Optimal Spatial Resolution for the Analysis of Human Mobility

Michele Coscia[2], Salvatore Rinzivillo[1], Fosca Giannotti[1], Dino Pedreschi[3]

[1] KDDLab ISTI-CNR, Via G. Moruzzi, 1, Pisa, Italy, Email: rinzivillo@isti.cnr.it
[2] CID - Harvard Kennedy School, 79 JFK Street, Cambridge, MA, US, Email: michele_coscia@hks.harvard.edu
[3] KDDLab University of Pisa, Largo B. Pontecorvo, 3, Pisa, Italy, Email: pedre@di.unipi.it

***Abstract*—The availability of massive network and mobility data from diverse domains has fostered the analysis of human behaviors and interactions. This data availability leads to challenges in the knowledge discovery community. Several different analyses have been performed on the traces of human trajectories, such as understanding the real borders of human mobility or mining social interactions derived from mobility and viceversa. However, the data quality of the digital traces of human mobility has a dramatic impact over the knowledge that it is possible to mine, and this issue has not been thoroughly tackled so far in literature. In this paper, we mine and analyze with complex network techniques a large dataset of human trajectories, a GPS dataset from more than 150k vehicles in Italy. We build a multiresolution grid and we map the trajectories with several complex networks, by connecting the different areas of our region of interest. Then we analyze the structural properties of these networks and the quality of the borders it is possible to infer from them. The result is a significant advancement in our understanding of the data transformation process that is needed to connect mobility with social network analysis and mining.**

I. Introduction

The availability of massive network and mobility data from diverse domains has fostered the analysis of human behaviors and interactions. Traces of human mobility can be collected with a number of different techniques. We can obtain Global Positioning System logs, or GSM data referring to which cell tower a cellphone, carried and used by a person, was connecting. The result is a huge quantity of data about tens of thousand people moving along millions of trajectories.

This data availability leads to challenges in the knowledge discovery community. Several different analyses have been performed on the traces of human trajectories. For example, [5], [11] are two examples of studies able to detect the real borders of human mobility: given how people move, the authors were able to cluster different geographical areas in which people are naturally bounded. Another analysis example connects mobility with social networking [13], [1]. The fundamental question in these cases is: do people go in the same places because they can find their friends there or do people become friends because they go in the same places?

However, there is an important issue to be tackled before performing any kind of social knowledge extraction from mobility data. It has been proved that the data quality of the digital traces of human mobility has a dramatic impact over the knowledge that it is possible to mine. For example, in [12] authors perform a trajectory clustering analysis, with GPS data that are successively transformed in GSM-like data. They prove that the knowledge extracted with the semi-obfuscated data is more prone to data noise and performs worse. The conclusion is that mobility analysis should be performed with the high data precision that only GPS is able to provide.

An open question is left unanswered, and it is the main focus of this paper. Given that we use GPS data, how can we connect it to the territory? In general, GPS does not need to be mapped on the territory, as it already provides the coordinates of the person moving. However, usually we are dealing with two kinds of constraints. First, we are studying vehicles mobility, thus the "data points" are not free to move on a bi-dimensional surface, but they are constrained by the road graph. Second, if we want to apply social network analysis techniques on these data, such as the ones applied in [5], [11] namely community discovery over a network of points in space to find the borders of mobility, we need to discretize the territory in cells, as it is impossible to translate a continuous surface into a graph.

These two considerations force us to discretize the continuous human trajectories into a grid and then operate social network analysis on that grid. Should we use external information about the territory, such as the political organization in towns and municipalities? Or should we create a regular grid? In this paper, we propose an empirical study aimed at tackling these questions. We collect data from 150k vehicles moving on Tuscany (a region of Italy) road graph. We create a multiresolution grid representing Tuscany, for each cell size we generate a cell-cell complex network: cells c_1 and c_2 are connected with a directed edge if there is at least one trajectory starting from c_1 and ending in c_2. The edge is weighted according to how many trajectories connect the two cells.

Given our set of networks, one for each resolution of the grid, we analyze each network's structure. Given a collection of network measures, we are able to describe the network result of the translation of human trajectories to the grid network. We then apply community discovery on these networks, following our previous work [6], to identify the borders of human mobility. Our focus is to evaluate which grid resolution is leading to the best results. We evaluate each network results quantitatively, using different quality scores, and qualitatively, looking at the resulting borders and confronting them with what we know about Tuscany mobility.

The rest of the paper is organized as follows. In Section II we present the works related to the present paper: the connections between mobility and social network analysis and mining. Section III contains our data description and the creation of the multiresolution grid. In Section IV we evaluate the mobility grid and finally Section V concludes the paper presenting also some future insights.

II. Related Work

In literature, there are several works exploring the application of social network analysis to mobility data. One example is [11], where it is proposed to represent trajectories with a graph, then community discovery techniques are applied to discover areas frequently connected by the same set of trajectories. The mobility data used is manually submitted information about the movements of one dollar bills in the US territory. In [5] the same approach is implemented, but using GSM cellphone data: each trajectory is composed by the cell tower to which a particular device was connected. The main problems of these approaches is that the data source leads to unavoidable approximations, significantly lowering the quality of the results [12]. We improve over these works by using a reliable data source, i.e. direct GPS tracks.

Another class of works is more focused on the links between mobility and social relationships. In [13] a new link prediction technique (to quantify how much likely is to observe new connections in a complex network [8]) is proposed. The authors use for the prediction not only the topology of the graph, but also mobility information about the nodes of the network. The orthogonal problem is tackled in [1]: given the social relationships among a set of individuals, the study aims to predict the trajectories of these individuals. Not only GSM data about real people are used: some studies focus on movements of virtual spaceships in an online game [9]. Our paper focuses on the prerequisites of these works, i.e. how to define the underlying movement graph.

Finally, as community discovery plays an important role in this paper, we report two surveys about it: [3], focused on an empirical evaluation of many different algorithms; and [2], aiming to classify the many different community discovery approaches according to the underlying definition of community they operate on. When clustering algorithms enable the multi-level identification of "clusters-in-a-cluster", they are defined "hierarchical". This is useful for mobility networks, as it is necessary to explore borders at different granularity levels: conglomerates of cities, cities and even neighborhoods.

Some interesting algorithms are [7], [4], [10], employing different community clustering strategies. We focus on [7], as it is the algorithm we used in the framework presented in this paper. The Infomap algorithm uses the probability flow of random walks on a graph as a proxy for information flows in the real system and decomposes the network into clusters by compressing a description of the probability flow. The algorithm looks for a cluster partition M into m clusters so as to minimize the expected description length of a random walk. The expected description length, given a partition M, is given by $L(M) = qH(Q) + \sum_{i=1}^{m} p_i H(P_i)$. $L(M)$ is made up of two terms: the first is the entropy of the movements between clusters and the second is entropy of movements within clusters. The entropy associated to the description of the n states of a random variable X that occur with probabilities p_i is $H(X) = -\sum_1^n p_i \log_2 p_i$. The entropy is weighted by the probabilities with which they occur in the particular partitioning. More precisely, q is the probability that the random walk jumps from a cluster to another on any given step and p_i is the fraction of within-community movements that occur in community i plus the probability of exiting module i. Accordingly, $H(Q)$ is the the entropy of clusters names and $H(P_i)$ the entropy of movements within cluster i, including the exit from it. The algorithm uses a deterministic greedy search and then refines the results with simulated annealing.

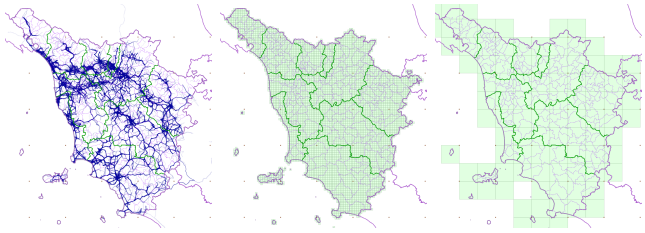


Fig. 1: *(Left)* A sample of the trajectory dataset used for the experiments. *(Center)* A partition based on a regular grid with cells of size $2000m$. *(Right)* A partition with a grid with 20,000m cell size.

III. Grid Creation

We used a dataset of spatio-temporal trajectories of private cars consisting of around 10M trips performed by 150,000 vehicles. The GPS tracks were collected by Octo Telematics S.p.A. Each trajectory is represented as a time-ordered sequence of tuples (id, x, y, t), where id is the anonymized car identifier, x and y are the latitude and longitude coordinates, t is the timestamp. The GPS tracks were collected from 1st May to 31st May 2011. The GPS device automatically starts collecting the positions when the car is turned on and it stops when it is turned off. Octo Telematics serves the 2% of registered vehicles in Italy. In our collection, we focus on the traces of the vehicles circulating in a bounding box containing Tuscany Region during the period of observation.

Given the spatial precision of GPS points, it is necessary to process the data in order to generalize neighbor points with a spatial region. Since the spatial precision of a GPS position can have an error of few meters, we need to determine the most suitable generalization for complex network analysis. Our approach consists in studying the properties of a complex network extracted from a regular grid composed of regular squares with edges of the same length.

As a starting point, we consider the bounding box containing our GPS trajectories, the minimum geographical rectangle containing all the points, say h and w respectively the height and width of the box. Chosen a length l for the edge of a cell, we divide the bounding box into a grid of cells with r rows and c columns, where $r = \lceil h/l \rceil$ and $c = \lceil w/l \rceil$. The resulting grid is aligned with the lower left corner of the original box.

There are several criteria to partition the territory for a spatial generalization step. In this research, we focus on the

spatial resolution of a regular division, since it enables us to control the granularity with a uniform distribution of the cells. Given a spatial partition, we can extract a network model to represent human movements on the grid. Each travel is mapped to a pair of cells: c_s, the starting cell, and c_e the destination cell. The network is determined by a set of nodes, representing the cells, and a set of edges, representing the travels between two cells. Each edge is weighted with the number of travels connecting the corresponding cells.

Varying the grid resolution, we are able to generate different network perspectives, and for each network we can derive basic statistics on its topology. Network basic statistics are a proxy to understand part of the topology of the network itself. Given the value of measures like average degree or path length, we understand if the network representation presents a topology likely to include a modular structure, thus community discovery can be used effectively.

To refer to distinct granularity, we call each network as "od_net_" followed by the cell size in meters of the underlying grid. Figures 2(a-b) depicts two different sets of statistics. Please note that the figures do not report the absolute value of the particular network measurement, but their relative value w.r.t the value obtained for the network with the largest grid cell, i.e. "od_net_40000". We cannot report the actual values for all networks for lack of space[1].

First, the number of nodes and edges, as depicted in Figure 2(a-b), drops dramatically by passing from a grid size of 200m to 10,000m, while sizes greater than 15,000m do not create much difference. Second, the number of edges drops with a different rate w.r.t the drop in the number of nodes. This is consistent to what we see in the Figure 2b: the average degree increases until a maximum density for a cell size in between 10-15,000m. The average path length drops consistently, while reciprocity and avg node weight increase: larger cells includes more trips and it is more probable to have reciprocal edges.

If we want significant results with community discovery we need dense networks with small-world properties with not too many small isolated components, and we want to achieve this objective with the smallest possible grid cell, thus with more nodes and edges, to have a more fine-grained description of reality. A preliminary conclusion may be that the optimal cell size should be around 5,000m: smaller cells generate networks with lower density and too many components.

Another important characteristic of the analyzed networks can be observed by when plotting their degree distributions (see Figure 2c). For clarity, we plotted only the degree distributions of the networks generated with a cell size of 500m, 1,000m, 2,000m, 5,000m, 10,000m, 20,000m and 40,000m. We can see that all the distributions present a heavy exponential cutoff. However, while the distributions for small cell sizes are similar, just on different scales, from cell sizes larger than 10,000m the exponential cutoff is increasingly stronger. This means that networks generated with larger cells lack of a peculiar characteristic of many large complex networks, i.e. the presence of hubs, a set of nodes very highly connected. As their average shortest path is still low, it means that their "small world" properties are not due to the network connectivity itself, but instead to the network small size. Thus, a cell size of 10,000m seems a reasonable upper bound for the cell size in our dataset. This upper bound can be explained by considering the distribution of lengths showed in Figure 2d: short-ranged travels (up to 10km) count for the 60% of the whole dataset. When increasing the grid size, small travels tend to be contained within the same cell (self-loop). This reduces the "power" of a cell of attracting other cells in its community, since there are less long-ranged trips.

[1]Complete table: http://www.di.unipi.it/∼coscia/borders/gridstatistics.htm

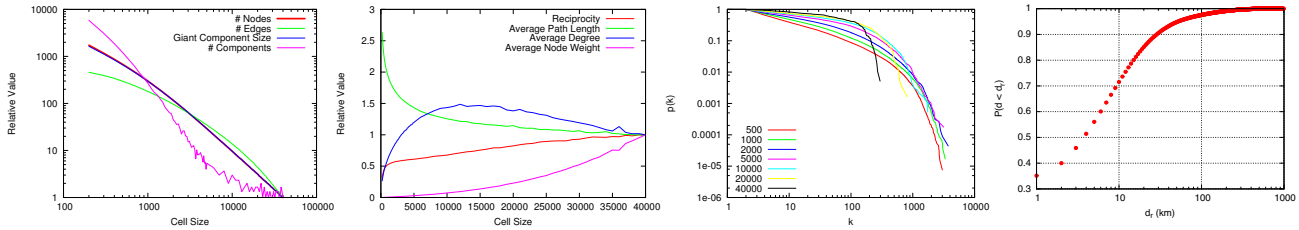
IV. Experiments

The communities extracted for each grid resolution are mapped back to the geography and they are used to compute thematic maps of the territory. Given a spatial resolution, for each community we retrieve all the cells associated to its nodes and we join them in a cluster, i.e. a geometric representation of the area covered by the community. An example of such thematic map is presented in Figure 3. Areas corresponding to different communities are rendered with different colors. There are holes in the reconstructed map, since these cells do not contain any travel. The phenomenon is more evident for smaller resolutions, as many small cells do not contain roads.

A. The Borders

We compare the resulting clusters with the existing administrative borders, in particular with the provinces, i.e. an aggregation of adjacent municipalities whose governance has the duty for traffic monitoring and planning. The borders of provinces are drawn with a thick green line in Figure 3*(Left)*. The emerging communities suggest small variations on the location of the actual borders. The four provinces of Pisa, Livorno, Lucca and Massa are aggregated in a single cluster, since the province of Lucca serves as collector of the mobility of the other three. Exploring the hierarchical aggregation of the communities resulting from Infomap (see Figure 3*(Right)*), it is evident the role of the central area of the province, where Lucca is located and where there exists a large vertical cluster (in blue) connecting the majority of the municipalities of the region. In fact, the cities of Pisa, Lucca, and Livorno form the so-called *area vasta* (i.e. large area), characterized by a large flow of commuters. Livorno is divided into two parts: the north part is included to the province of Pisa. A similar behavior is observed for the cluster containing the provinces of Firenze, Prato, and Pistoia. These big cities actually form a large metropolitan area with many commuters moving from one city to the other. The mobility is sustained a high capacity highway connecting the south with the north through the node of Firenze. The provinces of Siena and Arezzo maintain their own borders. The derived communities follow the borders of each municipality enforcing the internal role of each city as a minimum building block for human mobility borders.

Figure 4 shows the evolution of the clusters at different spatial granularity, namely with size 500m, 1,000m, 2,000m, 5,000m, 10,000m, and 20,000m. The first three snapshots show a coherent result, where the clusters identified within the high resolution grid of 500m are preserved in the successive steps. Starting from a cell size of 5,000m, the smaller clusters



(a) (b) (c) (d)

Fig. 2: Some statistics of the dataset: (a) number of nodes, edges and connected components, and giant component size; (b) reciprocity, average path length, degree and node weight; (c) the degree distributions for the networks generated with different cell sizes; (d) Cumulative distribution of trajectory lengths..

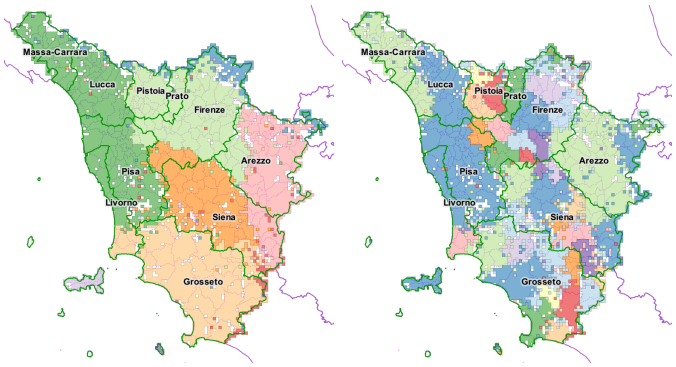


Fig. 3: *(Left)* The clusters obtained with grid cell size of 2000m. *(Right)* The clusters determined by the level 2 of the Infomap hierarchy for the same grid resolution.

disappear, e.g. the cluster between Siena and Grosseto, in red. When the spatial resolution became more and more coarse, we observe also a merging of distinct clusters in the same communities. In the clusters of resolution 5,000m, for instance, the cluster of Siena is merged with the cluster formed by Firenze, Prato, and Pistoia. In the other two successive steps the same phenomenon is repeated. At a resolution of 10,000m the cluster of Firenze is merged with the cluster of Pisa and Lucca. In the coarser version of the grid the resulting clustering actually contains all the grid cells in the same cluster.

From a qualitative evaluation of the resulting maps, we can infer an optimal grid cell size threshold of 5,000m: smaller granularities allow the identification of reasonable borders at the cost of a more complex computation and with the proliferation of very small local clusters.

B. Community Quality

Beside a visual comparison with the provinces, we analytically compared the partition derived by the community discovery approach and the partition determined by the actual administrative organization by means of two measures, namely *precision* and *recall*, largely used in information retrieval. Given two sets of objects, say C_1 and C_2, precision and recall are given by the formulas; $R(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1|}$, $P(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_2|}$. The recall measures how many of the objects in C_1 are present in C_2, while the precision measures the proportion of the object of C_1 in the cluster C_2. The recall of the set C_1 tends to one when all the elements of C_1 are present in C_2, it tends to zero otherwise. The precision of a cluster C_1 tends to zero when the proportion of elements

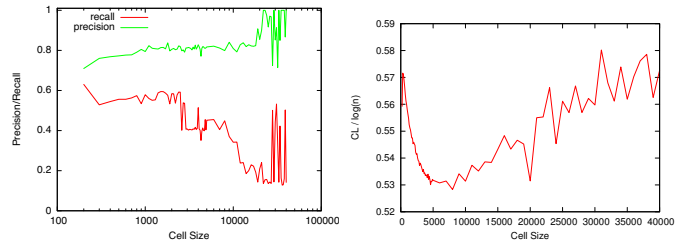
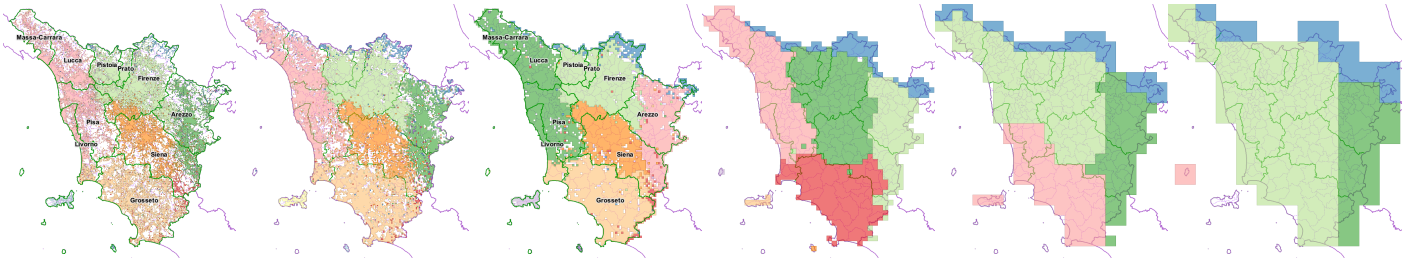


Fig. 5: (Left) The measures of precision and recall compared with the division of the territory into provinces; (Right) the adjusted codelength values of the extracted networks.

of C_1 is small with respect to the number of element in C_2, and it tends to one when the cluster C_2 contains only elements in C_1. In our setting, for each grid resolution we compare the sets of cells determined by the Infomap algorithm and the set of cells determined by the administrative borders. The administrative borders are represented by the set of grid cells whose cenrtoid is contained within the border interior (we use the centroid of the cell to avoid duplicated cells in different clusters). The measures expressed at the cluster level are extended to the global territory by means of the following procedure. Let denote with $\mathcal{C}_1$ the set of clusters determined by Infomap, and with $\mathcal{C}_2$ the clusters discovered by Infomap. First, for each cluster C_i in $\mathcal{C}_1$ we determine a cluster $C'_j = map(C_i) \in \mathcal{C}_2$, such that C'_j maximize the intersection with C_i among all the clusters in $\mathcal{C}_2$. Then, for each pair $(C_i, map(C_i))$ we determine precision and recall values. The overall similarity indexes is given by the weighted mean of each pair: $P(\mathcal{C}_1, \mathcal{C}_2) = \sum_{C_i \in \mathcal{C}_1} |C_i| P(C_i, map(C_i))$, $R(\mathcal{C}_1, \mathcal{C}_2) = \sum_{C_i \in \mathcal{C}_1} |C_i| R(C_i, map(C_i))$.

The resulting values for precision and recall are plotted in Figure 5 (left). The plot supports the observation made by means of the visual comparison of the clusters. Recall performs better for smaller grid size, namely up to 2,000m grid size, it decreases for values between 2,000m and 7,000m, and it has a drop for larger cell sizes. These results confirm and explain the clusters presented in Figure 4.

F-Measure is not the only evaluation test we can perform. Infomap calculates also the code length needed to codify the network given the community partition. Lower code lengths are better because they are the results of a better division in communities. Of course, the simple value of the code length is meaningless in our case, as the networks have very different scales (the number of nodes goes from 335k to 194 and the



(a) 500m (b) 1000m (c) 2000m (d) 5000m (e) 10000m (f) 20000m

Fig. 4: The resulting clusters obtained with different spatial granularities.

number of edges from 4M to 9k). Instead, we can adjust the code length with the number of nodes, as it is an information referred to how many bits are needed to represent all the nodes in the network. We adjust the code length as $CL_{adj} = \dfrac{CL}{\log_2 n}$, where n is the number of nodes in the network. The $\log_2 n$ term returns the number of symbols (bits) needed to code each node of the network taken separately, i.e. using a uniform code, in which all code words are of equal length. Since CL is the code length returned by Infomap, i.e. the number of symbols needed to code each node of the network given the community partition (that tries to exploit community information to use shorter code words), their ratio is telling us how CL improves over the baseline. If $CL_{adj} \geq 1$, then the community division is using the same number of symbols (or more) than the ones needed without the community, otherwise the compression is effective, and the lower value the better partition; therefore, CL_{adj} is scale independent.

The resulting plot of the CL_{adj} for all the networks generated is depicted in Figure 5 (right). As we can see, the adjusted code length decreases while approaching a cell size in the interval 5-10,000m, that is our minimum, and then increases again. At cell size 8,000m, the adjusted code length is slightly lower than 0.53, intuitively it means that the obtained code length is long as 53% of the baseline. This confirms the topology analysis of the networks performed in Section III, that identified the most promising cell sizes at values smaller than 10,000m. Moreover, the comparison of the plots in Figure 5 right and left show that the communities discovered for grid sizes up to 2,000m have comparable results at the cost of a complexity that decreases when the cell grid size increases. Beyond the grid size limit of 7-10,000m the spatial grid is no more able to capture local mobility behavior and the corresponding communities start getting worse.

V. Conclusion

In this paper we explore the influence of spatial generalization for the analysis of complex networks extracted from mobility data. We considered a large dataset of GPS trajectories, with a very precise spatial resolution, to derive a set of multi-resolution spatial grids. Each grid generates a mobility complex network where the nodes represent the cells of the grid and the edges represent the travels connecting the two cells. We studied several network statistics over the extracted networks and we applied a community discovery algorithm to derive the actual borders of human mobility. The extensive experiments show that the choice of the appropriate spatial resolution of the grid is critical for the generalization of mobility data. Finer resolutions create over detailed networks where smaller components are associate to several small clusters. Large cell sizes, on the contrary, generate an excessive aggregation of local movements. The optimal tradeoff is found within an interval of 2-7000m for grid cell size. This resolution allow the correct generalization of local trips, that represent the majority of human mobility, and the reduction of model complexity of the extracted communities, which yield a compact code representation.

Acknowledgements. Michele Coscia is a recipient of the Google Europe Fellowship in Social Computing, and this research is supported in part by this Google Fellowship. We received funding from the EU 7th Framework Programme (FP7/2007-2013), grant agreement n270833. We also acknowledge Octo Telematics S.p.A. for providing the dataset.

References

[1] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *KDD*, 2011.

[2] M. Coscia, F. Giannotti, and D. Pedreschi, "A classification for community discovery methods in complex networks," *Statistical Analysis and Data Mining*, vol. 4, no. 5, pp. 512–546, 2011.

[3] S. Fortunato and A. Lancichinetti, "Community detection algorithms: a comparative analysis: invited presentation, extended abstract," ser. VALUETOOLS '09. ICST, 2009, pp. 27:1–27:2.

[4] S. Papadimitriou, J. Sun, C. Faloutsos, and P. S. Yu, "Hierarchical, parameter-free community discovery," in *ECML PKDD*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 170–187.

[5] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz, "Redrawing the map of great britain from a network of human interactions," *PLoS ONE*, vol. 5, no. 12, pp. e14 248+, Dec. 2010.

[6] S. Rinzivillo, S. Mainardi, F. Pezzoni, M. Coscia, D. Pedreschi, and F. Giannotti, "Discovering the geographical borders of human mobility," in *Kunstliche Intelligenz*, 2012, p. In press.

[7] M. Rosvall and C. T. Bergstrom, "Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems," *PLoS ONE*, no. 6(4), p. e18209, Apr. 2011.

[8] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla, "When will it happen?: relationship prediction in heterogeneous information networks," in *WSDM*, 2012, pp. 663–672.

[9] M. Szell, R. Sinatra, G. Petri, S. Thurner, and V. Latora, "Understanding mobility in a social petri dish," *ArXiv e-prints*, Dec. 2011.

[10] L. Tang and H. Liu, "Relational learning via latent social dimensions," in *KDD*. New York, NY, USA: ACM, 2009, pp. 817–826.

[11] C. Thiemann, F. Theis, D. Grady, R. Brune, and D. Brockmann, "The structure of borders in a small world." *PloS one*, vol. 5, no. 11, pp. e15 422+, Nov. 2010.

[12] R. Trasarti, F. Pinelli, M. Nanni, and F. Giannotti, "Mining mobility user profiles for car pooling," in *KDD*, 2011, pp. 1190–1198.

[13] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *KDD*. New York, NY, USA: ACM, 2011, pp. 1100–1108.