. It means that, on average, each entity is linked with a considerable number of other entities. This highlights the good relationship between drivers and POIs: the mobility of each driver is well represented because a valuable number of POIs are taken into account.

6.4 Mobility complexity analysis

We applied MOR on the Drivers–Places networks of Pisa and Florence with a threshold tolerance of $1.0e^{-8}$ to stop the method. Figure 8 shows the semilog plots of the mobility complexity distribution for drivers and places for the two GPS datasets, the number of visits made and received, and the number of travels and stops (i.e., the

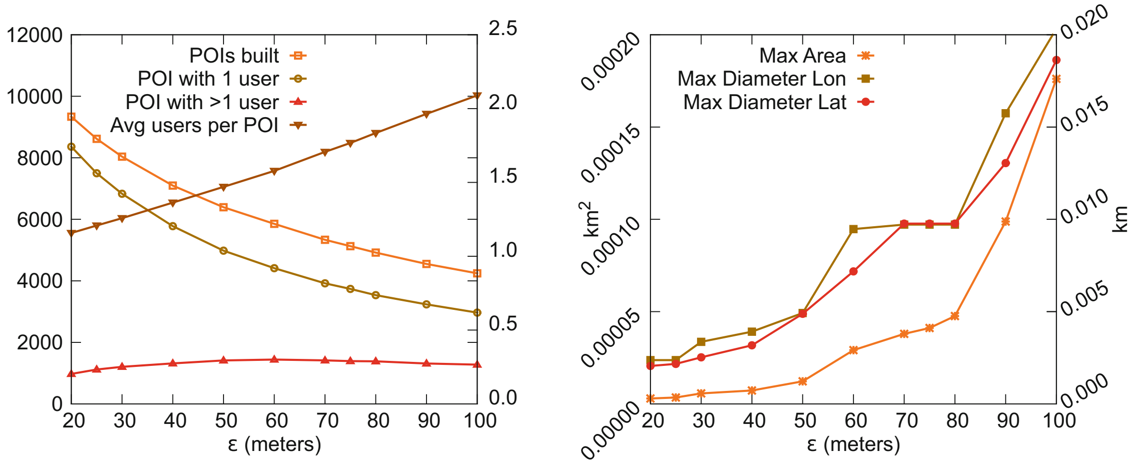


Fig. 6 Parameters evaluation for collective POIs construction. In the (*left*) plot we observe how ε parameter for POIs affects the number of POIs built, the number of POIs visited by only one driver and those visited by more than one driver ($y1$ axes), and the average numbers of drivers per POI ($y2$ axes). In the (*right*) plot we observe how ε affects the shape of the POIs (area on $y1$ axes and diameter on $y2$ axes)

Table 1 GPS Drivers–Places network statistics for Pisa and Florence

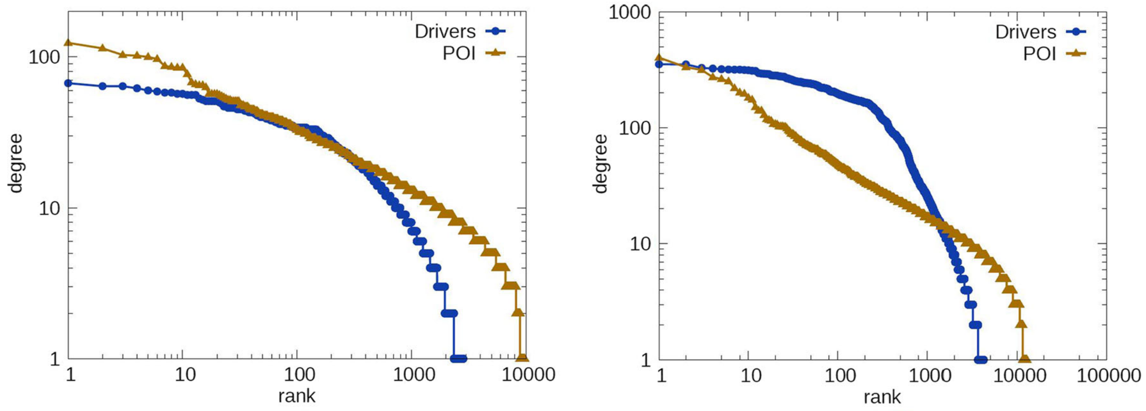
Province	|*D*|	|*P*|	|*E*|	Lift impact
Pisa	13,642	9760	148,027	0.07 %
Florence	12,848	27,765	415,447	0.16 %

D set of drivers, *P* set of POIs, *E* set of edges, *l* lift impact in the reduction of the number of edges

nodes degree). All the values are normalized between zero and one. In both provinces, the mobility complexity distributions are obviously long tailed. Thus, there are few complex drivers and many not very complex drivers. On the other hand, there are few complex places, and many not complex places. Some differences arise between the two datasets. In Pisa, there is more heterogeneity among the drivers with respect to the mobility complexity than in Florence, where most of the drivers have a similar mobility complexity. The same happens for the other curves. Regarding the POIs, the mobility complexity distribution is similar between Pisa and Florence, i.e., a similar number of POIs per users is visited, while the other curves have longer tails in Pisa than in Florence.

The couple of scores (*d*, *l*), i.e., the driver mobility complexity score and visited locations score, and the pair (*p*, *s*), i.e., the POI mobility complexity score and the stopped drivers score are obviously correlated. For example for Pisa, we have a Pearson (Galton 1886) coefficient $pearson(d, l) = 0.83$ and $pearson(p, s) = 0.79$, with *p* value smaller than 0.00001. This phenomenon is not surprising. Indeed, the degree is always correlated to rank analysis measures like PageRank and HITS. However, similarly to what happens for PageRank and HITS, our recursive definition of mobility complexity through MOR captures more than the simple diversity of POIs and visitors. Indeed, we do not consider only the diversity of places visited by a user to define it complex but also the complexity of his places. Similarly, for complex place, we do not take into account only the fact that it is a popular place (i.e., visited by a lot of users) but also the complexity of visitors. Figure 9 confirms our intuition about this fact. It reports the density scatter plot between the mobility complexity and the number of visited places (or visitors): the more the color of an area is red, the higher the density. We can notice how, according to the long tails, the denser areas are close to the origin. In the second column, we report a zoom of these areas. We can observe that the phenomenon is repeated in this smaller area. The outcome of this figure is that for a consistent group of nodes, both drivers and places, the two measures are correlated: their points are close to the black straight line representing the ideal situation in which the correlation is 1; on the other hand, for another consistent group of nodes lying far from this line the correlation is not so high. Take for example, the points *A* and *B* of every plot. *A* is a node (either driver or place) with a mobility complexity higher with respect to the number of places visited/number of visitors. In other words, the few places visited are very complex. On the other hand, considering *B*, the complexity is very low for the relative high degree. Thus, the many visited locations (or the many visitors) are not complex.

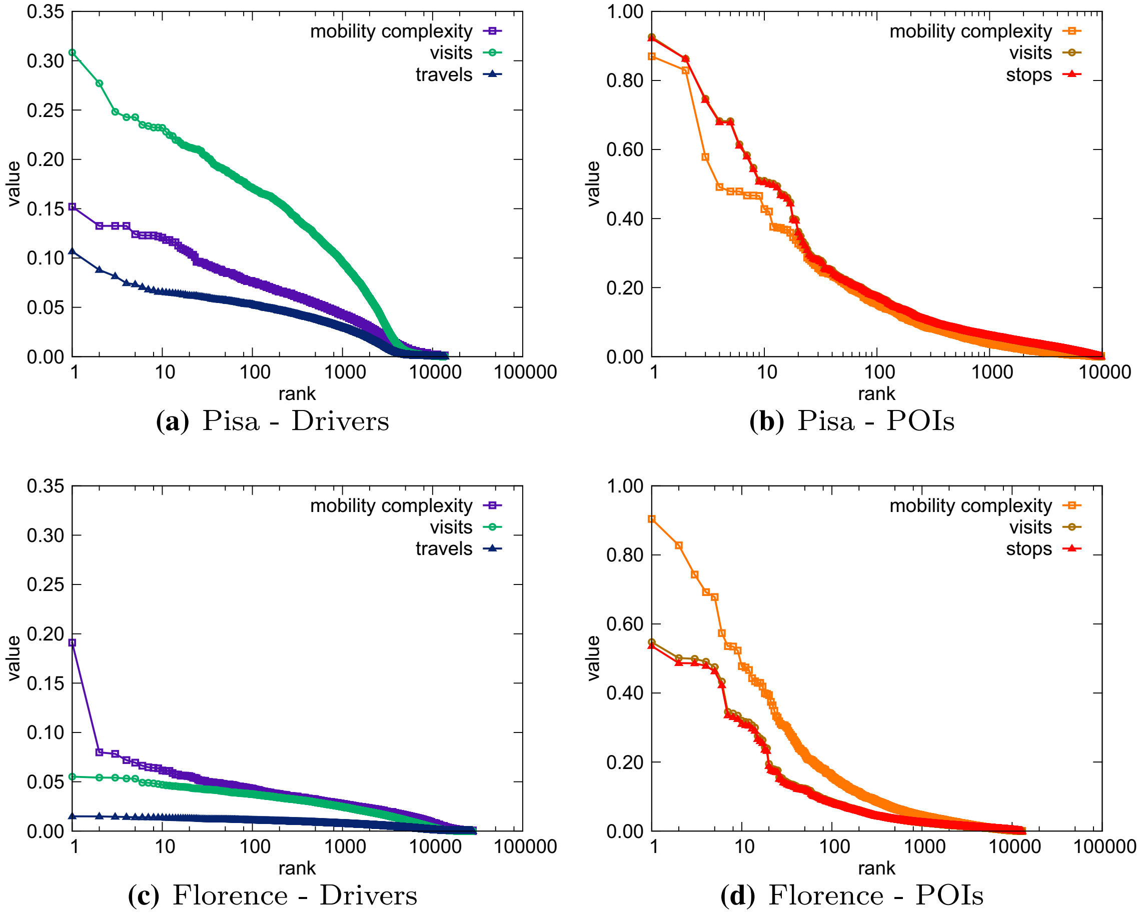
Plots in Fig. 10a, c depict the driver mobility complexity versus the average place mobility complexity. They highlight: (1) there are few drivers with a high mobility complexity visiting a lot of POIs with an average low mobility complexity; (2) there are few drivers visiting few POIs with an average high complexity, they probably visit only their own places and perhaps a complex POI such as a shopping center; and (3) we have many not complex drivers visiting POIs that are not very complex on average, i.e., they visit few complex POIs. Plots in Fig. 10b, d show the place mobility complexity versus the average driver



(a) GPS - Pisa

(b) GPS - Florence

Fig. 7 Distribution of the degree of the drivers (*blue circles*) and places (*yellow triangles*) in log–log scale for Pisa (**a**) and Florence (**b**) (color figure online)



(a) Pisa - Drivers

(b) Pisa - POIs

(c) Florence - Drivers

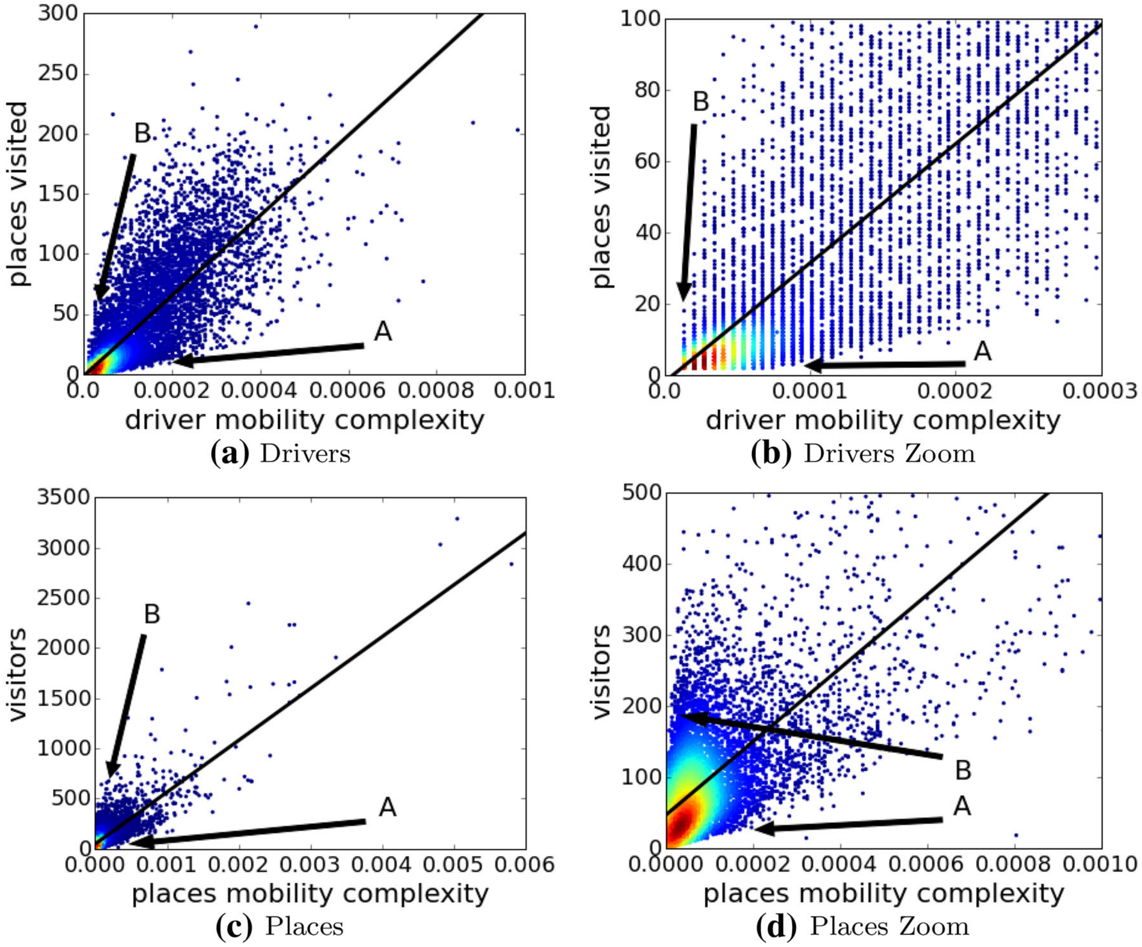
(d) Florence - POIs

Fig. 8 Distribution of the mobility complexity (*squares*), number of visits (*circles*), and number of travels/stops (*triangles*) in semilog scale. The driver mobility complexity is in **a**, **c**, while the POIs mobility complexity is in **b**, **d**

mobility complexity. In this case, it appears that few POIs are very complex and they are visited nearly by all drivers, thus they are visited both by complex and not complex drivers. Then, we have very few POIs not complex but visited by some complex drivers. Moreover, there are many places not so much complex because they are visited on average by not complex drivers.

In general, we highlight that a large amount of drivers have a low mobility complexity and visit not complex places. Inspired by Pappalardo et al. (2015), we could categorize them as *common drivers* because they do not travel very much, going systematically in many complex POIs and in few not complex POIs. Only few drivers have a low complexity but visit complex POIs: this means that they are more systematic than *common drivers* going only in their places, irrelevant for others, and in a complex POI such as a shopping center. We can claim this knowing the formula used in MOR. Thus, they could be called *systematic drivers*. Finally, few users have a high complexity visiting not complex POIs. The only way to achieve this is that they visit a lot of POIs not complex on average. This last category is a sort of *explorers* because they visit many places that are not very common. A similar reasoning can be done about places. In this case, it is clear that a large part of POIs are concentrated in the bottom left corner of Fig. 10b, d, meaning that they are private houses or not common workplaces. Only few places are very complex and a POI to be complex must be visited by many complex drivers. In fact, the most complex POI has a low average driver mobility complexity, and this is a signal that it is visited by drivers of any type. This reasoning illustrates how ranking measures might be helpful in classifying human mobility.

Is it interesting to observe that the most complex POIs are frequented by all kinds of drivers, both complex and not complex. Figure 11 shows the ten most complex POIs in Pisa and Florence. They are mainly big shopping centers, hospitals and car parks close to locations visited very often by many people. We underline that, in both provinces, there are some complex POIs out of the main town but always corresponding to car parks close to big malls.



(a) Drivers **(b)** Drivers Zoom

(c) Places **(d)** Places Zoom

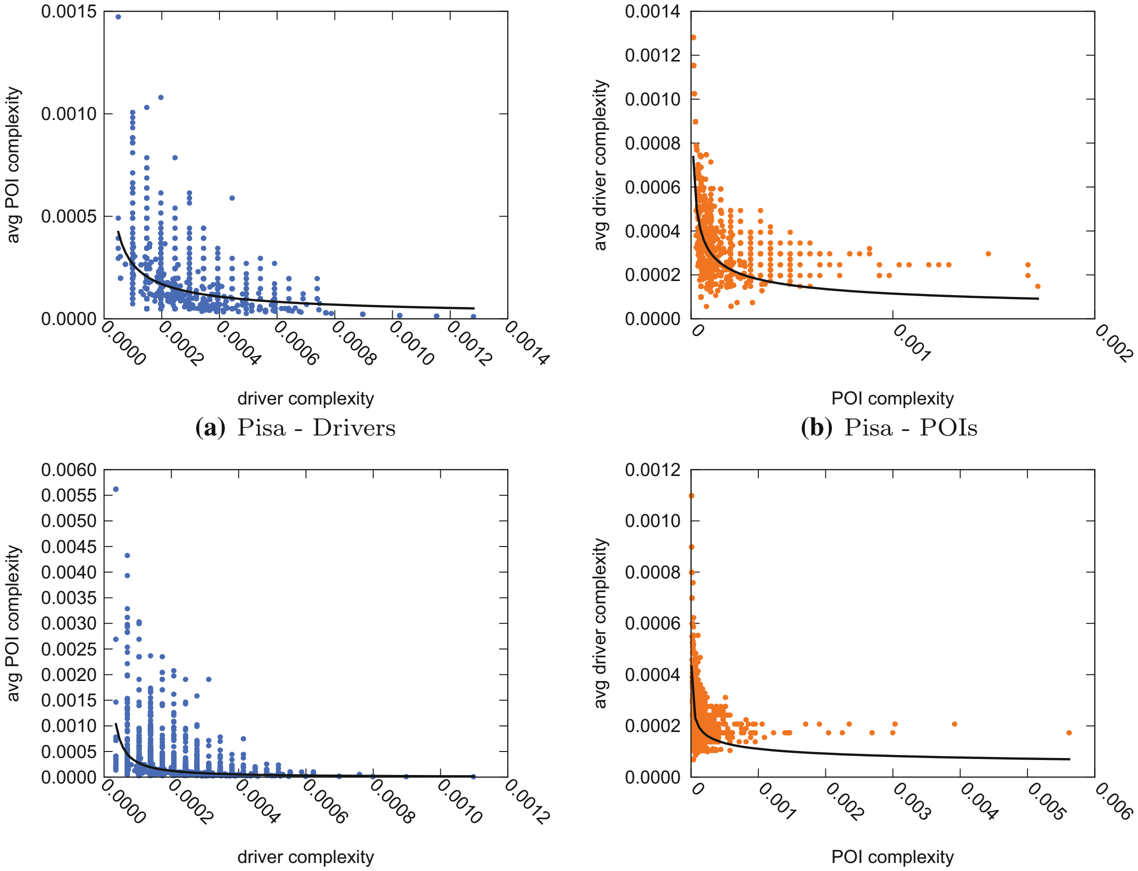
Fig. 9 Density scatter plots of mobility complexity against number visits for Pisa. The *black straight line* is the fitting function representing the equivalence between mobility complexity and node degree

6.5 Mobility communities

In our analysis, we are also interested in observing if it is possible to characterize some groups of similar places or drivers in terms of mobility complexity. In particular, we would like to understand if groups of places or drivers show the homophily phenomenon (McPherson et al. 2001) and, if this is the case, which is the relationship between the mobility complexity and the degree of homophily. To this end, we extract two projections from our Drivers–Places network. The first projection, *Drivers–Drivers*, connects two drivers i and i' to each other if they have stopped in at least a common POI. The second projection, *POIs–POIs*, links two POIs j and j' if they have been visited at least by a common driver.

It is worth to notice that, when doing projection, very high degree nodes in the bipartite network of the type that is not projected to, can cause large cliques in the one-mode network, i.e., the *Drivers–Drivers* and *POIs–POIs* networks. This can influence metrics and distributions of these networks. In Table 2, we report some features describing the projected networks. We observe how, in both networks of Pisa and Florence, the average degree μ and standard deviation σ are quite high. This is due to the effect described above. However, the skewness of the degree distribution ς is always positive, the medians v are much smaller than the means, and the density δ are very low. These indicators tell us that, even if the effect described above is present, it does not affect the structure of the network. In other words, even if there are places visited by the majority of the drivers, thus linking many drivers together in the projected network, the overall distribution of the degree remains long tailed: there are few nodes linked with many nodes and many nodes linked with few nodes.

We weighted the edges on the projections to evaluate the similarity between neighbors in order to estimate the level of homophily within a community. We use the Jaccard coefficient (Pang-Ning et al. 2006) to weight the similarity between each couple of linked nodes for each



(a) Pisa - Drivers

(b) Pisa - POIs

(c) Florence - Drivers

(d) Florence - POIs

Fig. 10 Scatter plots of mobility complexity versus the average score of the linked nodes

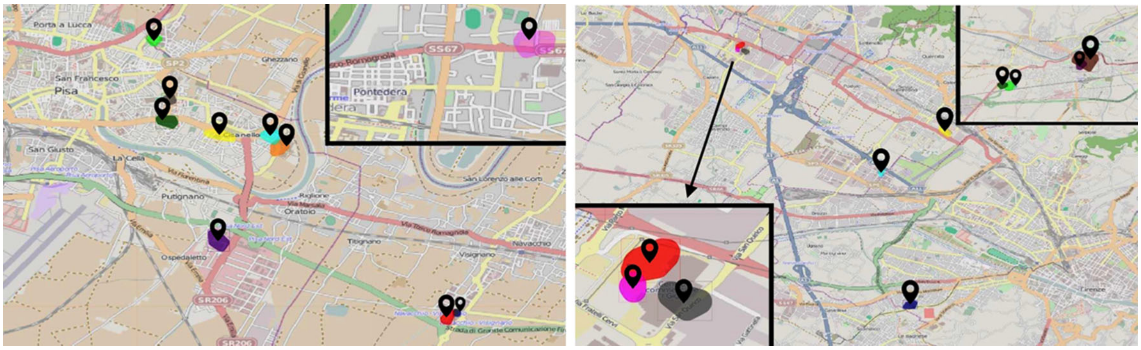


Fig. 11 Top ten POIs with respect to mobility complexity for Pisa (*left*) and Florence (*right*). They are large malls and shopping center or parking areas close to them

community in the two partitions. More formally, given two drivers i and i' and two places j and j' the corresponding weights are:

$$w_{ii'} = \frac{|N(i) \cap N(i')|}{|N(i) \cup N(i')|} \quad w_{jj'} = \frac{|N(j) \cap N(j')|}{|N(j) \cup N(j')|}$$

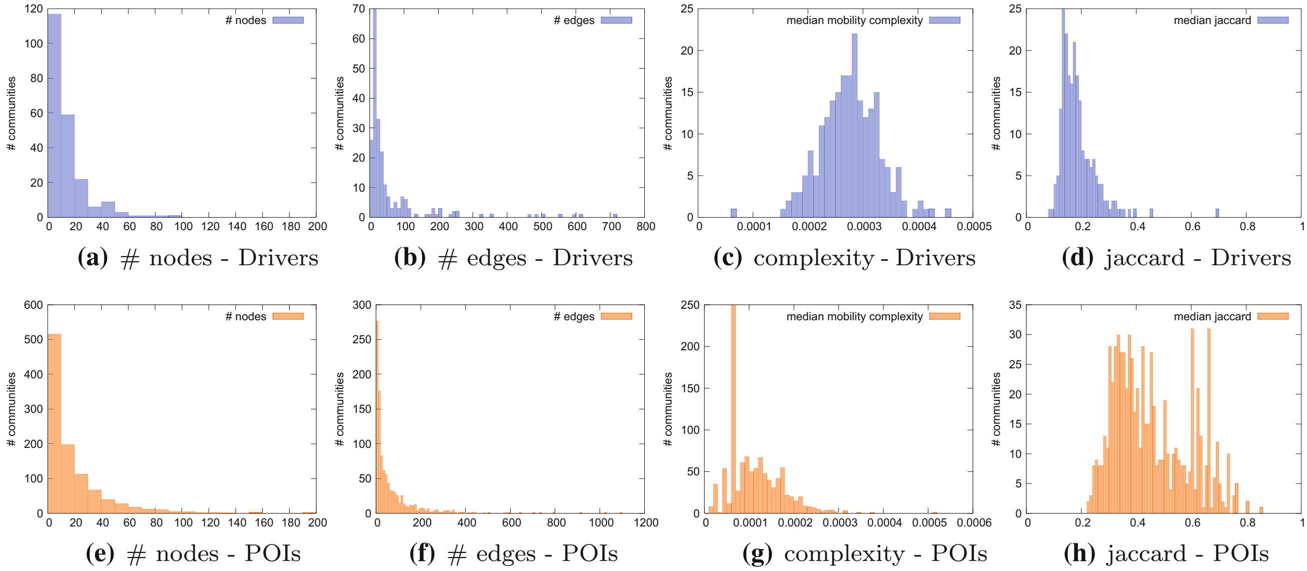
where we denote with $N(\cdot)$ the function that given a node v returns the set of neighbors of v, More formally, given a network $G = (V, E)$ the set of neighbors of a node $v \in V$ is defined as $N(v) = \{u \in V | \exists (v, u) \in E\}$.

In order to extract groups of similar drivers and similar places, we applied community detection on the Drivers–Drivers and Places–Places projected networks obtained from the Drivers–Places networks. Among several community detection algorithms such as Demon, Infohiermap and Louvain (Coscia et al. 2012; Rosvall and Bergstrom

Table 2 GPS Drivers–Drivers and POIs–POIs network statistics for Pisa and Florence

Province-type	$\|N\|$	$\|E\|$	μ	σ	ν	ς	δ	$\|C\|$
Pisa—Drivers–Drivers	13,642	3,144,699	461.13	613.79	175.00	1.76	0.0338	220
Pisa—POIs–POIs	9,760	1,353,382	277.33	384.54	137.50	3.36	0.0284	1,028
Florence—Drivers–Drivers	12,848	2,961,669	672.24	699.69	382.50	1.97	0.0359	256
Florence—POIs–POIs	27,765	3,597,603	560.02	653.26	338.00	2.65	0.0093	1,205

N set of nodes (drivers or places), *E* set of edges, μ average degree, σ degree standard deviation, ν degree median, δ network density, ς degree skewness, *C* set of communities extracted



(a) # nodes - Drivers **(b)** # edges - Drivers **(c)** complexity - Drivers **(d)** jaccard - Drivers

(e) # nodes - POIs **(f)** # edges - POIs **(g)** complexity - POIs **(h)** jaccard - POIs

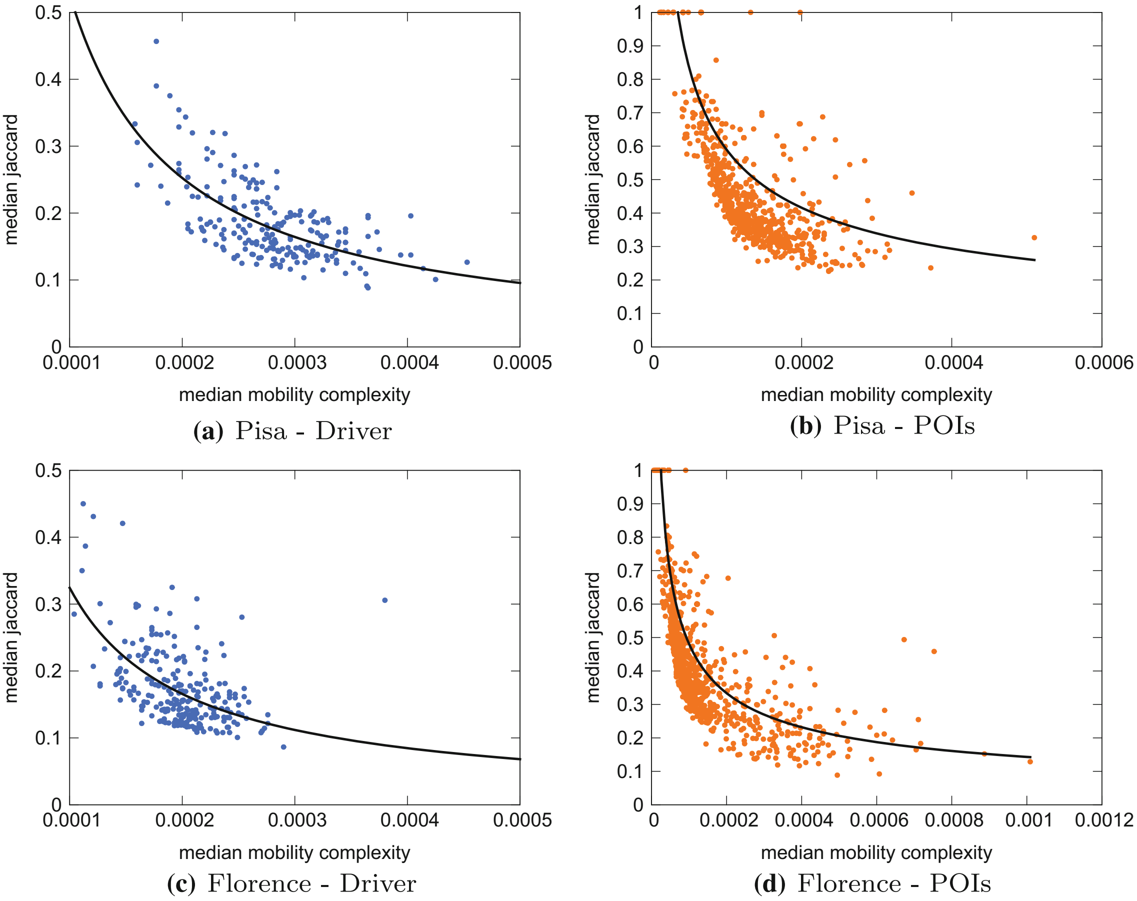
Fig. 12 Statistics about the communities extracted on the Drivers–Drivers network and POIs–POIs network in Pisa dataset. From *left* to *right*, we find the number of communities per number of nodes, number of edges, median mobility complexity and median Jaccard coefficient. The distributions for the Drivers–Drivers communities are in the top row, while the distributions for the POIs–POIs communities are in the bottom row

2011; Blondel et al. 2008), we adopted Demon on the two projected networks since the communities returned have a treatable size and there is not a dominant component as for the other methods. The communities returned are interesting because a community of drivers is composed of people who visit the same POIs while, a community of POIs is composed of places visited by the same group of drivers. By studying the size, number of nodes and number of edges community distribution, we notice that there are few small size communities, many medium size communities and a few large communities. Figure 12 shows the distributions for number of nodes, edges, median mobility complexity, and median Jaccard coefficient both for drivers and POIs in Pisa dataset for the communities extracted.

In the following, we denote homophilus communities (i.e., with a high median Jaccard coefficient) with a low median mobility complexity as *homogeneous communities*, while heterophilous communities (i.e., with a low median Jaccard coefficient) with a high median mobility complexity as *heterogeneous communities*. In other words, the first type of communities is those composed of very similar drivers or very similar places. On the contrary, the second type of communities is those composed of drivers or places with a low degree of similarity. Why are we interested in finding the relationship between mobility complexity of users (or places) and similarity of users (or places) deriving by the network component? Once getting the characterization of our communities we can use one of the two involved components (similarity or complexity) for inferring the other one. For example, based on our finding by knowing simply that the mobility complexity of nodes (drivers or POIs) in a community is high then, we can directly infer that similarity of those nodes is low without computing the similarity.

6.5.1 Drivers communities

A community of drivers is composed of people visiting similar places (POIs). Figure 13a, c shows the scatter plot of the median mobility complexity against median Jaccard



(a) Pisa - Driver

(b) Pisa - POIs

(c) Florence - Driver

(d) Florence - POIs

Fig. 13 Scatter plots of Driver and POIs mobility complexity versus Jaccard per community. In both cases and both dataset, we can observe an anti-correlation: high jaccard, i.e., similarity, means low mobility complexity, while high mobility complexity means low similarity

coefficient for drivers communities. We observe that the more complex is a community, the less similar is its drivers, and the less complex is a community the more similar are its drivers. Drivers visiting not complex places cannot have a high value of mobility complexity because, according to what exposed previously, they are quite systematic and do not visit complex POIs. On the other hand, if a community is made of complex drivers, they can be similar each other but only until a certain level because if all of them visited the same complex POIs, then, their mobility complex score would have been lower by definition. This means that their community would have been less complex. In other words, we found that homophilic communities tend to have a low mobility complexity. This information could be used to predict a new location visited by a certain driver. In fact, if a group of drivers frequent the same places, with a high probability, they have a similar lifestyle and/or similar interests. Therefore, it is plausible that similar drivers will visit similar places in the near future. This supposition becomes even more probable for nodes in *homogeneous communities*.

6.5.2 *Places communities*

A community of POIs is made of locations visited by similar drivers. Similarly to driver communities, the same results are exposed in Fig. 13b, d about POIs. The behavior of mobility complexity and Jaccard coefficients still holds for *homogeneous communities* and *heterogeneous communities*. However, this time most of the communities are concentrated in an area between low median mobility complexity and middle median mobility complexity, that is, there are more *homogeneous communities*. This indicates that these groups of places are visited from a set of drivers quite narrow and not very variable. So, we can observe that the homophily phenomenon is more evident in the Place–Place network. The POIs community information in conjunction with mobility complexity could be used to classify a place according to mobility criteria. In fact, if a group of places is visited by drivers with certain characteristics, then, it means that these places are suitable for this kind of people.

6.5.3 Communities summary

Summarizing, the main result emerging from the study of the communities on the projections is that: the more complex a community is the weaker are the ties among their nodes, i.e., the nodes do not tend to be homophilic; on the other hand, the less complex a community is, the stronger are the links and consequently the similarity among their nodes. These communities could be called *heterogeneous* when the median mobility complexity is high and *homogeneous* when the median mobility complexity is low. Therefore, the mobility society could be roughly split in subsets with a different mobility behavior: a set of (1) homophilic and not complex groups of drivers and POIs and (2) a set of groups of drivers and POIs which are not very similar and having a low level of complexity.

7 Case study on GSM data

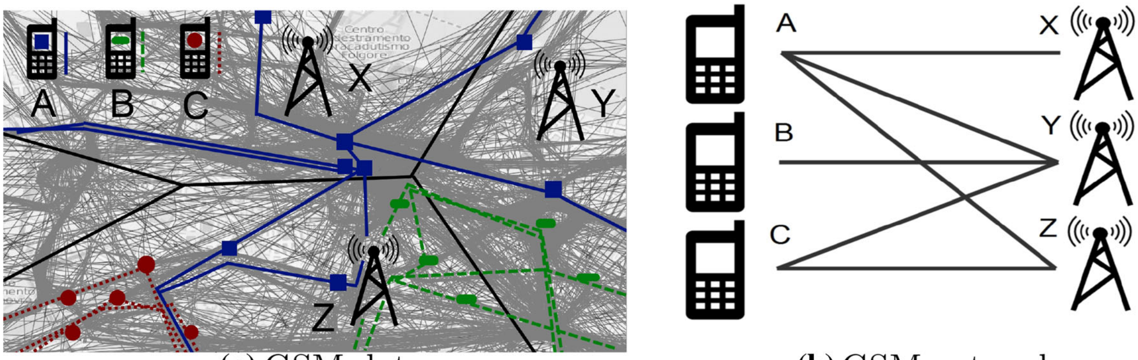
A Drivers–Places network can be constructed on the basis of mobile phone network traces that are commonly and massively available from telecom operators. In this setting, we do not need to extract POIs from the mobile phone traces, but we use directly the raw data of each user phone call composed of $\langle caller_{id}, cell_1, cell_2 \rangle$ (see Fig. 14). In particular, the phone cells are POIs and we add an edge for each cell in which the user appears during a call. Starting of this network, we can perform the same kind of mobility complexity analysis like that one, presented in Sect. 6.4, for GPS traces and compare the results. The GSM dataset used for our case study is composed of call data collected by a big telecom operator during October 2013 in Tuscany, in particular in the provinces of Pisa, Lucca, Livorno and Florence. It contains about 67.3 millions of calls made by 979,000 users . We focused our study on the data of Pisa and Florence province in order to make the analytical results comparable with those obtained in the previous case study on GPS data.

In the following, we analyze the GSM Drivers–Places networks of Pisa and Florence in order to understand what the mobility complexity analysis applied to it can reveal. Starting from GSM data, we obtained bigger networks due to the high numbers of drivers (see Table 3). Indeed, in this case, we have both occasional and systematic drivers who move from a cell to another one. On the contrary, we have only a limited number of cells (places). In this kind of network, the lift coefficient has a considerable impact both for Pisa and Florence (20.87 and 14.20 % respectively). Table 3 reports the dimensions of the bipartite networks of the two provinces. In both cases, we obtain networks with a low level of density. We note that the GSM Drivers–Drivers and Places–Places networks are denser than the GPS ones.

Figure 15 shows two different degree distributions for drivers and places. On the other hand, in GPS data, we obtained a bipartite network with comparable distributions of the degree for drivers and places. This happens because in the GSM Drivers–Places network every place is a *cell* and consequently has a very high degree due to the large spatial coverage (2–5 km^2). Indeed a GSM cell captures a considerably larger set of drivers in terms of visits if compared with the POIs extracted in the GPS case study (0.5–2 km^2).

We also analyzed the distribution of the mobility complexity for drivers and places of the two GSM datasets (for Pisa and Florence). Figure 16 shows the results. We can observe that, as for the GPS case study, also this time we have long tailed power low distributions. However, these curves are more uniform due to the fact that there is a considerable low number of places.

Finally, we performed the analysis of communities extracted from GSM Drivers–Places networks in order to study groups of similar drivers and places with respect to



(a) GSM data

(b) GSM network

Fig. 14 An example of Drivers–Places network extracted from GSM data. *Lines* represent sequences of calls for drivers A, B and C while the *towers* represents common cells X, Y and Z. The *gray background* is other trajectories of calls not considered in the example

mobility complexity. Unfortunately, we did not find any interesting result due to the small number of cells in this kind of networks.

8 Conclusion

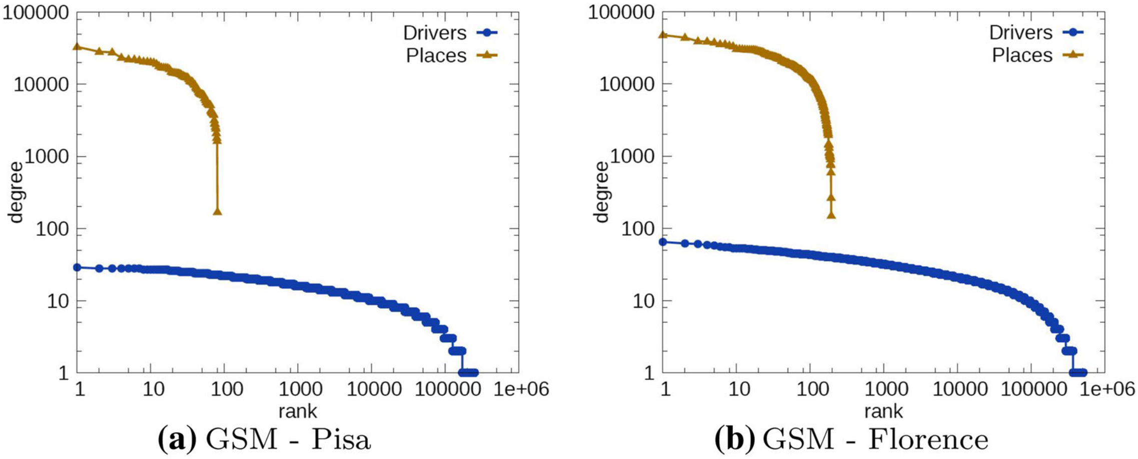
In this paper, we present a network analytics approach to study human mobility. From the observation of raw movements, we construct a high level representation of mobility by means of a bipartite network, the *Driver-Place* network. The network contains an edge between two nodes d and p when there is at least a visit of a driver d to the place p. Starting from this network, we depart from the analysis of degree distribution of nodes. We focus on the intuition that a deeper understanding of mobility phenomena should consider the mobility of a person in her whole. Thus, we propose to study the characteristics of the network with a link analysis approach, where each element of the network is related with the topological properties of its neighborhood. This approach improves the traditional studies on mobility by augmenting the quantitative estimation of indicators and patterns with a qualitative characterization of nodes. We are not solely interested on the volume of traffic attracted by a particular place (or generated by a driver), but we want to state the capability of a place to attract drivers that have visited many other places. To this aim, a driver visits many places and she influences each place she visits. A place is visited by many drivers and each driver gives a contribution according to her previously visited places.

We call such measure *mobility complexity*. The inherent estimation of this complexity if computed by means of the Method of Reflection (we prove a formal equivalence of MOR with HITS in Appendix). This methods provides a measure of relevance of the two families of nodes: complex drivers are persons that visits many complex destinations; complex places are zones visited by many complex drivers. The recursive definition of this measure allows to capture properties of mobility that a mere quantitative evaluation can not provide. In particular, if we compare the complexity of nodes, with their degree, we can notice that there are new evidences that emerge. For instance, in Sect. 6, we show that the two measures are related, but *mobility complexity* adds new levels of interpretation. For example, there are places with low visits (i.e., low degree) that have

Table 3 GSM Drivers–Cells network statistics for Pisa and Florence

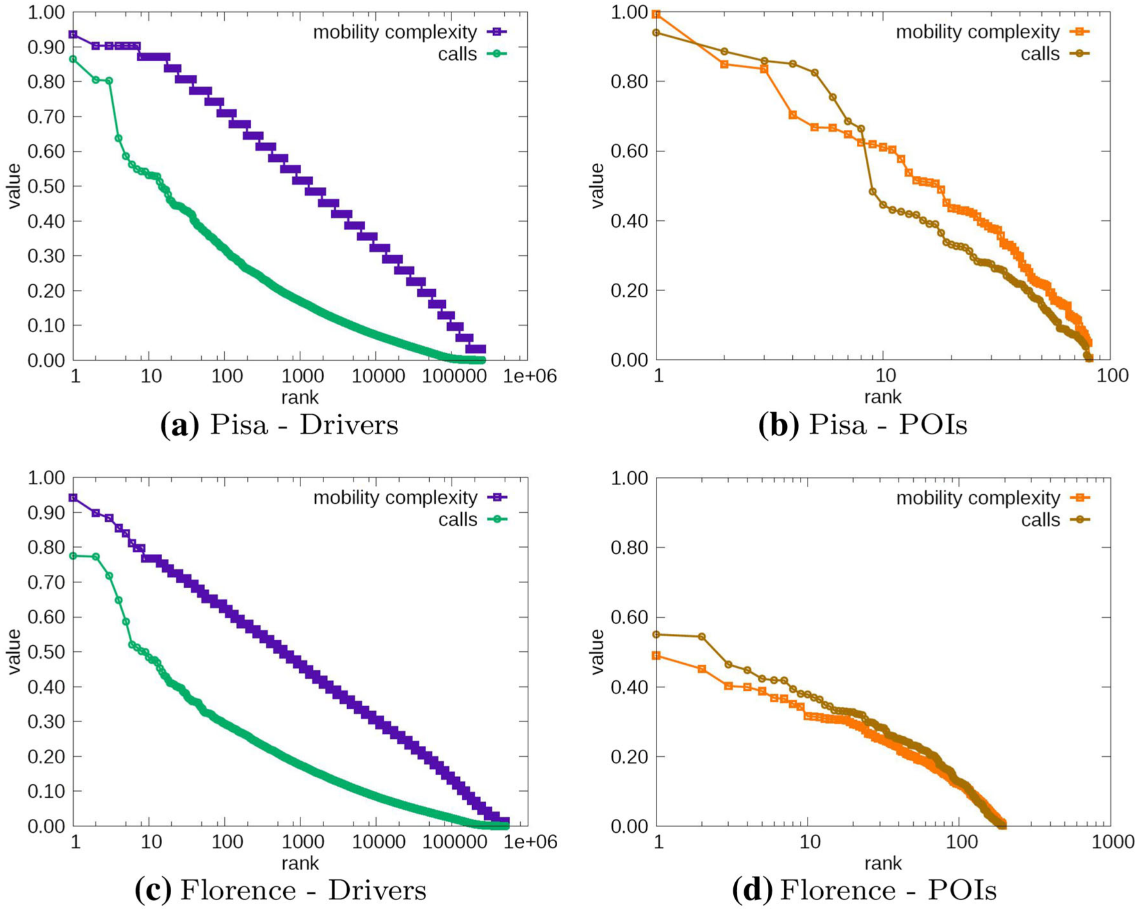
Province	Pisa	Florence
$\|D\|$	251,895	511,672
$\|P\|$	82	195
$\|E\|$	908,700	2,773,960
l	20.87 %	14.20 %
$\|E_D\|$	89,478,486	181,756,827
δ_D	0.0028	0.0014
$\|C_D\|$	74	16
$\|E_P\|$	3315	18,803
δ_P	0.9982	0.9941
$\|C_P\|$	3	3

D set of drivers, P set of places, E set of edges, l lift impact in the reduction of the number of edges, $E_{D,P}$ set of edges, $\delta_{D,P}$ density and $C_{D,P}$ of the Drivers–Drivers or Cells–Cells network, respectively



(a) GSM - Pisa **(b)** GSM - Florence

Fig. 15 Distribution of the degree for the GSM networks of the drivers (*blue circles*) and places (*yellow triangles*) in log–log scale for Pisa (**a**) and Florence (**b**)



(a) Pisa - Drivers

(b) Pisa - POIs

(c) Florence - Drivers

(d) Florence - POIs

Fig. 16 Distribution of the mobility complexity (*squares*) and number of calls (*circles*) in semilog scale for the GSM network. The driver mobility complexity is in **a**, **c**, while the POIs mobility complexity is in **b**, **d**

high complexity, whereas there are places with very high degree and low complexity.

This measure for mobility opens many application scenarios. From the point of view of traffic management, the complexity of places may support a mobility manager to reorganize the connections among places by means of public transportation service. It is also relevant to have a complexity estimation in emergency situation, when, for example, it is necessary to isolate part of the road network. The driver mobility complexity may be used to provide highly customized services to individuals. For example, an insurance company may offer different prices to different profiles of user.

We envisage other future developments of the approach. As a first exploration, we want to further develop the community analysis performed on the projected networks. The experimental results give a clear indication that there are group of drivers that are similar and visit similar places. This property may be refined to compare mobility behaviors in different regions of a country. It also interesting to investigate how external behaviors are mapped on complexity property. Consider for example the problem of simulating an epidemic scenario. The added value of mobility complexity may provide more reliable simulation, given the capability of having different exploration of the geographical space: complex places may be considered as high risk zone for contagion, whereas complex drivers are, very likely, vectors that can spread the epidemy faster.

Acknowledgments This work is partially supported by the European Communitys H2020 Program under the funding scheme FETPROACT-1-2014: 641191 CIMPLEX, and by the SMARTCITIES Project No. FP7-ICT-609042, PETRA.

Appendix: Method of reflection as particular case of HIT

In this section, we show that the Method of Reflection (MOR) (Caldarelli et al. 2011) can be seen as a particular case of HITS (Kleinberg et al. 1999). In HITS, we have the *authority* score, which estimates the value of a node, and its the *hub* score, which estimates the value of its links to other nodes. Authority and hub values are defined in terms of each other in a mutual recursion manner:

$$h_i^{(n)} = \sum_{j=1}^{|P|} A_{ij} a_j^{(n-1)} \forall i \qquad a_j^{(n)} = \sum_{i=1}^{|P|} A_{ij} h_i^{(n-1)} \forall j$$

where n is the iteration index, A is the adjacency matrix $A^{n \times n}$, h the of hub scores vector, a the authority scores vector, and P is the set of nodes. Hub and authority update rules can be viewed as matrix-vector multiplication:

$$h = Aa \quad a = A^T h \tag{5}$$

In practice, HITS performs a series of iterations by computing for each step:

$$h^{(n)} = AA^T h^{(n-1)} \quad a^{(n)} = A^T A a^{(n-1)}$$

that are normalized to guarantee the convergence. A common technique used to calculate the hub and authority scores is the *power iteration method* (Lanczos 1950).

Comparing the equations of HITS (5) with the equations of MOR (2) in Sect. 5, it is easy to observe that HITS and MOR are very similar. Indeed, we have the same kind of computation applied to different matrices: HITS uses a standard adjacency matrix A, while MOR uses an adjacency matrix M weighted with the degree of the nodes.

In the following, we formally prove this similarity.

Theorem 1 *Let G be a bipartite graph and $\bar{A}$ its weighted adjacency matrix. $\bar{A}^{(|D|+|P|)\times(|D|+|P|)}$ its weighted adjacency matrix. Applying HITS to G by using $\bar{A}$ is equivalent to apply MOR to G.*

Proof Since the graph $G = (D, P, E)$ is bipartite we have that $D \cap P = \emptyset$, thus the weighted adjacency matrix $\bar{A}^{(|D|+|P|)\times(|D|+|P|)}$ has the form

$$\bar{A} = \begin{bmatrix} 0 & \bar{M} \\ \bar{M}^T & 0 \end{bmatrix}$$

where $\bar{M}^{|D|\times|P|}$ is the same adjacency matrix used in MOR and $(\bar{M}^T)^{|P|\times|D|}$ is its transposed matrix. Now we have that $\bar{A}^T = \bar{A}$ since

$$\bar{A}^T = \begin{bmatrix} 0 & (\bar{M}^T)^T \\ (\bar{M})^T & 0 \end{bmatrix} = \begin{bmatrix} 0 & \bar{M} \\ \bar{M}^T & 0 \end{bmatrix} = \bar{A}$$

Applying HITS to G means $h^{(n)} = \bar{A}\bar{A}^T h^{(n-1)}$ and $a^{(n)} = \bar{A}^T \bar{A} a^{(n-1)}$ where $\bar{A}\bar{A}^T = \bar{A}^T\bar{A} = \hat{A}$, that is

$$\hat{A} = \begin{bmatrix} 0 & \bar{M} \\ \bar{M}^T & 0 \end{bmatrix} \begin{bmatrix} 0 & \bar{M} \\ \bar{M}^T & 0 \end{bmatrix} = \begin{bmatrix} \bar{M}\bar{M}^T & 0 \\ 0 & \bar{M}^T\bar{M} \end{bmatrix} = \begin{bmatrix} \mathcal{D} & 0 \\ 0 & \mathcal{P} \end{bmatrix}$$

By applying the power iteration method to $\hat{A}$ we obtain $min(|D|, |P|) = s$ eigenvalues with the following set $\lambda_1, \lambda_1, \lambda_2, \lambda_2 \ldots, \lambda_{s/2}, \lambda_{s/2}$ and $max(|D|, |P|) - min(|D|, |P|)$ eigenvalues equal to zero. Assuming that $\lambda_i > \lambda_j$ for $i < j$ for the convergence, there are $\frac{s}{2}$ eigenvalues each one associated with an *eigenpair*. Given the eigenpair a_i^d and h_i^p associated with the eigenvalue λ_i, it must hold that $a_i^d \neq h_i^p \neq 0$. The only possibility is that a_i^d and h_i^p have the form

$$a_i^d = \begin{bmatrix} d_i^* \\ 0 \end{bmatrix} \; h_i^p = \begin{bmatrix} 0 \\ p_i^* \end{bmatrix}$$

Therefore, the results are a_1^d and h_1^p because

$$\lambda_1 = \rho(\bar{A}) = \rho(\mathcal{D}) = \rho(\mathcal{P})$$
$$\bar{A}a_1^d = \lambda_1 a_1^d \quad \bar{A}h_1^p = \lambda_1 h_1^p$$

and removing the useless zeroes we get

$$\mathcal{D}d_1 = \lambda_1 d_1 \quad \mathcal{P}p_1 = \lambda_1 p_1$$

that is the result of MOR.

This statement proves that MOR is a particular case of HITS with a weighted adjacency matrix applied to a bipartite graph. It is worth to underline that the above theorem only proves the equivalence of the two algorithms under some conditions; and it clearly does not suggest that HITS could replace MOR in the calculus of ranking score on bipartite networks. Indeed, it would be useless since many multiplications per zero would be performed.

References

Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. ACM SIGMOD Rec 22:207–216

Ankerst M, Breunig MM, Kriegel HP, Sander J (1999) Optics: ordering points to identify the clustering structure. ACM Sigmod Rec 28:49–60

Ashbrook D, Starner T (2003) Using GPS to learn significant locations and predict movement across multiple users. Pers Ubiquit Comput 7(5):275–286

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):10,008

Brilhante IR, Berlingerio M, Trasarti R, Renso C, de Macedo JAF, Casanova MA (2012) Cometogether: discovering communities of places in mobility data. In: IEEE 13th international conference on mobile data management (MDM), pp 268–273

Caldarelli G, Cristelli M, Gabrielli A, Pietronero L, Scala A, Tacchella A (2011) Ranking and clustering countries and their products: a network analysis. arXiv preprint arXiv:1108.2590

Caldarelli G, Cristelli M, Gabrielli A, Pietronero L, Scala A, Tacchella A (2012) A network analysis of countries' export flows: firm grounds for the building blocks of the economy. PLoS ONE 7(10):e47278. doi:10.1371/journal.pone.0047278

Cao L, Luo J, Gallagher A, Jin X, Han J, Huang TS (2010) A worldwide tourism recommendation system based on geotagged-web photos. In: IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp 2274–2277

Coscia M, Rossetti G, Giannotti F, Pedreschi D (2012) Demon: a local-first discovery method for overlapping communities. In: ACM 18th SIGKDD international conference on Knowledge discovery and data mining, pp 615–623

Eubank S, Guclu H, Kumar VA, Marathe MV, Srinivasan A, Toroczkai Z, Wang N (2004) Modelling disease outbreaks in realistic urban social networks. Nature 429(6988):180–184

Furletti B, Cintia P, Renso C, Spinsanti L (2013) Inferring human activities from gps tracks. In: ACM 2nd SIGKDD International Workshop on Urban Computing, p 5

Galton F (1886) Regression towards mediocrity in hereditary stature. J Anthropol Inst Great Br Irel 15:246–263

Giannotti F, Nanni M, Pedreschi D, Pinelli F, Renso C, Rinzivillo S, Trasarti R (2011) Unveiling the complexity of human mobility by querying and mining massive trajectory data. Int J Very Large Data Bases 20(5):695–719

Guidotti R, Monreale A, Rinzivillo S, Pedreschi D, Giannotti F (2014) Retrieving points of interest from human systematic movements. In: Canal C, Idani A (eds) Software engineering and formal methods. Lecture Notes in Computer Science, vol 8938. Springer, pp 294–308

Guidotti R, Trasarti R, Nanni M (2015) TOSCA: two-steps clustering algorithm for personal locations detection. In: ACM 23nd SIGSPATIAL international conference on advances in geographic information systems

Hidalgo CA, Hausmann R (2009) The building blocks of economic complexity. Proc Natl Acad Sci 106(26):10570–10575

Hidalgo CA, Hausmann R (2010) Inferring macroeconomic complexity from country-product network data. In: 2010 AAAI Spring Symposium Series

Hossmann T, Spyropoulos T, Legendre F (2011) A complex network analysis of human mobility. In: IEEE conference on Computer communications workshops (INFOCOM WKSHPS), pp 876–881

Jiang S, Ferreira J, González MC (2012) Clustering daily patterns of human activities in the city. Data Min Knowl Disc 25(3):478–510

Kaluza P, Kölzsch A, Gastner MT, Blasius B (2010) The complex network of global cargo ship movements. J R Soc Interface 7(48):1093–1103

Kleinberg JM , Kumar R, Raghavan P, Rajagopalan S, Tomkins AS (1999) The web as a graph: measurements, models, and methods. In: Asano T, Imai H, Lee DT, Nakano S-i, Tokuyama T (eds) Computing and combinatorics. Lecture Notes in Computer Science, vol 1627. Springer Berlin, Heidelberg, pp 1–17

Lafferty J, McCallum A, Pereira FC (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: 8th International conference on machine learning, ICML 2001, Morgan Kaufmann Publishers Inc. pp 282–289

Lanczos C (1950) An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. United States Governm, Press Office, Los Angeles

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27:415–444. doi:10.1146/annurev.soc.27.1.415

Monreale A, Pinelli F, Trasarti R, Giannotti F (2009) Wherenext: a location predictor on trajectory pattern mining. In: ACM 15th SIGKDD international conference on Knowledge discovery and data mining, pp 637–646

Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the web. Technical Report, Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

Pang-Ning T, Steinbach M, Kumar V, et al (2006) Introduction to data mining. In: Library of Congress, p 74

Pappalardo L, Rinzivillo S, Qu Z, Pedreschi D, Giannotti F (2013) Understanding the patterns of car travel. Eur Phys J Spec Top 215(1):61–73

Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási A-L (2015) Returners and explorers dichotomy in human mobility. Nat Commun 6:8166. doi:10.1038/ncomms9166

Pennacchioli D, Coscia M, Rinzivillo S, Pedreschi D, Giannotti F (2013) Explaining the product range effect in purchase data. In: IEEE international conference on big data, pp 648–656

Rinzivillo S, Gabrielli L, Nanni M, Pappalardo L, Pedreschi D, Giannotti F (2014) The purpose of motion: learning activities from individual mobility networks. In: IEEE international conference on data science and advanced analytics (DSAA), pp 312–318

Rosvall M, Bergstrom CT (2011) Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. PloS One 6(4):e18,209

Trasarti R, Pinelli F, Nanni M, Giannotti F (2011) Mining mobility user profiles for car pooling. In: ACM 17th SIGKDD international conference on Knowledge discovery and data mining, pp 1190–1198

Zheng VW, Zheng Y, Xie X, Yang Q (2010) Collaborative location and activity recommendations with gps history data. In: ACM 19th international conference on world wide web, pp 1029–1038

Zheng Y, Xie X (2010) Learning location correlation from gps trajectories. In: IEEE 11th international conference on mobile data management (MDM), pp 27–32

Zhou C, Frankowski D, Ludford P, Shekhar S, Terveen L (2004) Discovering personal gazetteers: an interactive clustering approach. In: ACM 23rd annual international workshop on geographic information systems, pp 266–273