Spatial and Temporal Evaluation of Network-Based 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 spatial grid and we map the trajectories to several complex networks, by connecting the different areas of our region of interest. We also analyze different temporal slices of the network, obtaining a dynamic perspective over its evolution. We analyze the structural properties of the temporal and geographical slices and their human mobility predictive power. 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 (GPS) 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, [16], [22] 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 [25], [4]. 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 [23] 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.

Several open questions are left unanswered, and some of them represent the main focus of this paper.

The first is connected to the temporal dimension, that is intrinsically linked to any movement data. For example, in [22] authors want to define the borders of human mobility, but they create a rather static snapshot by putting together movements without considering when these movements took place. Also works that consider temporal information usually use it as a continuum without discontinuity points or phase transitions.

In the real world, different events may dramatically change how people move on the territory. Such events may be unpredictable or not frequent, like natural disasters, but most of them are not. The most natural regular and predictable event is the transition between working and non-working days. During Saturdays and Sundays, people usually abandon their working mobility routines for different paths, obeying to completely different criteria. Another example may be organized human social events, like manifestations in a particular town or sport events.

The aim of this paper is to systematically prove that to mine human mobility and to extract from it useful knowledge is necessary to take into account these phase transitions. A dataset of undifferentiated trajectories, without taking into account when they were performed, may lead to increased and unexpected noise effects, lowering the quality of the results and, in extreme cases, hiding interesting patterns.

The second open question is orthogonal to the temporal dimension and it involves the spatial dimension. 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 [16], [22] 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 discrete spatial tessellation and then operate social network analysis on that partition. 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 a region of Italy, namely Tuscany. First, we address the temporal dimension problem by analyzing with complex network techniques our GPS trajectories and then understand their predictive power of the movements of our observed vehicles over the time span of a month.

Second, we address the spatial dimension problem by creating a multiresolution regular grid that covers Tuscany. We use this grid to generate different network perspectives over Tuscany mobility: grid 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 then weighted according to how many trajectories connect the two cells.

Both questions are addressed with the same complex network analysis technique, namely community discovery. Community discovery in complex networks aims to detect a graph's modular structure, by isolating densely connected sets of nodes called communities. For the temporal dimension, the communities observed at time t are used to predict the communities observed at time $t+1$. For the spatial dimension, we verify how well the community partition of a network generated with a particular grid resolution is able to describe the general structure with the minimum amount of information loss.

In the proposed framework, we generate sets of network with different criteria (temporal and spatial). We then apply community discovery on these networks, following our previous works [17], [6], to identify the borders of human mobility. Our focus is to evaluate which temporal perspective and which grid resolution is leading to the best results. We evaluate each network results both quantitatively, using different quality scores, and qualitatively, by 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. We introduce the community discovery problem definition and our adopted solution in Section III. We address our temporal analysis in Section IV: we map movements using the political division of the territory, we generated different temporal slices and we predict the community from one slice to the other. The creation of the multiresolution grid is presented in Section V. Finally Section VI concludes the paper presenting also some future insights.

II. Related Work

As stated in the introduction, there are several works in the field of human trajectories data mining. A class of these work is focused on applying frequent pattern mining to mobility data [13], [24], even borrowing techniques from biology mining [9]. A popular application to these techniques is the privacy-preserving anonymization of human movements [3], [12]. Different data sources can be used to obtain mobility data ranging from GSM [16], to GPS [17], to RF tags [11]. Sometimes, techniques developed for trajectory mining are then applied in other scenarios [10]. A good taxonomy for mining trajectories can be found in [1].

In literature, there are several works exploring the application of social network analysis to mobility data. Two examples are [22], [16]. In [22] for the first time it is proposed to represent trajectories with a graph, then community discovery techniques are applied to the graph to discover areas that are frequently connected by the same set of trajectories. The mobility data used is the manually submitted information about the movements of one dollar bills in the US territory[1]. In [16] 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. As stated in the introduction, the main problems of these approaches is that the data source leads to unavoidable approximations, significantly lowering the quality of the results [23]. We improve over these works by using a more reliable data source, namely direct GPS tracks.

Another class of works is more focused on the links between mobility and social relationships. In [25] a new link prediction technique is proposed. Link prediction in social network is the problem of quantifying how much likely is to observe new connections in a complex network given the current topology of the graph (see for example [19]). The advancement proposed in [25] is to use for the prediction not only the current topology of the graph, but also mobility information about the nodes of the network. The orthogonal problem is tackled in [4]: given the social relationships among a set of individuals, the study aims to predict which trajectories these individuals will decide to take. Not only GSM data about real people are used, there are some studies focusing on movements of virtual spaceships in a massive multiplayer online game, with a wide "universe" to move in [20]. Our paper is focused on the prerequisites of this class of works, namely how to define the movement graph needed for the analyses.

Finally, as community discovery is used as mean to assess the quality of a network representing human mobility, we report some references about it. Two comprehensive surveys about community discovery are [8], focused on an empirical evaluation of many different algorithms, and [5], that aims to classify the many different community discovery approaches according to the underlying definition of community they operate on. Several interesting community discovery algorithms are [18], [15], [21], [7], employing different community clustering strategies. We focus particularly on [18], as it is the algorithm we used in the framework presented in this paper.

[1] http://www.wheresgeorge.com/

This paper is built on previous work [6]. The focus of [6] was mainly on analyzing the geographical dimension of our problem. We extend over it by introducing a temporal analysis and extended experiments.

III. Community Discovery

An important part of our framework is the application of graph clustering algorithm on our network of trajectories. For this reason, in this section we introduce the problem of community discovery in complex networks along with the solution that we adopted.

An extensive survey, providing more background about community discovery, can be found in [5]. From [5] we know that clustering algorithms can provide extremely different results, according to their definition of what is a community in a complex network. For example, modularity maximization algorithms aim to maximize a fitness function describing how internally dense are the clusters according to their edges. Other techniques use random walks to unveil the modular structure of the network, since the random walker is trapped in denser areas of the network.

When clustering algorithms enable the multi-level identification of "clusters-in-a-cluster", they are defined "hierarchical". With this type of clustering algorithms, we can explore each cluster at several levels and possibly choose the level which, for example, best optimize some fitness function. This is a critical function for mobility networks, as in this scenario it is necessary to explore borders at different granularity levels: conglomerates of cities, cities and even neighborhoods.

Among the hierarchical clustering algorithms available in the literature, we choose the Infomap [18], which is one of the best performing non-overlapping clustering algorithms [8].

The Infomap algorithm is based on a combination of information theoretic techniques and random walks. It 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 intuition behind the Infomap approach for the random walks compression is the following. The best way to compress the paths is to describe them with a prefix and a suffix. Each node that is part of the same cluster M of the previous node is described only with its suffix, otherwise with prefix and suffix. Then, the suffixes are reused in all prefixes, just like the street names are reused in different cities. The optimal division in different prefixes represent the optimal community partition. We can now formally present the theory behind Infomap. 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 $H(X) = -\sum_1^n p_i \log_2 p_i$. In (1) 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, or city names in our intuition presented before, and $H(P_i)$ the entropy of movements within cluster i, the street names in our example, including the exit from it. Since trying any possible partition in order to minimize $L(M)$ is inefficient and intractable, the algorithm uses a deterministic greedy search and then refines the results with a simulated annealing approach.

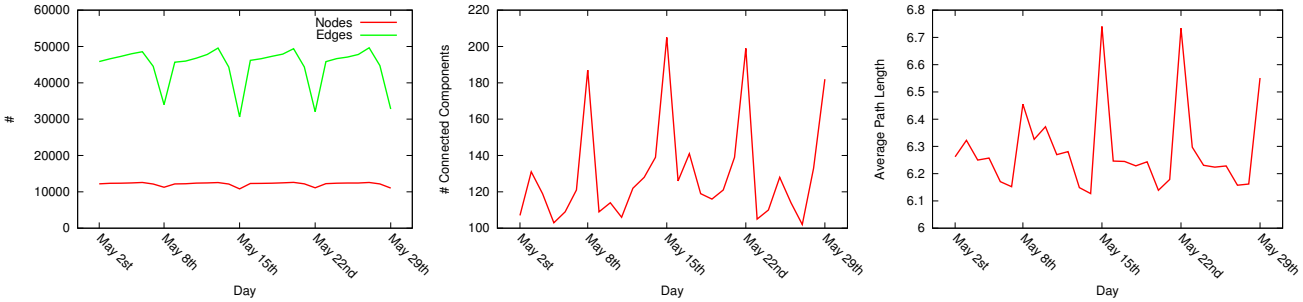
IV. The Temporal Dimension

In this section we explore the temporal issues of the application of complex network analysis to mobility data. As a proxy of human mobility, we used a dataset of spatio-temporal trajectories of private cars consisting of around 10M trips performed by 150,000 vehicles. These GPS tracks were collected by Octo Telematics S.p.A., a company that manages on-board GPS devices and data collection for the car insurance industry. 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 of the position. The GPS tracks were collected during a period of one month, 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. The log is transmitted to the server via GPRS connection. Octo Telematics serves the 2% of registered vehicles in Italy. In our collection, they collected the traces of the vehicles circulating in a bounding box containing Tuscany Region during the period of observation.

To apply complex network analysis on mobility data we first generalize the spatio-temporal positions by means of a spatial tessellation. This is already a challenge *per se*, and we deal more in deep with it in Section V. Since in this section we are focused on the temporal analysis of human mobility networks, we use a simple, sub-optimal, solution. We focus on the origin and destination of each travel of each vehicle. Using the spatial tessellation provided by ISTAT, the statistical bureau in Italy, we associate each origin (destination) to the census sector where the corresponding travel began (ended).

After this generalization step we can model human mobility by means of a graph where the nodes represent the census sectors and each edge represents the set of travels starting and ending within the corresponding census sectors. In particular, an edge connecting the nodes v_1 and v_2 is weighted with the number of travels starting from the sector associated to v_1 and ending at the sector associated with v_2. Moreover, since we are interested in studying the temporal evolution of the extracted network, we extracted several networks at different time intervals. In general, our method consists in selecting only the trajectories "alive" in the time period of study.

Which time interval should be adopted to analyze mobility from a temporal perspective? We fixed a minimum temporal



(a) Number of nodes/edges. (b) Number of components. (c) Average path length.

Fig. 1: Some statistics for the daily network snapshots.

interval of one day and then we generated daily snapshots of the movement graphs. We depict in Figure 1 some of the basic statistics of these daily networks. We can see that there are remarkable differences between weekday and weekend networks (we recall that May 8th, 15th, 22nd and 29th 2011 were Sundays). Saturdays and Sundays networks usually have less edges, somewhere between 62-75% of the edges of a weekday (Figure 1(a)); they have more components, i.e. the networks are more fragmented, with areas not connecting at all to each other (Figure 1(b)); and finally their average path length is significantly higher, May 8th presents a lower peak, but the whole preceding week was lower than the following, due to the fact of Italian national holiday of May 1st (Figure 1(c)).

We can conclude that we expect different results from the weekdays and weekend networks, as their topology is significantly different. Thus, we considered three distinct intervals for each week: weekdays, i.e. day from Monday to Friday, weekends, i.e. Saturday and Sunday, and the whole week, obtaining 12 networks for the four weeks considered.

A. Weeks, Weekdays and Weekends Network Statistics

We now take a look to the basic statistics of the extracted networks, as they are able to unveil preliminary differences between the different network views of the dataset. For a deeper explanation about concepts such as "connected component" or "average path length" we refer to [14]. In Table I we reported the following statistics: number of nodes (column $|V|$), number of edges (column $|E|$), average degree (column Avg Degree), number of connected components (column $|CC|$), relative size of the giant component (column GC Size %), reciprocity (column Reciprocity) and average path length (column ℓ). In each row of the table we grouped three kinds of networks: Week, Weekdays and Weekends. Each entry is the average value of the measure of the four networks in each network type.

As we can see, the number of nodes of the Week networks is slightly higher than the number of nodes of the Weekdays networks. This means that during weekends people sometimes choose to reach places that were never visited during weekdays, although in general their destination set is slightly narrower. A big difference between Weekdays and Weekend networks is highlighted by the average degree: during weekends the paths chosen by users are significantly less than what expected by the smaller set of destinations. This means that during weekends the same few paths are under a higher mobility pressure.

Weekends networks appears to be more fragmented (the networks on average present 69 components against the 26 for Weekdays networks), however almost 98% of destinations are still part of the network's giant component. The giant component size is important because if most of the census sectors are actually isolated from each other, the community discovery loses significance. Also, we know that in Weekends networks we will find 68 very small and isolated communities, that can be ignored for our analytical purposes.

Reciprocity is the ratio of bidirectional edges over the total number of edges. This measure is lower during weekends, implying that in that period of the week people are more likely to stay in the places they reach. Finally, the average path length unveils the fact that we are dealing with classical small-world networks [26]: the average number of edges to be crossed to go from any node to any other node is below 5. An exception is represented again by Weekends networks: although the average path length ℓ is low, it is higher than the other network view, and with a lower number of nodes. We can conclude that the long-range connectivity in Weekends network is weaker than expected.

We depict in Figure 2 the degree distributions of our 12 networks. We colored in red the Week networks, in blue the Weekend networks and in green the Weekdays networks. The distributions represent an argument in favor of our chosen methodology. The three kinds of networks present very similar degree distributions, while they differ from each other. While the Weekday networks still can approximate the Week ones, the same does not hold for the Weekend network, that dramatically differ from the previous two. The statement that the Weekend network cannot be useful in predict general patterns of the week, and vice versa, proves to be intuitive. We provide evidences in favor of this statement in Section IV-C.

B. Evaluation

To evaluate how much the communities discovered in a particular temporal interval are meaningful, we check if they are preserved in different time periods, by comparing each other by means of the measures of precision and recall. We

Network	$\|V\|$	$\|E\|$	Avg Degree	$\|CC\|$	GC Size %	Reciprocity	ℓ
Weeks	17468.8	218474.0	25.01	20.25	98.8351%	0.276039	4.25788
Weekdays	16568.2	167425.0	20.21	26.00	98.7612%	0.263951	4.50722
Weekends	13895.5	72055.8	10.37	69.00	97.9868%	0.247907	5.33465

TABLE I: The average statistics of the different network views of the dataset.

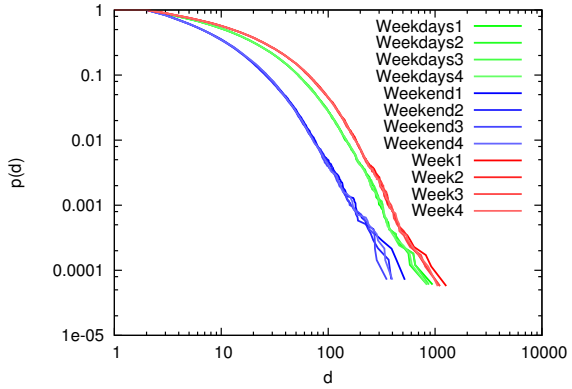


Fig. 2: The cumulative degree distribution of our networks.

call *clustering* the aggregation of a set of objects into subgroup and each subgroup is called a *cluster*. Formally, a clustering $\mathcal{C}$ is the union of its own clusters $\{C_1, C_2, \ldots, C_n\}$. Given two clusters, say C_1 and C_2, precision and recall are given by the formulas;

$$R(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1|}; \quad P(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_2|}$$

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 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.

To extend the measures from the cluster level to the global evaluation of the two clusterings, we propose the following procedure. 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 maximizes 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 average of each pairs:

$$P(\mathcal{C}_1, \mathcal{C}_2) = \sum_{C_i \in \mathcal{C}_1} |C_i| P(C_i, map(C_i))$$

$$R(\mathcal{C}_1, \mathcal{C}_2) = \sum_{C_i \in \mathcal{C}_1} |C_i| R(C_i, map(C_i)).$$

C. Experiments

We now take a look at the results of the application of our workflow to the real world data presented in Section IV.

1) The Human Mobility Borders: Weekdays vs Weekends: We start by taking a look at the top-level clusters extracted by the hierarchical version of Infomap algorithm. In Figure 3 we show a matrix of all the clusterings, for each week and for each network type (Weekday, Weekend and Whole Week). In general, the clusters look mostly compact, with the exceptions of the areas where we can find the highway entrances, as they are of course catalyst hubs of long-range trips. The white areas are regions where no trip started or ended and for this reason are excluded from the network representation. In general, some useful insights can be extracted to improve human mobility management, as the merging of Pisa and Livorno provinces (cluster on the middle-left, light green color in the Weekday map for the week 1, top left corner of Figure 3): the two cities are divided only for political reasons, but they are very close and part of a strongly connected area, as witnessed by the way GPS traces move. At least on the higher level, those areas need to coordinate.

In this case, we can exploit the power of the cluster hierarchy to have a finer description of the mobility borders. In Figure 4 we zoomed into the Pisa-Livorno cluster for the Weekday network of week 1: on the left side we have the cluster at the top level of the hierarchy, on the right side the cluster at the second level. As we can see, at this level the provinces of Pisa and Livorno are correctly split, meaning that there is a border at least at the city level, and our framework is able to detect it by exploring the cluster hierarchy.

Let us now focus on the differences between the Weekdays and the Weekends clusters. The Weekends clusters look as compact as the Weekdays clusters and the *quantity* of differences looks lower than expected from the intuition that the network statistics gave us (see Section IV-A). However, the *quality* of the differences is very important: in week 1 the Pisa-Livorno cluster expanded and now includes also the cities of Lucca and Viareggio (black and brown clusters north of Pisa in Figure 3, respectively), that are naturally separated from Pisa and difficult to reach. The inclusion is probably due to a higher rate of long-range trips to neighboring towns usually difficult to reach, but appealing to spend some free time during the weekend. Also, the Florence cluster (orange in Figure 3(a)) is split in two (pink and blue cluster in Figure 3(b)). These changes are very important qualitatively, as these clusters involves a large share of all Tuscany trips. In general, the strong noise effect created by weekend movements is evident for week 3. The Whole Week clusters tend to look in general more alike the Weekdays, but Weekend clusters perturb their borders: the Weekday Lucca cluster (light purple) is split in three parts in the Weekend cluster and this causes its disappearance also from the Whole Week clusters, divided between the pre-existing Florence (blue) and Massa-Carrara-Viareggio (green) clusters. Similar instances of these problems

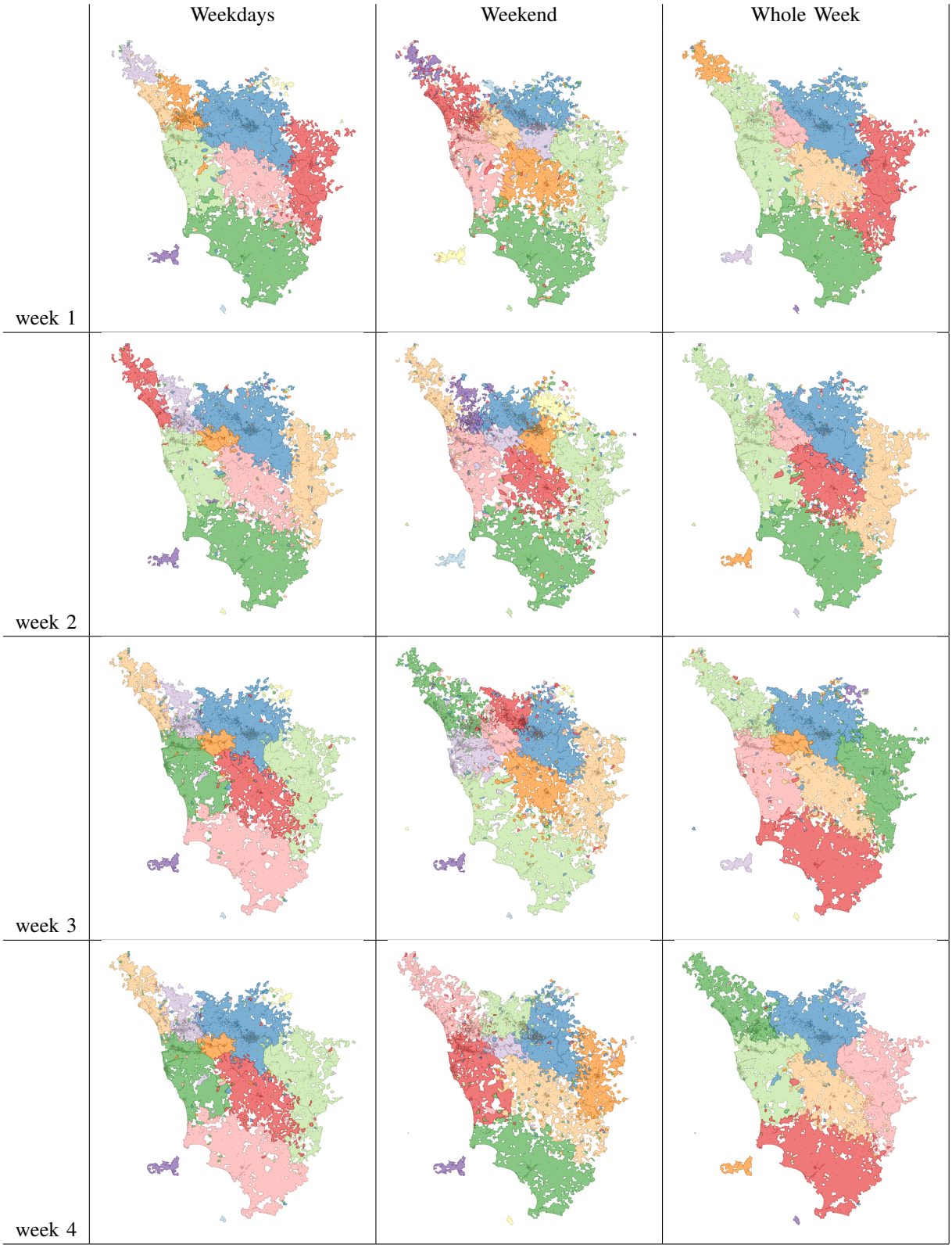


Fig. 3: The Tuscany mobility clusters (top level of the hierarchy).

(a) Weekend 2 Clusters at the highest level of hierarchy.

(b) Pisa-Livorno Cluster at level 1.

(c) Pisa-Livorno Cluster at the second level of hierarchy.

Fig. 4: Exploring the second level of hierarchy clusters.

are present in each week, intuitively proving the noisy effect of weekend trajectories.

2) Weekdays and Weekends Quality Evaluation: We now evaluate the predictive power quality of the cluster extracted from the various networks. We make use of the Precision and Recall measures as defined in Section IV-B. The general procedure is the following: we consider the clusters extracted in the network representing the first week and then we calculate the Precision and the Recall for each of the other networks. A high score means that the target network contains similar clustered information, therefore is predictable using the source network. The results are depicted in Figure 5.

To understand how to read Figure 5, let us consider its leftmost scatter plot: in this case the source clustering is calculated using each of the Weekday network. Each dot represent the quality results, according to Precision (x axis) and Recall (y axis), for each of the other network considered in this article. The dot color represent the kind of network to which we are applying the prediction: green for Weekday, blue for Weekend and red for Week. Since we are dealing with four weeks and three different network views for each week (Weekday, Weekend and Week) we have a total of 48 points, 4 of which scores 1 for both Precision and Recall as they are clusterings applied to themselves: since we are considering the leftmost plot, the 4 perfect scores are all green dots, each representing a Weekday clustering applied to itself.

Now we can find evidences about the lower quality of the Weekend predictions by considering all the three plots. As we can see, the central plot, the one representing the prediction results using the Weekend clusters, scores lower performances for all networks, both in Precision and Recall. Not only Weekend clusterings are not able to predict Weekday and Week clustering: they also score poorly in predicting themselves, proving that from one weekend to another the trajectories vary significantly, and therefore they cannot be predicted efficiently using the simple assumption that the same period in the week should behave in the same way across time.

The other side of the story also holds: not only Weekend cannot predict with high scores, but it also cannot be predicted. By considering the leftmost and the rightmost plot, we see that the distribution of the colors of the dots is not random, but they are clustered in precise areas of the plot. Focusing on the blue dots (Weekend), we notice that they always tend to be clustered in the lower side of the plot, i.e. the one characterized with lower Recall scores. In conclusion, Weekend clusterings are behaving like an unpredictable, and unreliable for prediction, class of phenomena.

However, we also notice that unexpectedly Prediction scores for blue dots in the leftmost and rightmost plots are not the lowest in absolute terms. The explanation lies in the nature of the Week datasets: by definition it also includes the trajectories originated during weekends. This inclusion is lowering the Precision scores for the prediction Weekday to Week and from Week to Weekday. In fact, in the leftmost plot the green dots (Weekday to Weekday predictions) tend also to score better according to prediction, while this does not hold for red dots (Weekday to Week predictions). For the rightmost plot, being the Week-based prediction affected by the weekend data, we have a more confused evaluation. We can conclude the following thing: to integrate weekday data with weekend data is equivalent to manually introduce noisy data points, and it should be avoided. It is not true that weekday data can correct the noise of weekend data, or that weekend data can somehow compensate or integrate weekday data. If we want to have reliable models for the majority of human movements, then we should use only weekday data. If we want to have also a model for the irregular human movements during weekends, we need to sacrifice the prediction quality.

3) Systematic vs Occasional Trajectories: To evaluate the influence of systematic movements over human mobility, we propose here a method to select the very frequent movements among the travels of each vehicle. Given a vehicle v we select

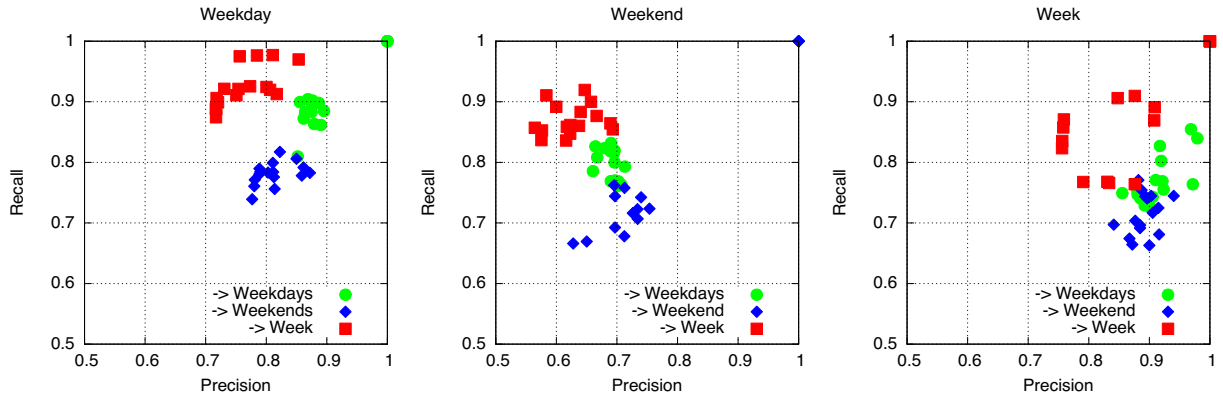


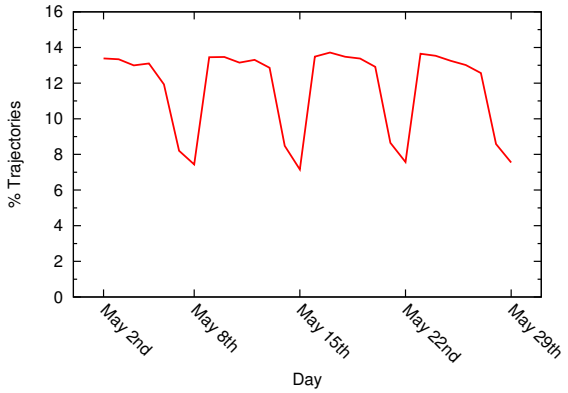
Fig. 5: The Precision and Recall values for the predictions using Weekday (Left), Weekend (Center) and Week (Right).

all the travels associated to v and we cluster them according to their starts and ends, i.e. the trips starting from similar places and ending in similar places are aggregated in the same cluster. To extract the clusters from the set of origins and destinations of each vehicle we adopt a density based clustering method, namely OPTICS [2]. OPTICS is one of the best candidates clustering methods since it is very robust to noise, it does discover the natural number of clusters in the dataset analyzed and it can be customized by providing specific distance functions. In our case, we defined a distance function based on the relative distance between origin and destination point of each trajectory. In particular, given two trajectories t_1 and t_2 with end points respectively (s_1, e_1) and (s_2, e_2), the distance between t_1 and t_2 is defined as

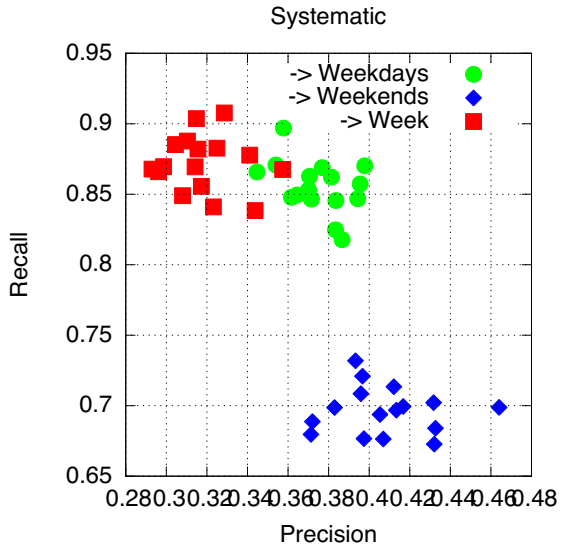
$$d(t_1, t_2) = \frac{d(s_1, s_2) + d(e_1, e_2)}{2}.$$

The OPTICS algorithm start exploring the dataset by evaluating the neighborhood of each trajectory according to the distance function provided and to a distance threshold ϵ, which defines a minimum radius around the current object, and a minimum number of point *MinPts* expected to be found within the given radius. When a trajectory has enough neighbors in its radius, it is said to be a core trajectory and its cluster is expanded as far as other density points are reachable. In our experiments we focused on the analysis of very compact clusters that could represent systematic travels. Thus, we used a distance threshold of $250m$. The cluster with the highest cardinality is selected as the most frequent and, hence, as the systematic movement of the vehicle. By repeating this procedure for all the vehicles, we can select a subset of movements that are frequently performed by them. Starting from this subset we apply the same method presented in the previous sections: the trajectories are generalized to the spatial tessellation, they are projected in a specific time interval and a complex network is extracted.

In Figure 6(a) we report the relative distribution of systematic trajectories in our dataset. For each day, we divide the number of trajectories classified as "systematic" by the number of the total trajectories that were followed during that day. We can see that there is a strong difference between weekdays and weekends. During weekdays, more than 13% of trajectories



(a) The daily distributions of systematic trajectories.



(b) The quality measure of the communities extracted from the systematic networks.

Fig. 6: The daily distributions of systematic trajectories: for each day the share of trajectories that are systematic.

are systematic. An exception is Friday, as pre-weekend day, although always at least 12% trajectories are systematic during that day. During weekends, these shares drop to around 8.5% during Saturdays and 7.5% during Sundays. Therefore we can safely state that our assumption, i.e. that systematic trajectories are followed more commonly during weekdays, is sustained by evidence.

The impact of the systematic component of mobility is also evident from the results of community discovering on these network. Figure 6(b) show the measures of precision and recall resulting from the comparison of the systematic networks with the networks explored in Section IV-C2. The separation between weekend prediction and week/weekday prediction is here even more evident. In general, the values of recall are very low if compared with Figure 5. This is due to a sparse network extracted from a limited number of trajectories. We can verify this by looking at the statistics of the networks extracted from the systematic trajectories. In Table II we report the same statistics of Table I, but for the systematic networks instead of the networks created with the complete set of trajectories. We can see that there is an increased number of connected components (ten times more) and the giant component is significantly smaller. Each separated component generates an isolated community, thus greatly lowering the Recall score. The values of precision, in this case, are neatly separated: weekday and week networks maintain similar values, whereas weekend networks have poor prediction performances.

V. The Geographical Dimension

As we saw in the previous section, 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, i.e. the minimum geographical rectangle that contains all the points, say h and w respectively the height and width of the box. Chosen a length l for the edge of each 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.

By varying the grid resolution as shown in Figure 7, we are able to generate different network perspective of human mobility, and for each network we can derive basic statistics on its topology. Network basic statistics are an important proxy to understand part of the topology of the network itself. Given the values of measures like average degree or path length, we can understand if the network representation presents a topology that is likely to include a modular structure, thus community discovery can be used effectively.

To refer to distinct granularities, we call each network as "od_net_" followed by the cell size in meters of the underlying grid used to generate the network. Figures 8 and 9 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[2].

Looking at Figures 8 and 9 we can state some interesting things about the networks generated with different grid resolution levels. First, the number of nodes and edges 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: we can see in Figure 8 that the green line starts from below the red line, then it is higher in the interval 4,000m-17,000m then drops again. This is consistent to what we see in Figure 9: the average degree increases until a maximum density for a cell size in between 10-15,000m, then slightly lowers. The average path length drops consistently, while reciprocity and average node weight increase: this is expected as bigger cells includes more trips and it is more probable to have reciprocal edges.

If we want significant results with community discovery we need generally 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, or with too many components.

Another important characteristic of the analyzed networks can be observed by when plotting their degree distributions (see Figure 10). 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

[2]The complete table can be retrieved at the following URL: http://www.di.unipi.it/~coscia/borders/gridstatistics.htm

Network	$\|V\|$	$\|E\|$	Avg Degree	$\|CC\|$	GC Size %	Reciprocity	ℓ
Weeks	11861.2	26349.8	4.443	240.75	94.7033	0.0290827	7.85268
Weekdays	11059	22748.2	4.11398	269.5	93.8127	0.0270297	8.40738
Weekends	7375.25	8745.5	2.37158	667.75	76.4822	0.0172919	14.4058

TABLE II: The average statistics of the different network views of the dataset, using only systematic trajectories.

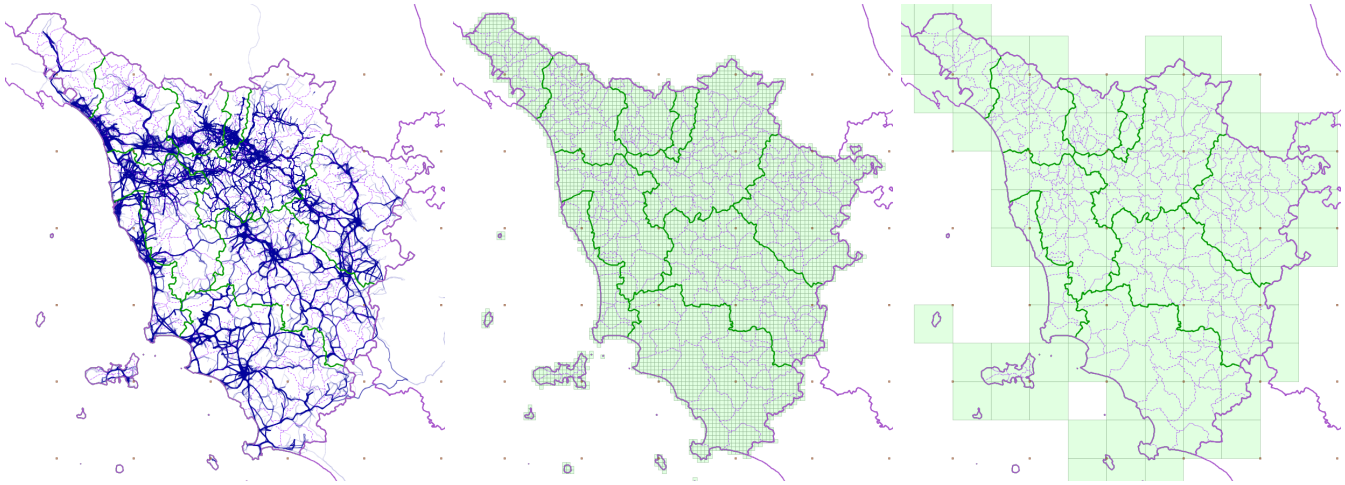


Fig. 7: *(Left)* A sample of the trajectory dataset used for the experiments. *(Center)* A partition based on a regular grid with cells of size 2000m. *(Right)* A partition with a grid with 20,000m cell size.

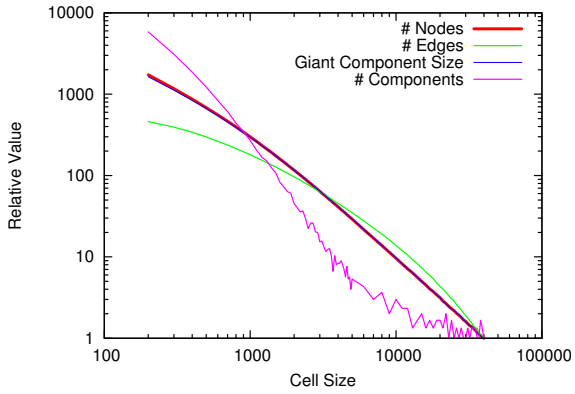


Fig. 8: Some statistics of the extracted networks, relative to the values of the "od_net_40000" network: number of nodes, edges and connected components, and giant component size.

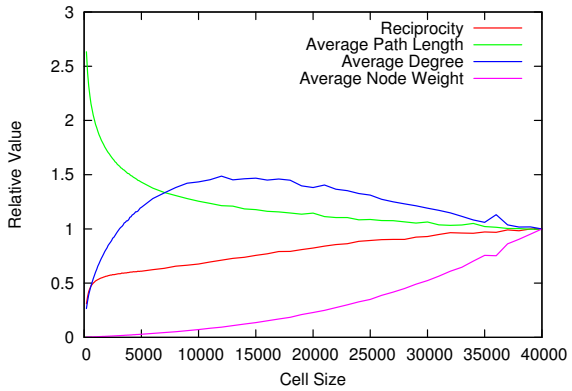


Fig. 9: Some statistics of the extracted networks, relative to the values of the "od_net_40000" network: reciprocity, average path length, degree and node weight.

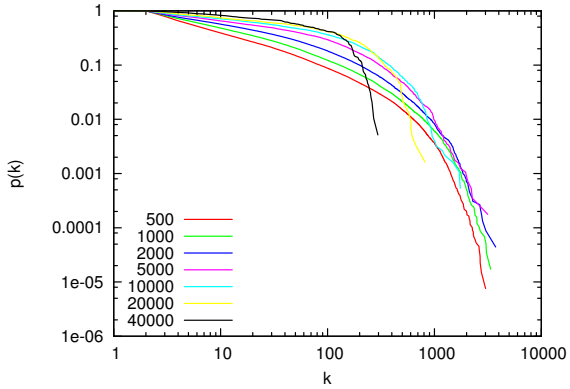


Fig. 10: The degree distributions for the networks generated with different cell sizes.

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 11: 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, generating a self-link in the resulting network. This reduces the "power" of a cell of attracting other cells in its community, since there are less long-ranged trips.

A. *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,

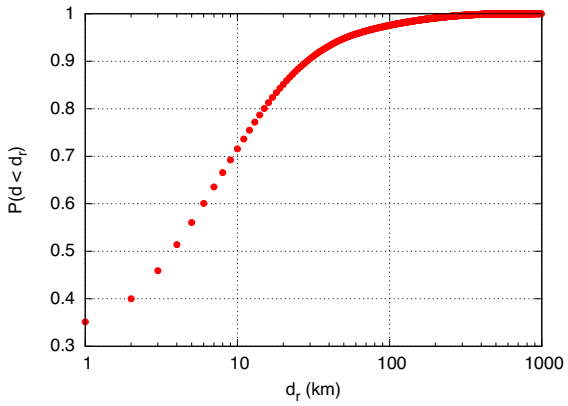


Fig. 11: Cumulative distribution of length of trajectories in our dataset.

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 12. For clarity, areas corresponding to different communities are rendered with different colors. It can be noted the presence of holes in the reconstructed map, since there cells of the spatial partition that do not contains any travel. This phenomenon is more evident for smaller resolutions, where it is possible to find cells that do not contains any road and, thus, any car travel.

1) 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 12*(Left)*. From the figure it is evident how the emerging communities suggest small variation on the location of the actual borders. For example, 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 12*(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 (highlighted 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), which is characterized by a large flow of commuters. The province of Livorno is divided into two parts, where the north part is included to the province of Pisa and, by transitivity, with the other twos. 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 a huge number of commuters moving from one city to the other. This mobility is also sustained by the high capacity of the highway that connects the south with the north through the node of Firenze. The citizen of the city, moreover, have a special reduction for the toll. The provinces of Siena and Arezzo maintain their own borders. It is worth noting that 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 13 shows the evolution of the clusters at different spatial granularities, 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 disappear, like for example the cluster between Siena and Grosseto, highlighted 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.

2) 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 the two measures of *precision* and *recall* introduced in Section IV-B. 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 resulting values for precision and recall are plotted in Figure 14. 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 13.

Precision and Recall are not the only evaluation measures we can exploit. 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 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 with the following formula:

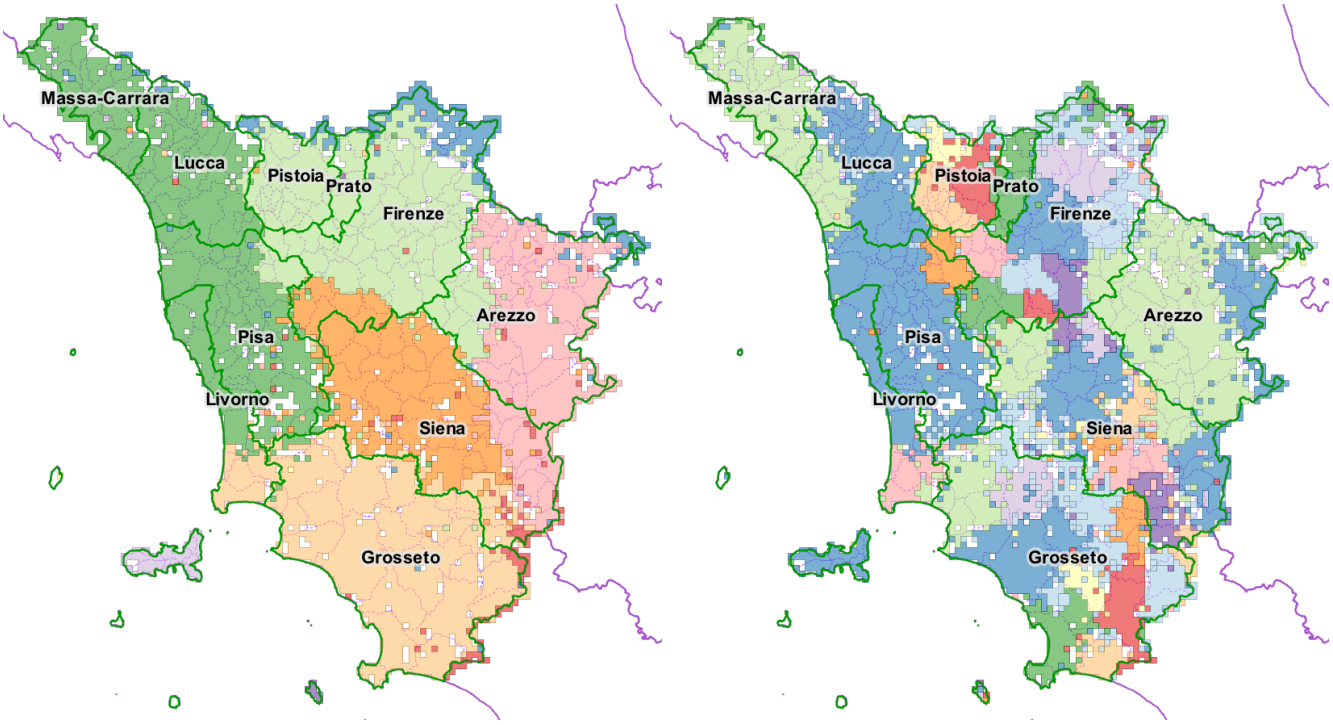


Fig. 12: *(Left)* The clusters obtained with grid cell size of $2000m$. *(Right)* The clusters determined by the level 2 of the Infomap hierarchy for the same grid resolution.

$$CL_{adj} = \frac{CL}{\log_2 n},$$

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 much better is CL 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. For this reason, CL_{adj} is scale independent.

The resulting plot of the CL_{adj} for all the networks generated is depicted in Figure 15. 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 at the beginning of this section, that identified the most promising cell sizes at values smaller than 10,000m. Moreover, the comparison of the plots in Figure 15 and Figure 14 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 and their complexity start getting worse.

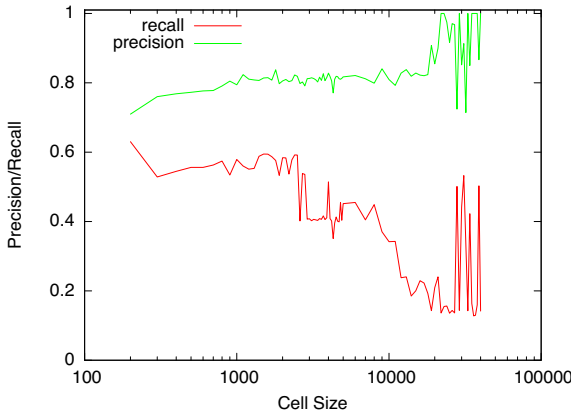
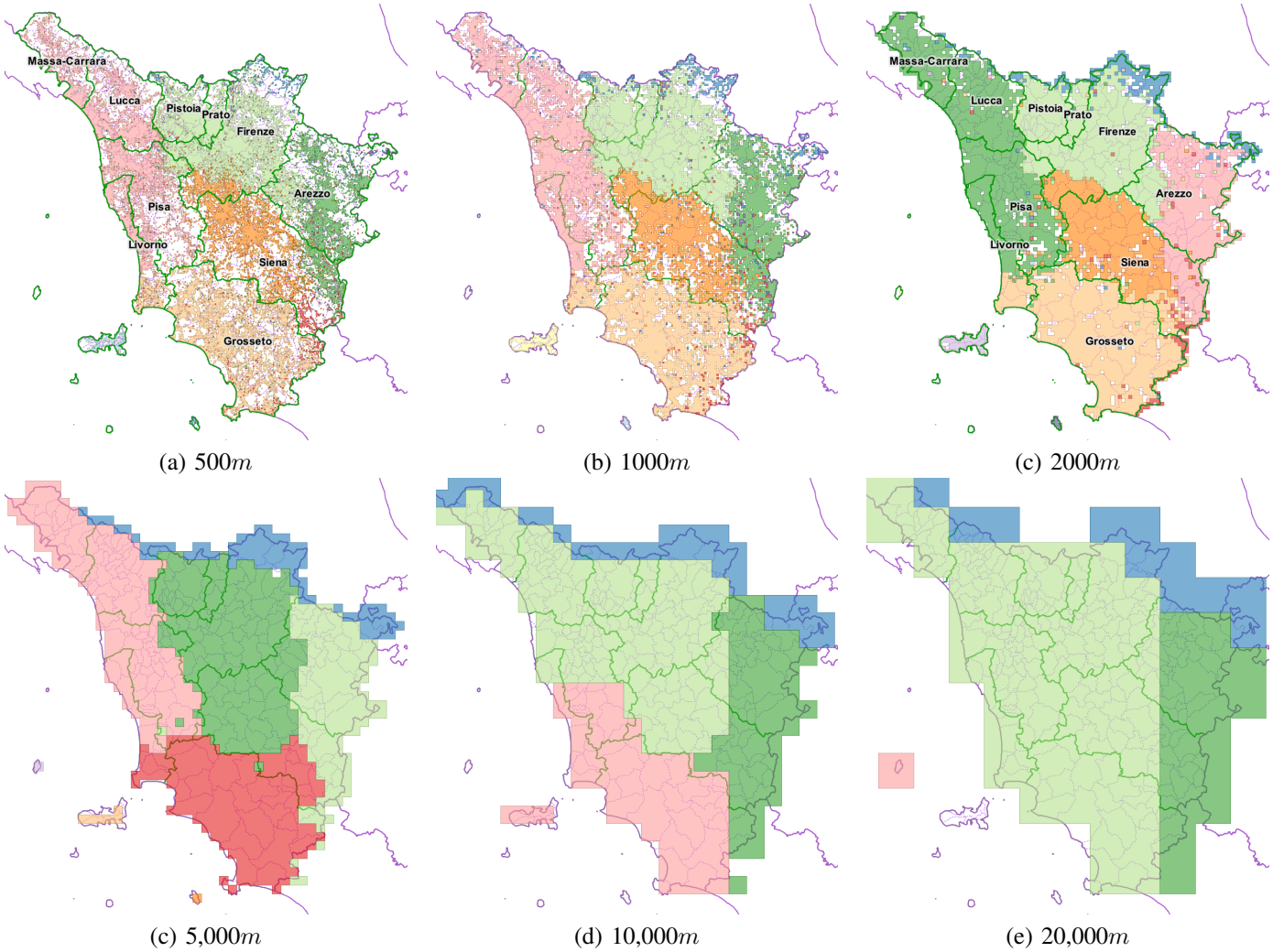


Fig. 14: The measures of precision and recall compared with the division of the territory into provinces

VI. Conclusion

In this paper we explore the influence of the temporal and spatial dimension for the analysis of complex networks extracted from mobility data. We considered a large dataset of GPS trajectories, with a very precise temporal and spatial resolution. From these trajectories, we derive different network perspectives: the first set is generated by defining time intervals (i.e. weekdays and weekends), the second set is generated by defining a set of multi-resolution spatial grids. We studied several network statistics over the extracted networks and we



(a) 500m (b) 1000m (c) 2000m

(c) 5,000m (d) 10,000m (e) 20,000m

Fig. 13: The resulting clusters obtained with different spatial granularities.

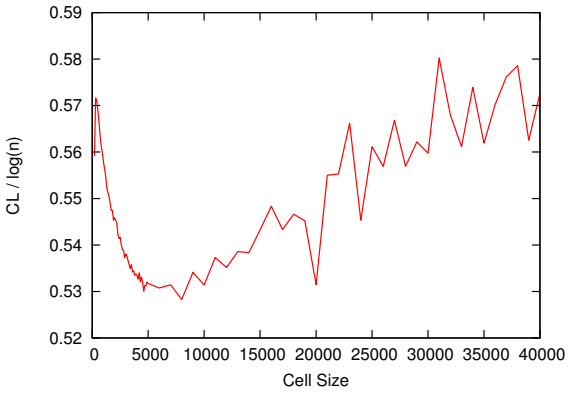


Fig. 15: The adjusted codelength values of the extracted networks.

applied a community discovery algorithm to understand how the temporal and the spatial dimension affect the problem of detecting the actual borders of human mobility. The extensive experiments show that the choice of the appropriate temporally bounded data and spatial resolution of the grid is critical for the network study of mobility data. Temporally, data from periods of increased unpredictability can introduce noise and lower the quality of mobility prediction. Spatially, 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. We provided a framework to understand how to detect the optimal spatiotemporal tradeoff. We detected the optimal spatial resolution, that allows 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. We also detected that to maximize the usefulness of the mobility clusters, one has to rely on systematic trajectories, that are more frequent and predictable.

Acknowledgements. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n270833. We also acknowledge Octo Telematics S.p.A. for providing the dataset.

REFERENCES

[1] G. L. Andrienko, N. V. Andrienko, P. Bak, D. A. Keim, S. Kisilevich, and S. Wrobel, "A conceptual framework and taxonomy of techniques for analyzing movement," *J. Vis. Lang. Comput.*, vol. 22, no. 3, pp. 213–232, 2011.

[2] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: ordering points to identify the clustering structure," *SIGMOD Rec.*, vol. 28, no. 2, pp. 49–60, Jun. 1999. [Online]. Available: http://doi.acm.org/10.1145/304181.304187

[3] F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang, "Trajectory anonymity in publishing personal mobility data," *SIGKDD Explorations*, vol. 13, no. 1, pp. 30–42, 2011.

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

[5] 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.

[6] M. Coscia, S. Rinzivillo, F. Giannotti, and D. Pedreschi, "Optimal spatial resolution for the analysis of human mobility," *ASONAM*, 2012.

[7] M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, "Demon: a local-first discovery method for overlapping communities," in *KDD*, 2012, pp. 615–623.

[8] 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.

[9] A. Jawad, K. Kersting, and N. V. Andrienko, "Where traffic meets dna: mobility mining using biological sequence analysis revisited," in *GIS*, 2011, pp. 357–360.

[10] G. Krempl, Z. F. Siddiqui, and M. Spiliopoulou, "Online clustering of high-dimensional trajectories under concept drift," in *ECML/PKDD (2)*, 2011, pp. 261–276.

[11] Y. Liu, L. Chen, J. Pei, Q. Chen, and Y. Zhao, "Mining frequent trajectory patterns for activity monitoring using radio frequency tag arrays," in *PerCom*, 2007, pp. 37–46.

[12] A. Monreale, G. L. Andrienko, N. V. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, and S. Wrobel, "Movement data anonymity through generalization," *Transactions on Data Privacy*, vol. 3, no. 2, pp. 91–121, 2010.

[13] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, "WhereNext: a location predictor on trajectory pattern mining," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '09. New York, NY, USA: ACM, 2009, pp. 637–646. [Online]. Available: http://dx.doi.org/10.1145/1557019.1557091

[14] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

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

[24] F. Verhein and S. Chawla, "Mining spatio-temporal patterns in object mobility databases," *Data Min. Knowl. Discov.*, vol. 16, no. 1, pp. 5–38, Feb. 2008. [Online]. Available: http://dx.doi.org/10.1007/s10618-007-0079-5

[25] 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.

[26] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks." *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998. [Online]. Available: http://dx.doi.org/10.1038/30918