Künstl Intell
DOI 10.1007/s13218-012-0181-8

FACHBEITRAG

Discovering the Geographical Borders of Human Mobility

Salvatore Rinzivillo · Simone Mainardi · Fabio Pezzoni · Michele Coscia · Dino Pedreschi · Fosca Giannotti

Received: 31 October 2011 / Accepted: 12 March 2012

Abstract The availability of massive network and mobility data from diverse domains has fostered the analysis of human behavior and interactions. Broad, extensive, and multidisciplinary research has been devoted to the extraction of non-trivial knowledge from this novel form of data. We propose a general method to determine the influence of social and mobility behavior over a specific geographical area in order to evaluate to what extent the current administrative borders represent the real basin of human movement. We build a network representation of human movement starting with vehicle GPS tracks and extract relevant clusters, which are then mapped back onto the territory, finding a good match with the existing administrative borders. The novelty of our approach is the focus on a detailed spatial resolution, we map emerging borders in terms of individual municipalities, rather than macro regional or national areas. We present a series of experiments to illustrate and evaluate the effectiveness of our approach.

1 Introduction and Related Work

In recent years the analysis of human behavior has received increasing attention by the scientific community. This is partly due to the availability of massive network and mobility data from diverse domains together with novel analytical paradigms which place human relationships or their mobility patterns at the center of investigation. Inspired by application domains such as social networks [1, 5], human mobility [12], the interplay between the two [24], and so on, over the last few years, broad, multidisciplinary, and extensive research has been devoted to extracting non-trivial knowledge from network and mobility data. Predicting future links between actors of a network [4, 19], detecting and studying the diffusion of information between them [13, 27], mining frequent patterns of user behavior [3, 7, 25], and predicting human mobility patterns [18] are only a few examples of the problems studied by researchers including physicists, mathematicians, computer scientists, and sociologists.

In this paper we address a set of fascinating questions that were recently posed in [23]: “Are there geographical borders that emerge from the way people use the territory for their daily activities?”, “If so, how can these borders be found?”, “Do these borders match the administrative borders?”. Two recent studies have tackled questions on a large geographical scale based on both mobile activity in the US [23] and on social interactions in the UK [21]. Thiemann et al. [23] analyzed the human mobility network extracted from the logs

S. Mainardi · F. Pezzoni
Department of Information Engineering, University of Pisa, Pisa, Italy

S. Mainardi
e-mail: simone.mainardi@iit.cnr.it

F. Pezzoni
e-mail: fabio.pezzoni@iit.cnr.it

S. Rinzivillo (✉) · M. Coscia · F. Giannotti
ISTI-CNR, Pisa, Italy
e-mail: rinzivillo@isti.cnr.it

M. Coscia
e-mail: michele.coscia@isti.cnr.it

F. Giannotti
e-mail: fosca.giannotti@isti.cnr.it

D. Pedreschi
Department of Informatics, University of Pisa, Pisa, Italy
e-mail: pedre@di.unipi.it

Published online: 29 March 2012

provided by the project *Where's George?*[1]: using a stochastic method, they extracted a partition of regions according to a fitness function based on modularity maximization. The experiments were performed in a large scale setting in which the minimum spatial granularity was given by a zip code area in the United States. Ratti et al.'s approach [21] also adopts the modularity function as an objective function to delineate borders emerging from the network extracted from a large database of telecommunication records. However, it is well known in the literature that modularity has an inherent resolution problem, which causes small communities to be ignored and merged together [11].

In this research, we address the problem of finding the borders of human mobility at the lower spatial resolution of municipalities or counties. The aim of discovering borders at a meso-scale is to provide decision-support tools for policy makers, capable of suggesting optimal administrative borders for the government of the territory. To this purpose, we need fine-grained results since we are working with smaller areas than those used by Thiemann et al. [23] and Ratti et al. [21]. We therefore use another state-of-the-art community discovery algorithm, namely Infomap [22], which has been shown to perform better than any other modularity maximization algorithm [17].

We study the problem of finding the geographical borders that emerge from the mobile activity of people and compare them with the existing administrative borders of cities, municipalities and provinces. "Do people move and interact within specific areas?", "Are those areas bounded somehow?", "Do these boundaries correspond to the administrative borders, which are defined *a priori*, usually without taking into account the social connections, the everyday needs of commuters, families, and so on?", "Do the borders change during the day, or during the week?", "Can we spot some seasonality?". Motivated by these questions, we apply Social Network Analysis techniques to mobility data. Our aim was to better understand human mobility patterns, in a new fashion, based not on the interaction of humans themselves, but rather on the underlying, hidden connections between different places. We apply Community Discovery algorithms to the network of geographic areas (i.e., each node represents a cell or region of movements) in order to find areas that are densely connected by the visits of different users.

The main contribution of the paper consists in the extraction of a fine-grained mobility network to model human behavior along with the use of a state-of-the-art community discovery algorithm to detect relevant communities corresponding to geographical areas. In addition, we provide several experiments based on a real-life scenario of GPS tracked vehicles.

The remainder of the paper is organized as follows. In Sect. 2 we present a general method for extracting a complex network from mobility data using a multi-scale approach. Section 3 introduces the Infomap algorithm. Section 4 shows the settings of our experiments and our main results. We conclude with a brief discussion in Sect. 5.

2 Mapping Mobility to Complex Networks

Our objective is to determine the influence of social behavior in a territory, in particular to evaluate how the current administrative borders represent the real basin of human movements. In general, we want to determine groups of regions such that the inner movements within a group are more frequent than the movements towards the other groups. We thus propose a general framework based on the following steps: (1) the territory is partitioned by means of a non-overlapping spatial tessellation whose regions will serve as spatial references; (2) the movements are generalized to the spatial tessellation; (3) they are then coded by means of a directed weighted graph; and (4) the graph is then analyzed to extract the communities within it.

A spatial tessellation serves as the basic level of detail to represent movements. The spatial granularity of the tessellation strictly depends on the precision of the data available. The movement of people can be tracked using various technologies such as GPS devices, GSM network logs, Wi-Fi fingerprints, and RFID tag readings. Each of these tracking technologies has its own spatial precision and uncertainty: for example, GSM data usually has a spatial granularity corresponding to the spatial extent of each cell. GPS based locations, on the other hand, are so precise that it is very unlikely that two different positions will share the same coordinates. It is thus useful to generalize each point to a spatial area, either using existing spatial coverage, such as cadastral data, census sectors, or cellular network coverage, or by aggregating together similar points by means of convex hulls, buffers, or clustering [2, 9, 16].

In a broad sense, the movement of an object can be described as a sequence of trips, i.e. the movements from an origin to a destination. Depending on the capabilities of the tracking device and the application scenario, each trip can be described in terms of a trajectory, i.e. a sequence of time-stamped locations collected along the route of the trip. In a scenario where GSM data are used, it is very likely that the movement is described in terms of a pair of cells: a first cell where the call began, and a second cell where the call ended [21]. In rare cases, it is possible to follow the devices moving in the network on the base of the cells crossed. This sampling frequency issue also generally applies to other movement data collections. For example, GPS devices have the potential to collect several points per second; however, to preserve the battery life of devices and to

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

minimize the quantity of data exchanged, the sampling frequency is determined according to the application scenario.

Here we consider two different approaches to represent movement: on the one hand we consider each movement as a pair consisting of the *origin* and *destination*; on the other hand we maintain the detailed information, according to the capabilities of the collection device used, regarding the route followed between the two locations. In the first case, movements are transformed into a sequence of visited places which are annotated with the corresponding temporal information. This type of representation provides a precise vision of movement dynamics and, at the same time, allows the data to be handled on a large scale. In addition, the emphasis on the data is placed on *where* people move rather than *how* they reach their destinations. Thus, given a trip—a detailed description of how we determine trips is given in Sect. 4—of a user, we only map the origin and the destination to the corresponding regions (we call this mapping strategy *Origin-Destination mapping*). In the second case, we map the entire route on the spatial tessellation. Depending on the technology used to log movements, the continuous path is often approximated with a sequence of sampled time-referenced observations. In this case, mapping to the spatial tessellation is performed by mapping each sampled point to the corresponding cell in the tessellation (we refer to this strategy as *Segments mapping*).

Once each position has been generalized according to the spatial tessellation, the transformation of the movements to a graph $G(V,E)$ is straightforward: each region R is mapped to the vertex $v_R \in V$ and the flow from a region R to a region Q is mapped to the edge (v_R, v_Q) whose weight is proportional to the density of movements between the two regions.

The original problem of finding clusters consisting of areas with a dense exchange of travelers between them and a low exchange of travelers across this set of areas can then be reduced to the problem of finding clusters of nodes that are densely connected internally and sparsely connected with the rest of the network. This last formulation is the most popular problem definition of many community discovery algorithms [8, 10].

3 Identifying Clustered Structure

Community algorithms can provide extremely different results depending on their definition of what a community in a complex network is [8]. For example, modularity maximization algorithms aim to maximize a fitness function describing how internally dense the clusters are according to their edges. Other techniques use random walks to unveil the modular structure of the network, with denser areas of the network where the random walker is "trapped".

When clustering algorithms enable the multi-level identification of "clusters-in-a-cluster", they are defined as being "hierarchical". With this type of clustering algorithm, we can explore each cluster from several levels and possibly choose the level, for example, which best optimizes a particular fitness function. Among the hierarchical clustering algorithms available in the literature, we choose the Infomap, which is one of the best performing non-overlapping clustering algorithms [17].

The Infomap algorithm is based on a combination of information-theoretic techniques and random walks. It uses the probability flow of random walks [20] 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 as follows. Each node is described with a prefix and a suffix. The prefix refers to the cluster the node belongs to. The suffix univocally identifies the node within its cluster. The suffixes are then reused in all prefixes, just like street names are reused in different cities. If a node n in a path belongs to the same cluster of its predecessor then n is described only by its suffix, otherwise both prefix and suffix are used. The optimal division into different prefixes represents the optimal community partition.

We can now formally present the theory behind Infomap. The expected description length, given a partition M, is given by:

$$L(M) = qH(Q) + \sum_{i=1}^{m} p_i H(P_i). \tag{1}$$

$L(M)$ is made up of two terms: the first is the entropy of the movements between clusters and the second is the entropy of movements within clusters. The entropy associated with 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 one 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. Thus, $H(Q)$ is the entropy of clusters names, or city names (as presented above), 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 [6] and then refines the results with a simulated annealing approach [14].

4 Experiments and Discussion

As a proxy for human mobility, we used a dataset of GPS tracked vehicles in the area around Pisa. The vehicles have a GPS tracker on board as required by a special insurance policy that vehicle owners are required to subscribe to. The GPS tracker collects timestamped points and transmits them to the insurance server at an average rate of one point every 30 seconds when the vehicle is moving or, at most, every two kilometers.

However, for each vehicle the server only has a sequence of received points without any semantic annotation. Thus, it is necessary to partition that sequence into sub-sequences that represent a single journey each. We used a time threshold to determine journeys: if a point in the sequence has been collected at least 20 minutes after the previous point, the current journey ends and a new one begins [26].

We observed approximately 38,000 vehicles for a period of five weeks (from June 14th to July 19th, 2010). The frequency of the time sampling enabled us to explore different temporal resolutions when generalizing the data to a given spatial tessellation. As presented in Sect. 2, we adopted two different strategies to generalize the timestamped locations. We used Origin-Destination (OD) mapping to simplify each trip by only considering the first and the last points. Secondly, we used Segment (SEG) mapping to generalize each timestamped point of a trajectory to the spatial tessellation.

We adopted a spatial tessellation based on existing census sectors as provided by the ISTAT, the Italian National Bureau of Statistics. The reasons for this are manifold: this data is publicly available and contains information such as population, commuters and segmentation by age; it provides a hierarchical representation of the territory (e.g. the administrative area of a city can be described as the union of all its statistical sectors) and thus it enabled us to compare directly the analytical results with the existing administrative borders, i.e. the existing aggregation of census sectors. In addition, the extent of each sector is proportional to the population density distribution, thus in the urban centers the sectors are very fine-grained, whereas in rural areas the extent is very large.

It would be possible to adopt a regular rectangular grid to generalize movements, however, new challenges could arise. First, the regular partition does not take population distribution into account. This could create biases within the cells, since many of them would not contain any trajectory, which would generate holes in the final clustered coverage. Secondly, it is not clear which spatial resolution should be adopted, since a very fine-grained grid could increase the biases in the cells and a coarser partition could fail to take important areas into account. Thus, to generate a suitable regular grid for this kind of analysis, it is necessary to have a multi-resolution grid that enables the extent of each cell to be adjusted dynamically. For example, in [15] a traffic generalization framework is shown that exploits this multi-relational approach using a dynamic traffic unit to aggregate trajectories. Census sectors can be aggregated into a four level hierarchy: the base level contains the census sectors in which each area corresponds approximately to a city block. Several adjacent sectors make a *comune* (hereafter, a municipality). Several adjacent municipality make up a *provincia* (hereafter, a province).

Table 1 Features of the OD and SEG mapping graphs

	OD mapping	SEG mapping
Nodes	7,878	8,156
Edges	474,964	292,524
Avg. node weight	350.03	4,279.65
Avg. edge weight	2.91	57.88
Avg. shortest path	2.6850	6.13534
Clust. coeff.	0.1705	0.4221
Diameter	7	17

The census sector level is used for the generalization in accordance with the two mapping strategies. The network derived by the OD mapping contains a link between two nodes v_R and v_S if at least one vehicle starts from region R and stops at region S, where R and S are the regions associated with v_R and v_S respectively. The weight of the link is given by the number of all the vehicles starting and stopping in the two nodes. The network determined by the SEG mapping has a link between two nodes if at least one trajectory of a vehicle exists whose two consecutive points can be mapped to v_R and v_S respectively. The generalized sectors are then clustered according to the community discovery method and the result is compared with the aggregation of sectors at a *town* level.

Table 1 shows some features of the OD and the SEG mapping. Although the census sectors we considered did not change from one mapping to another, SEG has about 300 more nodes than OD. These nodes correspond to “transit” census sectors, which are neither the source nor the destination of any journey. Conversely, the difference in the number of edges between SEG and OD means that there are adjacent census sectors crossed by many journeys. For example, consider two adjacent census sectors encompassing a highway. Many vehicles will pass through these sectors when traveling on the highway, regardless of their source (destination). Despite this, only one edge linking these two highway sectors exists. Indeed, information on the number of journeys passing through these two sectors can still be read from the weight associated with the edge interconnecting them.

By observing the average node weight in the OD mapping, we can see that on average each sector is the source (destination) of approximately 350 journeys. Similarly, the

Table 2 Level-1 Clusters with PageRank greater than 5 % in the OD mapping. Clusters are indicated with the name of a *Comune* they encompass

Cluster	PageRank %
Pisa	16.93
Viareggio	13.04
Lucca	12.07
Empoli	11.98
Livorno	8.33
Pistoia	8.14
Pontedera	7.13
Montecatini Terme	6.42

average edge weight indicates that two sectors are the source (destination) of on average about three journeys. If we note the average node weight in the SEG mapping, we can see that each census sector is reached and/or left approximately 4,000 times. This apparently huge number is due to the fact that many sectors are crossed in each journey and this directly translates into an increment of the weight associated with incoming and outgoing edges. Finally, the average edge weight indicates that about 60 vehicles travel between each two adjacent census sectors.

4.1 Origin-Destination Mapping

The clustering method produced a 4-level hierarchy of clusters for the OD mapping. At the first level there are 96 clusters, which are further divided into smaller clusters at lower levels of the hierarchy (e.g. 513 at the second level). Figure 1 shows the resulting level-1 clusters. Out of these 96 clusters, we select 19 with a PageRank value greater than 0.1 % and in particular eight with a PageRank value greater than 5 % (see Table 2). These clusters are named after the largest municipality that they contain. We will always refer to each cluster by that name, when not ambiguous. Thus, the majority of the journeys involve very few clusters—a journey has 98.13 % chance of beginning (ending) in a sector of the 19 highest-PageRank clusters. These few clusters are also the most geographically extended, spanning almost all the territory we considered—containing 7,527 census sectors, i.e. 95.54 % of the total. Furthermore, they consist of *geographically adjacent* census sectors, although OD mapping contains many connections between non-adjacent areas.

To validate our results, we will now discuss the main clusters using background knowledge of the interested areas, starting from the *Pisa* cluster, which is highlighted in a dark blue in Fig. 1. This cluster consists of the majority of the statistical sectors in Pisa plus the sectors of its adjacent municipalities, i.e. Cascina, Calci, San Giuliano Terme and Vecchiano. Traditionally, these towns are referred to as “*Area Pisana*”[2], which can be considered as an enlarged metropolitan area centered around Pisa. Recently, the regional government promoted a strategic development project for this area (named “*Piano Strategico dell’Area Pisana*”) with the objective of designing an integrated mobility plan for the five municipalities.

The other clusters with high PageRank can also be interpreted by means of well-known geographical and socio-demographic features. The reasons for these relations are due to both the historical relationship and the morphology of the territory. For example, the cluster of Viareggio, located in the north-west and in green in Fig. 1, covers an area widely known as the “*Versilia*”[3]. Other examples include, but are not limited to, the cluster of Lucca and the “*Piana di Lucca*”[4], Montecatini Terme and the “*Valdinievole*” as well as Empoli and the “*Valdarno Inferiore*”[5]. Thus, we can state that mobility patterns reflect the strength of the socio-economic relations between geographical areas very well.

It is worth noting how the cohesion of sectors within the same municipality is maintained after the clustering, apart from one small exception. For example, the sectors belonging to the administrative border of Pisa are assigned to different clusters, in particular the south-west sectors are associated with the adjacent cluster of Livorno. These sectors, in fact, correspond to the beaches and are a frequent destination for people from Livorno during the summer period. The main seaside destination, on the other hand, for people in Pisa is the west of the city, adjacent to the estuary of the river Arno and the beaches in Vecchiano.

Finally, it is important to note that it is not a necessary condition for a cluster to consist of geographically adjacent sectors. In fact, the OD mapping has many edges that represent long-range trips. However, the clusters consist of adjacent sectors, in particular the urban zones, where the local mobility is very dense and, hence, very effective in attracting the zones. Figure 1 *(Right)* shows an example of a cluster within non adjacent sectors. The teasels are rendered with a color proportional to the volume of mobility flows. It should be noted that some satellite areas are assigned to the cluster.

4.2 Segments Mapping

The clustering method for the SEG mapping produced a 5-level hierarchy of clusters. At level 1 there are 11 clusters, which are shown in Fig. 2 *(Left)*. At this level, the number of

[2] http://it.wikipedia.org/wiki/Pisa#Area_pisana.

[3] http://en.wikipedia.org/wiki/Versilia.

[4] http://it.wikipedia.org/wiki/Piana_di_Lucca.

[5] http://it.wikipedia.org/wiki/Valdarno#Valdarno_inferiore.

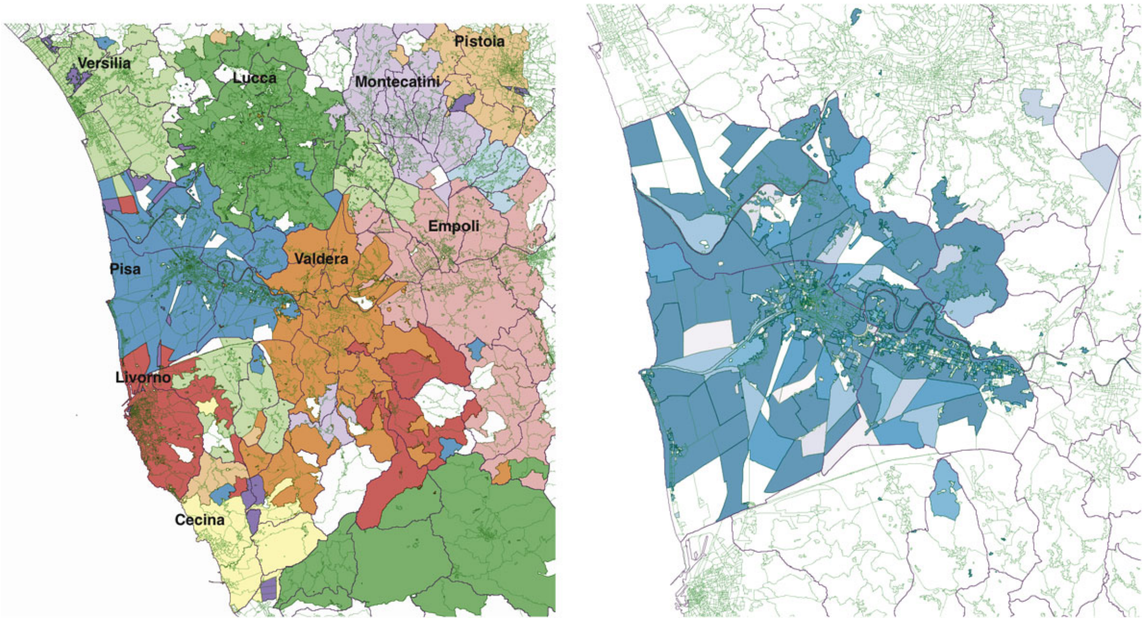


Fig. 1 Visualization of the clusters identified by the OD mapping. In reference to the existing administrative borders, the perimeter of each town is drawn with a thicker line. (*Left*) The regions within the same cluster are given the *same color*. (*Right*) Visualization of the level 2 sub-clusters of the Pisa cluster with different levels of brightness according to the internal volume of trajectories: the sub-clusters with the higher mobility flows have a *darker color*

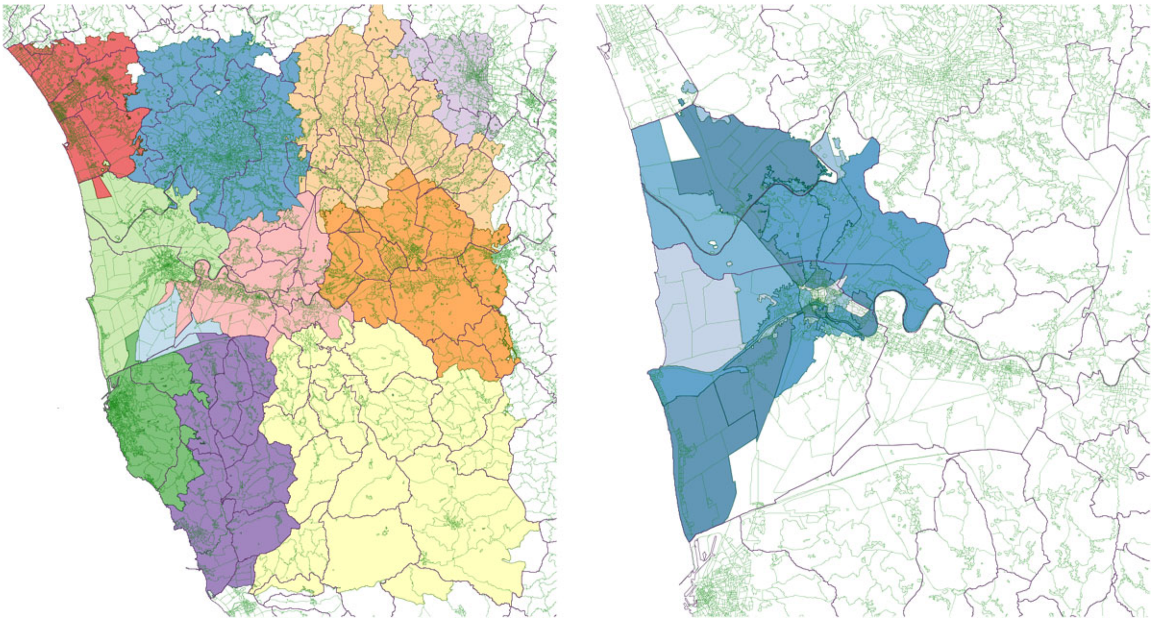


Fig. 2 Visualization of the clusters determined from the SEG mobility network. In reference to the existing administrative borders, the perimeter of each town is drawn with a thicker line. (*Left*) The regions within the same cluster are given the same color. (*Right*) Visualization of the level 2 sub-clusters of the Pisa cluster with different levels of brightness according to the internal volume of trajectories: the sub-clusters with the higher mobility flows have a *darker color*

clusters is significantly less than in the OD mapping. Hence, the clustering method aggregates census sectors better. This is reasonable since it is a direct consequence of the majority of very short-ranged edges, which allow the connection only among geographically adjacent sectors. Moreover, their PageRank never assume values less than 0.6 %, whereas in the OD mapping there are 77 clusters whose PageRank is less than 0.1 %. In contrast to the OD mapping, in SEG cluster coverage has an interesting and meaningful size at level 2 as well.

At the second level of clustering it is possible to investigate how the sectors are aggregated. An example of the hierarchical aggregation of a single level 1 cluster is shown in Fig. 2 (Right). In this case, all the clusters consist of adjacent sectors, as opposed to the OD mapping. SEG clustering produces very compact clusters, all centered around urban centers as in the OD clustering. The clusters of Viareggio, Pistoia, Lucca, Livorno and Empoli have approximately the same geographical extension. The clusters of Montecatini Terme and Volterra, on the other hand, are bigger, encompassing geographical areas which, in OD, are considered as different clusters. The clusters of Pisa and Pontedera are significantly different compared to the OD mapping because the municipalities of Cascina and Calci belong to the cluster of Pontedera.

5 Conclusions and Future Work

In this paper we have presented a general method to discover geographical areas determined by the mobility behavior of people. The method is based on the extraction of a multiscale mobility network, representing the flows of movement between a set of regions. The network is analyzed using one of the best performing non-overlapping community discovery algorithms. We presented an extensive experimental setting where the results are discussed and commented on with reference to the domain knowledge of the territory. The clusters discovered have two main properties: (1) the sectors of the same municipality are mainly mapped to the same cluster, maintaining their adjacency; (2) a cluster is a composition of several municipalities, i.e. a municipality that self-contains its mobility flows does not exist. We believe that these clusters prove that our method is effective, since it does not destroy the original cohesion, and is useful since it suggests a better organization of mobility management, which is different from the organization currently used in a province.

The quality of the resulting clusterings strictly depends on the quality of the mobility network and, hence, on an accurate spatial generalization of trips. In this work we have focused on an existing spatial division provided by the census sector partition, thus with a fixed spatial resolution. An interesting extension of the approach would be to study how spatial resolution and clustering quality are related. We plan to set up a systematic experiment to evaluate the clustering result by varying the spatial generalization resolution. We also plan to emphasize the temporal dimensions of the mobility network. Our aim is to consider the movements in different temporal windows and to map these movements to different OD and SEG mappings. We will thus be able to compare the changes in the clustering map over time intervals, for example, mobility borders generated during weekdays and weekends or even variations within a single day. These two new directions require an objective procedure to state the quality of the clustering. Thus, it is necessary to define new measures to compare the clustering results obtained by using different spatial and temporal resolutions and different community discovery algorithms.

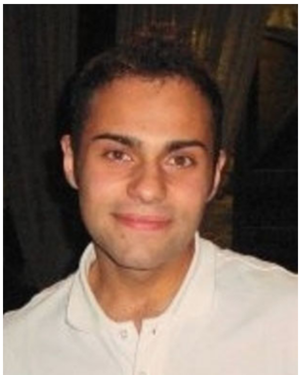
Acknowledgements The authors wish to thank Alessandro Grossi and Michele Berlingerio for their technical support. We also acknowledge Octo Telematics S.p.A. for providing the datasets. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 270833.

References

1. Aiello W, Chung F, Lu L (2000) A random graph model for massive graphs. In: STOC. ACM, New York, pp 171–180
2. Ankerst M, Breunig MM, Kriegel HP, Sander J (1999) Optics: Ordering points to identify the clustering structure. In: SIGMOD, pp 49–60
3. Benevenuto F, Rodrigues T, Cha M, Almeida VAF (2009) Characterizing user behavior in online social networks. In: Internet measurement conference, pp 49–62
4. Bringmann B, Berlingerio M, Bonchi F, Gionis A (2010) Learning and predicting the evolution of social networks. IEEE Intell Syst 25:26–35
5. De Castro R, Grossman JW (1999) Famous trails to Paul Erdös. Math Intell 21:51–63
6. Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E, Stat Nonlinear Soft Matter Phys 70:066111
7. Cook DJ, Crandall AS, Singla G, Thomas B (2010) Detection of social interaction in smart spaces. Cybern Syst 41(2):90–104
8. Coscia M, Giannotti F, Pedreschi D (2011) A classification for community discovery methods in complex networks. Stat Anal Data Min 4(5):512–546
9. Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: SIGKDD. AAAI Press, Menlo Park, pp 226–231
10. Fortunato S (2010) Community detection in graphs. Phys Rep 486:75–174
11. Fortunato S, Barthélemy M (2007) Resolution limit in community detection. Proc Natl Acad Sci USA 104(1):36–41
12. Giannotti F, Nanni M, Pinelli F, Pedreschi D (2007) Trajectory pattern mining. In: SIGKDD, pp 330–339
13. Gomez-Rodriguez M, Leskovec J, Krause A (2010) Inferring networks of diffusion and influence. In: SIGKDD, pp 1019–1028
14. Guimera R, Nunes Amaral KA (2005) Functional cartography of complex metabolic networks. Nature 433(7028):895–900
15. Hecker D, Körner C, Stange H, Schulz D, May M (2011) Modeling micro-movement variability in mobility studies. In: Advancing geoinformation science for a changing world, LNG&C, vol 1, pp 121–140
16. Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York
17. Lancichinetti A, Fortunato S (2009) Community detection algorithms: a comparative analysis. Phys Rev E, Stat Nonlinear Soft Matter Phys 80:5
18. Monreale A, Pinelli F, Trasarti R, Giannotti F (2009) Wherenext: a location predictor on trajectory pattern mining. In: KDD, pp 637–646
19. Nowell DL, Kleinberg J (2003) The link prediction problem for social networks. In: CIKM, pp 556–559

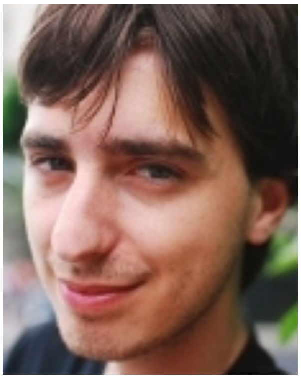
20. Page L, Brin S, Motwani R, Winograd T (1998) The pagerank citation ranking: bringing order to the web
21. Ratti C, Sobolevsky S, Calabrese F, Andris C, Reades J, Martino M, Claxton R, Strogatz SH (2010) Redrawing the map of great Britain from a network of human interactions. PLoS ONE 5(12):5:e14248
22. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105:1118–1123
23. Thiemann C, Theis F, Grady D, Brune R, Brockmann D (2010) The structure of borders in a small world. PLoS ONE 5:11
24. Wang D, Pedreschi D, Song C, Giannotti F, Barabasi AL (2011) Human mobility, social ties, and link prediction. In: SIGKDD, pp 1100–1108
25. Yan X, Han J (2002) gspan: Graph-based substructure pattern mining. In: ICDM.
26. Yan Z, Chakraborty D, Parent C, Spaccapietra S, Aberer K (2011) Semitri: a framework for semantic annotation of heterogeneous trajectories. In: EDBT/ICDT, pp 259–270
27. Yang J, Leskovec J (2010) Modeling information diffusion in implicit networks. In: ICDM, pp 599–608

Salvatore Rinzivillo is a researcher at the Information Science and Technology Institute of the National Research Council at Pisa, Italy. His main research interests are data mining and knowledge discovery in spatial and mobility databases.



Simone Mainardi is a Ph.D. student at the Department of Information Engineering (IET) of University of Pisa and at the Institute for Informatics and Telematics (IIT-CNR) in Pisa, Italy. His research deals with parallel and distributed algorithms, complex network analysis and network evolutionary models.

Fabio Pezzoni is a Ph.D. student at the Institute for Informatics and Telematics (IIT-CNR) in Pisa, Italy. He has a Master degree in Computer Science from the University of Padua, Italy in 2010. His research interests include social network modeling and opportunistic networks.



Michele Coscia is a Ph.D. student at the University of Pisa and a member of the KDDLab. He is also a Google Fellow in Computational Social Science. His interests include representing complex phenomena of the real world as multidimensional networks and studying them using data mining approaches.

Dino Pedreschi is a full professor of Computer Science at the University of Pisa. His current research interests are in data mining and logic in databases, and particularly in data analysis, in spatio-temporal data mining, and in privacy-preserving data mining.

Fosca Giannotti is a senior researcher at the Information Science and Technology Institute of the National Research Council at Pisa, Italy. Her research interests include data mining query languages, mobility data mining, privacy preserving data mining, and complex network analysis.