Noname manuscript No.
(will be inserted by the editor)

Defining the borders of mobility

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

the date of receipt and acceptance should be inserted later

Abstract The availability of massive network and mobility data from diverse domains has fostered the analysis of human behaviors and interactions. Wide, 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 behaviors over a territory to evaluate how the actual administrative borders represent the real basin of human movements. Starting from the a real life dataset of GPS tracked vehicles, we extract a network representation of the movements of the persons, in order to apply a community discovery algorithm to extract relevant clusters, which are then mapped to the geography. We present an extensive experimental settings to prove the quality of our approach.

1 Introduction and Related Work

In recent years, the analysis of human behaviors has been receiving increasing attention by the scientific community, also due to the availability of massive network and mobility data from diverse domains, and the outbreak of novel analytical paradigms, which pose relations among people, or their mobility pattern, at the center of investigation. Inspired by real-world scenarios such as social networks [1,5], human mobility [12], the interplay between the two [23], and so on, in the last years two years, wide, multidisciplinary, and extensive research has been devoted to the extraction of non trivial knowledge from network and mobility data. Predicting future links among the actors of a network ([18,4]), detecting and studying the diffusion of information among them ([13,26]), mining frequent patterns of users' behaviors ([3,24,7]), predicting human mobility patterns ([17]), are only a few examples of problems studied in these scenarios, that includes, among all, physicians, mathematicians, computer scientists, and sociologists.

Recently, the efforts of a few works on social and mobility behaviors was devoted on understanding how people move, and where are their prevalent mobility patterns located. Thiemann *et al.* [22] analyze the human mobility network extracted from the logs provided by the project *Where's George?*[1]: using a stochastic method, they extract a partition of regions according to a fitness function based on modularity maximization. The experiments are performed on a large scale setting, where the minimum spatial granularity is given by a zip code area in the United States. The approach of Ratti *et al.* [20] also adopts the modularity function as objective function to delineate borders emerging from the network extracted from a large database of telecommunication records. However, it is well known in literature that modularity present a resolution problem, i.e. the communities identified via modularity maximization tend to be large and smaller communities are ignored and clustered together [11]. In our case, we need a higher granularity resolution to obtain meaningful results, since we are working with smaller areas than the one used by Thiemann *et al.* and Ratti *et al.*, therefore we use another state-of-the-art community discovery algorithm, namely Infomap [21], that has been show to be better performing than any modularity maximization algorithm [16].

In this context, we want to study the problem of detecting the *borders* of human mobility patterns, com-

Address(es) of author(s) should be given

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

paring them with the *administrative borders* of the cities, provinces, regions, and so on. Do people move and interact within specific areas? Are those areas bounded somehow? Do these bound 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 the questions above, we apply Social Network Analysis techniques to mobility data, with the aim of reaching a better understanding of human mobility patterns, in a new fashion, based not on the interaction of humans themselves, but rather on the underlying, hidden connections that resides among different places. In order to do so, we apply Community Discovery algorithms to the network of geographic areas (i.e., where each node represents a cell or region of movements), with the aim of finding areas that are densely connected by the visits of different users.

The main contribution of the paper consist in the extraction of a fine-grained mobility network to model human behavior and the use of a state-of-the-art community discovery to discover relevant communities corresponding to geographical areas. Moreover, we provide several experiments based on a real-life scenario of GPS tracked vehicles.

The remainder of the paper is organized as follows. In Section 2 we present a general method to extract 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.

2 Mapping Mobility to Complex Networks

The objective of our analysis is to determine the influence of social behavior over the territory, in particular to evaluate how the actual 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. To pursuit such analysis, we propose a general framework based on the following steps: *(1)* the territory is partitioned by means of a non-overlapping spatial grid, whose regions will serve as spatial references; *(2)* the movements are generalized to the spatial grid; *(3)* then they are coded by means of direct weighted graph; and then *(4)* the graph is analyzed to extract the communities within.

A spatial grid serves as the basic level of details to represent the movements. The spatial granularity of the grids strictly depends on the precision of the available data. The movement of people can be tracked using different technologies like, GPS devices, GSM network logs, Wi-Fi fingerprints, RFID tag readings, and so on. 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 extends of each cell. On the contrary, GPS based locations are so precise that is very unlike that two different positions share the same coordinates. For this reason, it is useful to generalize each point to a spatial area, either using existing spatial coverages, like cadastral data, census sectors, or cellular network coverages, or by aggregating together similar points by means of convex hulls, buffers, or clustering [9, 2, 15].

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 is used, it is very likely that the movement is described in term of a pair of cells: a first cell where the call began, and a second cell where the call ended [20]. In rare cases, it is possible to follow the devices moving in the network on the base of the cells crosses. In general, this sampling frequency issue is present also for 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 minimize the quantity of data exchanged, the sampling frequency is determined according to the application scenario. We consider here two different approach to represent movement: on one hand we consider each movement as a pair of *origin* and *destination*; on the other hand we maintain the detailed information, according to the capabilities of the collection device used, about the route followed to move between the two locations.

In the first case, the daily movements of a person can be transformed into a sequence of visited places annotated with the corresponding temporal information. This type of representation provide a precise vision of movements dynamics and, at the same time, allows the handling of the data at a large scale. Moreover, the emphasis of the data is posed on *where* the people move rather then *how* they reach their destinations. Thus, given a trip of a user, the origin of the trip is mapped to the corresponding region in the spatial grid, i.e. the region that contains the first point, as well as the destination of the movement. We call this mapping strategy *Origin-Destination mapping*.

In the second case, we map the entire path on the spatial grid. Depending on the technology used to log the movement, the continuos path is often approximated with a sequence of sampled time-referenced observation. In this case, the mapping to the spatial grid is performed by mapping each sampled point to the corresponding cell in the grid. Since the finer granularity consists of segments of consecutive moves, we will refer to this mapping strategy as *Segments mapping*.

Once each position has been generalized according to the spatial grid, 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 composed by areas with a dense exchange of travelers between them and a low exchange of travelers among this set of areas can then be reduced to the problem of finding clusters of nodes internally densely connected and sparsely connected with the rest of the network. This last formulation is exactly the most popular problem definition of many community discovery algorithms [10, 8].

3 Identifying Clustered Structure

Community discovery algorithms can provide results with many properties. The one we are particularly interested in is the possibility of having hierarchical results. This means that the algorithm is not returning a simple flat partition, but we can actually navigate up and down in the hierarchy to tune the granularity of the communities. One algorithm able to satisfy this property is Infomap [21].

The Infomap algorithm is based on a combination of information-theoretic techniques and random walks. Authors want to explore the graph structure with a number of random walks of a given length and with a given probability of jumping to a random node. This approach is equivalent to the random surfer of the PageRank algorithm [19]. Intuitively, the random walkers are trapped into a community and exit from it very rarely. Thus, if we have a division in communities, we can efficiently describe these random walks as a series of intra community steps followed by an inter community jumps. The formal equation described by these concepts is the following:

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

where L is the lower bound for the number of bits needed in the description of the nodes of the network, M is the community partition, q is the probability that the random walk jumps from a community to another on any given step, $H(Q)$ is the entropy of the description of the community, m is the number of communities in the network, p_i is the fraction of within-community movements that occur in community i and $H(P_i)$ is the entropy of the within-community movements, including the exit code for community i.

Trying any possible community partition in order to minimize $L(M)$ is inefficient and intractable. Authors narrow the space of the candidate partitions with several iteration of a greedy modularity community discoverer [6]. They then refine the partition of the graph with simulated annealing [14]. This optimization is lead by the information-theoretic principle of reducing the number of bits needed to encode the information of the structure (i.e. they minimize $L(M)$). This is done by assigning an Huffman coding to the nodes of the network.

The description of the nodes of the network is divided into two levels. Authors retain unique names for large-scale objects, the communities identified within the network in the first step, and they reuse the names associated with fine-grain details, the individual nodes within each community. This two-level description allows to describe the path in fewer bits, relying on the fact that a random walker is statistically likely to spend long periods of time within certain clusters of nodes.

We chose to use Infomap algorithm not only for its ability to return both flat and hierarchical partition. Infomap, in fact, is one of the best performing non overlapping community discovery algorithms, as studied in [16]. Infomap was tested against the benchmark by Girvan and Newman and on random graphs. As a result Infomap has been judged to have an excellent performance, with the additional advantage of low computational complexity.

4 Experiments and Discussion

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

Nevertheless, for each vehicle the server has only a sequence of received points, without any semantic annotation. Thus, it is necessary to partition that sequence into sub-sequences that represent each single journey. We based on 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 [25].

We observed approximately 38 thousands vehicles for a period of 5 weeks (from June 14 to July 30, 2010). The frequency of the time sampling enabled us to explore different temporal resolutions when generalizing the data to a given spatial grid. As presented in Section 2, we adopted two different strategies to generalize the timestamped locations. As a first approach, we used the Origin-Destination (OD) mapping to simplify each trip by considering only the first and the last point. Secondly, we used the Segment (SEG) mapping to generalize each timestamped point of a trajectory to the spatial grid.

For this analysis we adopted a spatial grid based on existing census sectors, as provided by the ISTAT, the Italian National Bureau of Statistics. The motivations are manifold: this coverage is publicly available and contains many statistical information (e.g. population, commuters, segmentation by age, etc.); 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.

Census sectors can be aggregated according to a four level hierarchy: the base level contains the census sectors, where each area corresponds approximatively to a city block. Several adjacent sectors form a *Comune* — which is the italian term for a town or a city. Several adjacent *Comune* form a *Provincia*. An aggregation of adjacent *Provincia* determines a *Region*. In the following, we will use the english words province and region when referring to a *Provincia* and a *Regione*, respectively.

The census sector level is used for the generalization according to 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 respectively with v_R and v_S, and its weight 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 exist at least a trajectory of a vehicle 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 to the *town* level.

	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

Table 1 Features of the OD and SEG mapping graphs.

Table 1 shows some features of the OD and the SEG mapping. Although the census sectors we considered do not change from one mapping to another SEG has about 300 nodes more 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 an highway. Many vehicles will pass through these sectors when travelling on the highway, regardless of their source (destination). Despite this, only one edge linking these two highway sectors could exists. Indeed, information on the number of journeys passing through these two sectors can still be read from the weight associated to 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 average edge weight indicates that two sectors are the source (destination) of about 3 journeys on average. If we keep an eye on the average node weight in the SEG mapping, we can note that each census sector is reached and/or leaved 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 in an increment of the weight associated to in- and out-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 deeper levels of the hierarchy (e.g. 513 at the second level).

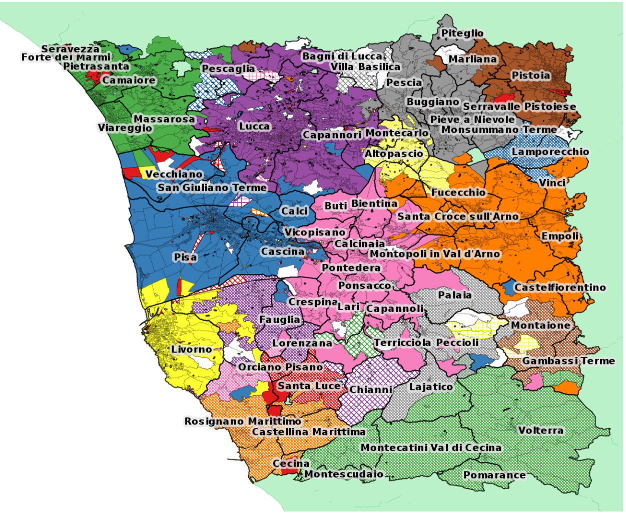


Fig. 1 Clusters in the area of study. Administrative borders of *Comuni* are drawn with thick black curves and sectors borders with gray ones. Colors and textures have been used in order to visually represent the 36 clusters with the highest PageRank values. The remainder of the clusters have been left white. Clusters with the highest PageRank values have no texture and are filled with solid colors.

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

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.

Figure 1 shows the resulting level-1 clusters. Among these 96 clusters, we select 19 of them with a PageRank value greater than 0.1% and in particular 8 of them have a PageRank value greater than 5% . Thus, the majority of the journeys involve very few clusters — a journey has the 98.13% chance to begin (end) 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 — they contain 7,527 census sectors, i.e. the 95.54% of the total. Furthermore, they are composed of *geographically adjacent* census sectors, despite the OD mapping contains many connections between non-adjacent areas.

In Tab. 2 we show clusters with a PageRank value greater than 5%. These clusters are named according to the largest *Comune* they contain. In the following and when not ambiguous, we will always refer to each cluster by that name.

To validate our results, we disuss now the main clusters by means of the background knowledge about the interested areas, starting from the *Pisa* cluster, which is highlighted with a dark blue color fill in Fig. 1. This cluster is composed by the majority of the statistical sectors of the city of Pisa plus the sectors of its adjacent towns, in particular Cascina, Calci, San Giuliano Terme and Vecchiano. Traditionally, these towns are referred as *"Area Pisana"*[2], which can be considered as an enlarged metropolitan area centered in Pisa. Recently, the Regional government has promoted a strategic development project for this area, named *"Piano Strategico dell'Area Pisana"*, which involves these five towns with the objective of designing an integrated mobility plan for the five municipalities.

The other highest-PageRank clusters can also be interpreted by means of well-known geographical and socio-demographic features. The reasons behind those relations are due both to historical relationship and to the morphology of the territory. For example, the cluster of Viareggio, located in the north-west and filled with 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 very well the strength of the socio-economic relations between geographical areas.

It is worth noting how the cohesion of sectors within the same *Comune* is maintained after the clustering, even with 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 sector, in fact, correspond to the beaches and they are a frequent destination for people from Livorno during the summer period. On the contrary, the main destinations for the seaside for people in Pisa is the west region of the city, adjacent to the estuary of Arno and the beaches in Vecchiano.

Finally, it is important to note that it is not a necessary condition for a cluster to be composed of geographically adjacent sectors. This is a direct consequence of the human mobility which we can conclude is short-range in the territory considered. On the contrary, if the mobility were been long-range, clusters would have been made up of distant sectors and the coloring of Fig. 1 would have been much less uniform.

[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

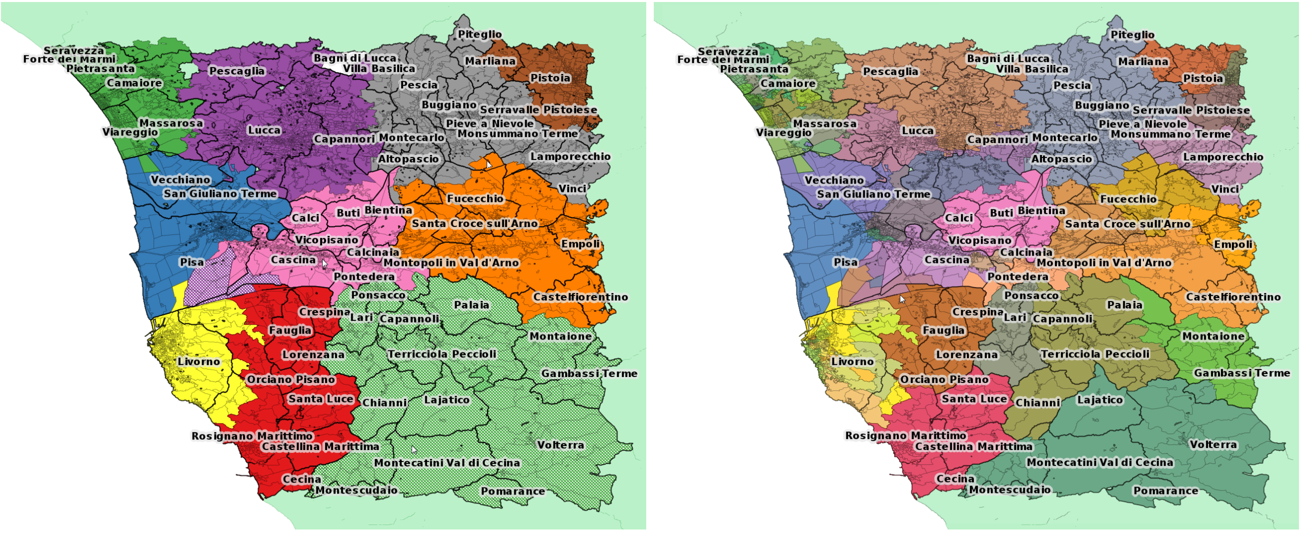


Fig. 2 Clusters in the north-western patch of the *Regione* Toscana obtained from the SEG mapping, at level 1 and 2. *Comuni* borders are drawn with thick black curves and sectors borders with narrower ones. Colors and textures have been used in order to visually represent the clusters. Clusters with the highest PageRank values have no texture and are filled with solid colors. The rest of the clusters have textures reflecting their PageRank values: the higher the tightness of a texture, the higher the PageRank. *(Left)* Level-1 Clusters. *(Right)* Level-2 Clusters

Cluster	PageRank %
Lucca	17.42
Pisa	13.67
Livorno	12.57
Pontedera	11.54
Volterra	11.39
Montecatini Terme	9.32
Empoli	8.91
Pistoia	6.20
Fauglia	4.33
Viareggio	3.86
Pisa (south east)	0.68

Table 3 Level-1 Clusters in the SEG mapping. Clusters are indicated with the name of a *Comune* they encompass.

4.2 Segments Mapping

The clustering method for the SEG mapping produced a 5-level hierarchy of clusters. At the level 1 there are 11 clusters, which are shown in Fig. 4.2 *(Left)*. A this level, the number of clusters is significantly less than in the OD mapping. Hence, the clustering method better aggregates census sectors. Nevertheless, this is reasonable since only geographically adjacent sectors can be connected in SEG. Moreover, their PageRank, which is reported in Tab. 3, never assume values less than 0.6%, whereas in the OD mapping there are 77 clusters whose PageRank is less than 0.1%. Differently from the OD mapping, in SEG cluster coverages have an interesting and meaningful size also at level 2. The 78 clusters identified at the level 2 are shown in Fig. 4.2 *(Right)*.

The clusters of the level 1 of the hierarchy can be compared with the biggest clusters obtained with the OD mapping. The clusters of Viareggio, Pistoia, Lucca, Livorno and Empoli maintain approximatively the same geographical extension. On the contrary, the clusters of Montecatini Terme and Volterra are bigger, encompassing geographical areas which, in OD, are considered as different clusters. The clusters of Pisa and Pontedera are significantly different in respect with the OD mapping because the *Comuni* of Cascina and Calci belong to the cluster of Pontedera. The cluster with Fauglia covers the geographical area known as "Colline Livornesi" [6].

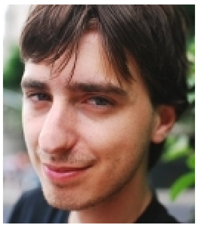
5 Conclusions and Future Work

In this paper we have presented a general method to discover geographical areas determined by mobility behaviors of people. The method is based on the extraction of a multi-scale mobility network, representing the flows of movement between a set of regions. The network is analyzed by means of one of the best performing community discovery algorithms among the non-overlapping ones. We presented an extensive experimental setting where the results are discussed and commented with reference to the domain knowledge of the territory. In the future, we plan to emphasize the temporal dimension of the mobility network by providing specific temporal projection during each single day or week.

[6] http://it.wikipedia.org/wiki/Colline_Livornesi

References

1. W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *STOC*, pages 171–180. ACM, 2000.
2. M. Ankerst, M. M. Breunig, H.P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *SIGMOD*, pages 49–60, 1999.
3. F. Benevenuto, T. Rodrigues, M. Cha, and V.A.F. Almeida. Characterizing user behavior in online social networks. In *Internet Measurement Conference*, pages 49–62, 2009.
4. B. Bringmann, M. Berlingerio, F. Bonchi, and A. Gionis. Learning and predicting the evolution of social networks. volume 25, pages 26–35, 2010.
5. R. De Castro and J.W. Grossman. Famous trails to paul erds. *Mathematical Intelligencer*, 21:51–63, 1999.
6. Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.
7. D.J. Cook, A.S. Crandall, G. Singla, and B. Thomas. Detection of social interaction in smart spaces. *Cybernetics and Systems*, 41(2):90–104, 2010.
8. Michele Coscia, Fosca Giannotti, and Dino Pedreschi. A Classification for Community Discovery Methods in Complex Networks. *Statistical Analysis and Mining*, 2011.
9. Martin Ester, Hans-Peter Kriegel, Joerg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad, editors, *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.
10. S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, February 2010.
11. Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, January 2007.
12. Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 330–339, New York, NY, USA, 2007. ACM.
13. Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *KDD*, pages 1019–1028, 2010.
14. Roger Guimera and Luis A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.
15. L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, 1990.
16. A. Lancichinetti and S. Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5):056117–+, November 2009.
17. Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. Wherenext: a location predictor on trajectory pattern mining. In John F. Elder IV, Françoise Fogelman-Soulié, Peter A. Flach, and Mohammed Javeed Zaki, editors, *KDD*, pages 637–646. ACM, 2009.
18. D.L. Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03*, pages 556–559. ACM, 2003.
19. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1998.
20. Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H. Strogatz. Redrawing the map of great britain from a network of human interactions. *PLoS ONE*, 5(12):e14248, 12 2010.
21. M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Science*, 105:1118–1123, January 2008.
22. Christian Thiemann, Fabian Theis, Daniel Grady, Rafael Brune, and Dirk Brockmann. The structure of borders in a small world. *PLoS ONE*, 5(11):e15422, 11 2010.
23. Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1100–1108, New York, NY, USA, 2011. ACM.
24. X. Yan and J. Han. gspan: Graph-based substructure pattern mining. ICDM '02, pages 721–, 2002.
25. Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. Semitri: a framework for semantic annotation of heterogeneous trajectories. In *Proceedings of the 14th International Conference on Extending Database Technology*, EDBT/ICDT '11, pages 259–270, New York, NY, USA, 2011. ACM.
26. Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In Geoffrey I. Webb, Bing Liu, Chengqi Zhang, Dimitrios Gunopulos, and Xindong Wu, editors, *ICDM*, pages 599–608. IEEE Computer Society, 2010.



Michele Coscia is a PhD student at the Computer Science Department of the University of Pisa and a member of the Knowledge Discovery and Data Mining Laboratory (KDDLab), a joint research group with the Information Science and Technology Institute of the National Research Council in Pisa. He is a Google Fellow in Computational Social Science, and visiting student at the Barabasi Lab in the Center for Complex Network Research, Boston. His interests include representing complex phenomena of the real world as multidimensional networks and studying them using data mining approaches.

Fosca Giannotti is a senior researcher at the Information Science and Technology Institute of the National Research Council at Pisa, Italy, where she leads the Knowledge Discovery and Data Mining Laboratory (KDD-Lab), a joint research initiative with the University of Pisa. Her recent research interests include data mining query languages, mining spatio-temporal and mobility data, privacy preserving data mining, and complex network analysis. She has been the coordinator of various European-wide research projects, including GeoPKDD: Geographic Privacy-aware Knowledge Discovery and Delivery. She is the author of more than one hundred publications and served as PC chair and PC member in the main conferences on Databases and Data Mining. She is the co-editor of the book ”Mobility, Data Mining and Privacy”, Springer, 2008.

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. He is a member of the program committee of the main international conferences on data mining and knowledge discovery and an associate editor of the journal Knowledge and Information Systems. He served as the coordinator of the undergraduate studies in Computer Science at the University of Pisa, and as a vice-rector of the same university, with responsibility in teaching affairs. He has been granted a Google Research Award (2009) for his research on privacy-preserving data mining and anonymity-preserving data publishing. He is the co-editor of the book ”Mobility, Data Mining and Privacy”, Springer, 2008.