GeoPKDD
Geographic Privacy-aware Knowledge Discovery

Fosca Giannotti[2], Mirco Nanni[2], Dino Pedreschi[1], and Chiara Renso[2]

[1] Pisa KDD Laboratory, Computer Science Department, University of Pisa, Italy
e-mail: {pedre}@di.unipi.it
[2] Pisa KDD Laboratory, ISTI - CNR, Pisa, Italy
e-mail: { fosca.giannotti, mirco.nanni, chiara.renso}@isti.cnr.it

Abstract. A flood of data pertinent to moving objects is available today, and will be more in the near future, particularly due to the automated collection of privacy-sensitive telecom data from mobile phones and other location-aware devices. Such wealth of data, referenced both in space and time, may enable novel classes of applications of high societal and economic impact, provided that the discovery of consumable and concise knowledge out of these raw data is made possible. The goal of the *GeoPKDD* project is to develop theory, techniques and systems for geographic knowledge discovery and delivery, based on new privacy-preserving methods for extracting knowledge from large amounts of raw data referenced in space and time. More precisely, we aim at devising knowledge discovery and analysis methods for trajectories of moving objects; such methods will be designed to preserve the privacy of the source sensitive data. The fundamental hypothesis is that it is possible, in principle, to aid citizens in their mobile activities by analysing the traces of their past activities by means of data mining techniques. For instance, behavioural patterns derived from mobile trajectories may allow inducing traffic flow information, capable to help people travel efficiently, to help public administrations in traffic-related decision making for sustainable mobility and security management, as well as to help mobile operators in optimising bandwidth and power allocation on the network. However, it is clear that the use of personal sensitive data arouses concerns about citizen's privacy rights. Obtaining the potential benefits by means of a trustable technology, designed to prevent infringing privacy rights, is a highly innovative goal; if fulfilled, it would enable a wider social acceptance of many new services of public utility that would find in the advocated form of geographic knowledge a key driver, such as in transport, environment and risk management.

1 The GeoPKDD Vision

In this paper, we present the vision underlying the GeoPKDD project – Geographic Privacy-aware Knowledge Discovery and Delivery, project number 01495 within the Future Emerging Technologies program of FP6-IST.

The general goal of the GeoPKDD project is to develop theory, techniques and systems for knowledge discovery and delivery, based on new automated privacy-preserving methods for extracting user-consumable forms of knowledge from large amounts of raw data referenced in space and in time. Particular emphasis is placed upon:

- Devising methods for representing, storing and managing moving objects, which change their position in time, and possibly also their shape or other features, together with their trajectories, with varying levels of accuracy and certainty;
- devising spatio-temporal knowledge discovery and data mining methods and algorithms for moving objects and their trajectories;
- devising native techniques to make such methods and algorithms intrinsically privacy-preserving, as data sources typically contain personal location-aware sensitive data.

The motivations for undertaking this direction of research are rooted in the consideration that spatio-temporal, geo-referenced datasets are, and will be, growing rapidly, due to, in particular, the collection of privacy-sensitive telecommunication data from mobile phones and other location-aware devices, as well as the daily collection of transaction data through database systems, network traffic controllers, web servers, sensors.

The large availability of these forms of geo-referenced information is expected to enable novel classes of applications, where the discovery of consumable, concise, and applicable knowledge is the key step. As a distinguishing example, the presence of a large number of location-aware wirelessly connected mobile devices presents a growing possibility to access space-time trajectories of these personal devices and their human companions: trajectories are indeed the traces of moving objects and individuals. These mobile trajectories contain detailed information about personal and vehicular mobile behaviour, and therefore offer interesting practical opportunities to find behavioural patterns, to be used for instance in traffic and sustainable mobility management, e.g., to study the accessibility to services. Clearly, in these applications privacy is a concern. In particular, how can trajectories of mobile individuals be stored and analysed without infringing personal privacy rights and expectations? How can, out of privacy-sensitive trajectory data, patterns be extracted that are demonstrably privacy-preserving, i.e., patterns that do not disclose individuals' sensitive information? These questions, which call for answers at a combined technical, legal and social level, address a crucial point, both from the ethical point of view and that of social acceptance - solutions that are not fully trustworthy will find insuperable obstacles to their deployment. On the other hand, demonstrably trustworthy solutions may open up tremendous opportunities for new knowledge-based applications of public utility and large societal and economic impact. As a prototypical application scenario, assume that source data are log transactions from mobile cellular phones, reporting user's movements among the cells in the network; these are streams of raw data (log entries) about users entering a cell – *(userID, time, cellID)* – or, in the near future, even user's position within a cell – *(userID, time, cellID, X, Y)* and, in the case of GPS/Galileo equipped devices, user's absolute position. Indeed, each time a mobile phone is used on a given network, the phone company records real-time data about it, including time and cell location. If a call is taking place, the recording data-rate may be higher. Note that if the caller is moving, the call transfers seamlessly from one cell to the next. In this context, a novel geographic knowledge discovery process may be envisaged, composed of three main steps: trajectories reconstruction, knowledge extraction, and delivery of the information obtained, described in the following (see Figure 1).

Trajectory reconstruction: in this basic phase, a stream of raw data about moving people has to be processed to obtain ready-to-use trajectories, building a privacy-aware

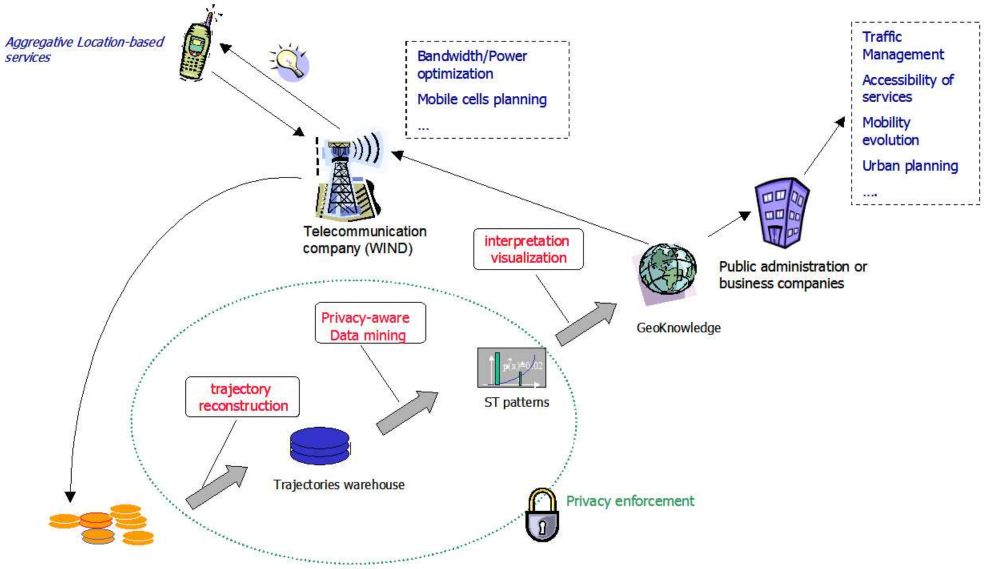
Aggregative Location-based services
Bandwidth/Power optimization
Mobile cells planning
...
Traffic Management
Accessibility of services
Mobility evolution
Urban planning
….
Telecommunication company (WIND)
interpretation visualization
Public administration or business companies
GeoKnowledge
Privacy-aware Data mining
trajectory reconstruction
p(x)=0.02
ST patterns
Trajectories warehouse
Privacy enforcement

Fig. 1. The GeoPKDD process

trajectories warehouse. Reconstruction of trajectories is per se a challenging problem. The reconstruction accuracy of trajectories, as well as their level of spatio-temporal granularity, depend on the quality of the log entries, since the precision of the position may range from the granularity of a cell of varying size to the relative (approximated) position within a cell. Indeed, each moving object trajectory is typically represented as a set of localization points of the tracked device, called sampling. This representation has intrinsic imperfection due to mainly two aspects. The first source of imperfection is the error measurement of the tracking device. As an example, a GPS-enabled device introduces an error measurement of some meter, whereas the imprecision introduced in a GSM/UMTS network is the dimension of a cell, that could be of some kilometer. In addition, the second source of imperfection is related to the sampling rate and involves the trajectory reconstruction process that approximates the movement of the objects between two localization points. Although for some application linear interpolation can be an acceptable approximation of the real trajectory, we believe that this could be a too coarse approximation and more sophisticated techniques are to be investigated to take into account the spatial, and possibly temporal, imperfection in the reconstruction process. This whole process is greatly dependant on the privacy concerns which must be addressed in manipulating data, ranging from a context where the full raw data are available, which allows to reconstruct individual trajectories at different spatial granularity (at cell level, or more accurate positions), to the opposite one, where user IDs are obfuscated, limiting the reconstruction accuracy to an aggregated traffic flow information. Efficient and effective storage of trajectories into ad hoc privacy-aware warehouses should also be devised, capable of dealing with continuous incoming streams of raw log data, and equipped with suitable access methods to support analysis and mining tasks. Such access methods, moreover, should be controlled by mechanisms that prevent the disclosure of sensible data, both explicitly (e.g., providing users' identity) and implicitly (providing non-sensible data from which sensible information can be inferred). In this context, all the privacy issues on traditional databases, which have been actively studied in recent years by the research community, will play an important role, together with new problems that arise from the peculiarities of trajectories data.

Knowledge extraction: spatio-temporal data mining methods must be developed to extract useful patterns out of trajectories. However, spatio-temporal data mining is still in its infancy, and even the most basic questions in this field are still largely unanswered: what kinds of patterns can be extracted from trajectories? Which methods and algorithms should be applied to extract them? How can such patterns be effectively used to improve the comprehension of the application domain and to deliver better services? The following basic examples give a glimpse of the wide variety of patterns and possible applications it is expected to manage :

- *Clustering*, the discovery of groups of "similar" trajectories, together with a summary of each group (see Figure 2). Knowing which are the main routes (represented by clusters) followed by people during the day can represent a precious information for improving several different services to citizens. E.g., trajectory clusters may highlight the presence of important routes not adequately covered by the public transportation service.

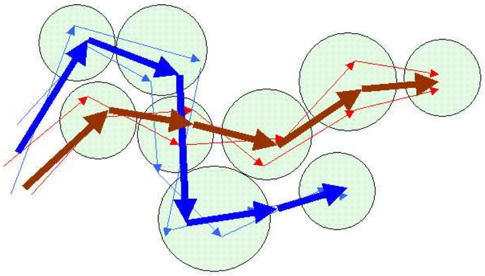

Fig. 2. Example of clusters in traffic data

- *Frequent patterns*, the discovery of frequently followed (sub)-paths (see Figure 3). Such information can be useful in urban planning, e.g., by spotlighting frequently followed inefficient vehicle paths, which can be the result of a mistake in the road planning.
- *Classification*, the discovery of behaviour rules, aimed at explaining the behaviour of current users and predicting that of future ones (see Figure 4). Urban traffic

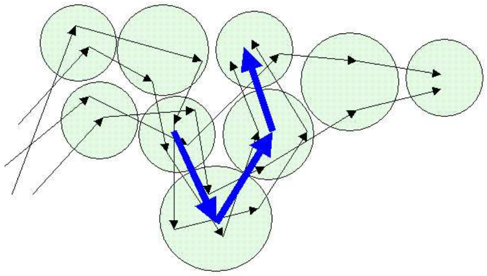

Fig. 3. Example of frequent patterns in traffic data

simulations are a straightforward example application for this kind of knowledge, since a classification model can represent a sophisticated alternative to the simple ad hoc behaviour rules, provided by domain experts, on which actual simulators are based.

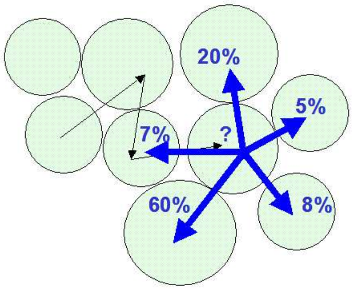

Fig. 4. Example of classification in traffic data

Spatio-temporal data introduce new possibilities and, correspondently, novel issues in performing these tasks. Clustering moving object trajectories, for example, requires to find out both a proper spatial granularity level (strongly dependant on the application) and a significant temporal sub-domain (e.g., rush hours might be informative for defining a clustering structure over traffic data, while other time periods might simply add noise to the clustering process). Whichever kind of pattern it is extracted, if no control procedure is applied there will always be the risk of extracting patterns that violate in some extent the privacy of an individual or a community. As an example, if an extremely homogeneous cluster is obtained, knowing its ”summary” (e.g., a representative trajectory) can be sufficient to reconstruct information on the individuals it contains (e.g., their trajectories would coincide almost perfectly with the representative trajectory of the cluster). In the same way, an overfitted classifier could contain rules generated on the base of only one or very few individuals, thus disclosing some (potentially sensible) information on their characteristics. Therefore, the generated patterns should be precisely characterized as privacy-preserving or not, in the sense that a formal set of constraints on spatio-temporal data and patterns should be available to express privacy requirements, along with data mining algorithms that demonstrably yield only privacy-preserving patterns - e.g., patterns that do not disclose any sensitive information of the individual trajectories they come from. Other flavours of privacy-preserving methods, such as data perturbation and multi-party secure computation should also be taken into account, whenever appropriate to deal with different privacy requirements.

Knowledge delivery: extracted patterns are very seldom geographic knowledge prêt-a-porter: it is necessary to reason on patterns and on pertinent background knowledge, evaluate patterns' interestingness, refer them to geographic information, find out appropriate presentations and visualizations. Once suitable methods for interpreting and delivering geographic knowledge on trajectories are available, several application scenarios become enabled, at different levels (societal, individual, business):

towards the society: Typically, we have here applications for sustainable mobility, such as dynamic traffic monitoring, traffic management, public transportation and urban planning. There are several examples of this kind of applications: identifying places most visited by people, measuring accessibility of services, modelling people decisions on how to move, study the impact of promotional actions, analysis of zonal characteristics for the impact on housing price.

towards the network operator: Here, the analysis of the people movements is exploited by the network operator to support optimization of the network infrastructure. An example could be the design of an adaptive system for dynamic bandwidth and/or power allocation to cells.

towards the individual: In this class of applications the extracted knowledge may be sent back to the mobile user for a form of personalized location-based services (LBS). For example, area traffic report and predictions about expected traffic flows can support the user in route planning.

Each level of applications may require a different interaction schema between mobile devices and a service server: e.g., the first two levels above typically exploit mobile devices for location acquisition, in which only the server needs to know where the device is located, while the third level of applications requires a bi-directional interaction, where the server receives location information but also returns services directly to the mobile device. In particular, focus will be on those applications which can be enabled by aggregative information extracted from positioning data at the server level in a safe, privacy-preserving way, and delivered in the appropriate form to various end users - public administrations, businesses, individuals.

In our vision, the fundamental hypothesis is that it is possible, in principle, to aid humans in their mobile activities, and in related decision making, by deriving new aggregated knowledge from the traces of their past activities. It is now time to undertake this research: soon we will be flooded by this form of spatio-temporal data, and more accurate spatial positioning will become available as GPS- and Galileo-enabled location-aware portable devices and phones get into widespread use.

The GeoPKDD consortium is illustrated in the following list which reports the partners and their key investigators:

- KDD Lab. joint research group of ISTI-CNR, Istituto di Scienza e Tecnologie dell'Informazione, Pisa. www.isti.cnr.it/ and Univ. Pisa, Dept. of Computer Science www.di.unipi.it
  (Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Chiara Renso, Franco Turini)
- Univ. Limburg, Theoretical Computer Science Group. www.luc.ac.be/theocomp
  (Bart Kuijpers, Jan Van den Bussche)
- EPFL, Lab. DB, Lausanne. lbdwww.epfl.ch/e/
  (Stefano Spaccapietra, Christine Parent)
- Fraunhofer Institute for Autonomous Intelligent Systems, Sankt Augustin.
  www.ais.fraunhofer.de/
  (Michael May, Natalia Andrienko, Codrina Lauth)
- Wageningen UR, Centre for GeoInformation ALTERRA.
  cgi.girs.wageningen-ur.nl/
  (Monica Wachowicz)
- Research Academic Computer Technology Institute, Research and Development Division. www.cti.gr/ and Univ. Piraeus, Dept. of Informatics www.unipi.gr
  (Yannis Theodoridis, Vassilios Verykios)
- Sabanci University, Faculty of Engineering and Natural Sciences.
  www.sabanciuniv.edu/
  (Yücel Saygin)
- Wind Progetti Finanziati and Technology Scouting. www.wind.it
  (Riccardo Mazza)

Besides the European level Consortium, there is an Italian Working Group involving researchers from the Universities of Calabria, Milano, Pisa and Venezia, active also on other related important topics, such as distributed data mining, mining data streams, and workflow mining.

Finally, KDD Lab in Pisa, the coordinator of GeoPKDD, concentrated so far on the following topics:

- *Synthetic Generation of Cellular Network Positioning Data*, aimed at providing a system to build benchmark datasets for cellular device positioning data, which typically will not be publicly available for scientific research. We provide a system, called CENTRE (CEllular Network Trajectories Reconstruction Environment [3]), able to randomly generate movement data by modelling different movement behaviors as specified by some user preferences, exploiting user defined topologies and cellular network requirements and referencing such data on a geographic scenario, which provides further constraints and background knowledge.
- *Density-based Clustering of trajectories of moving objects*: In this work [2], we addressed two distinct questions: (i) what is the most adequate clustering method for trajectories, and (ii) how can we exploit the intrinsic semantics of the temporal dimension to improve the quality of trajectory clustering. Both questions can be suitably addressed by generalizing the density-based approach to clustering.

- *Sequential Patterns with Temporal Annotations*: Most approaches to sequence mining focus on sequentiality of data, using time-stamps only to order items and, in some cases, to constrain the temporal gap between items. We propose an extension of the sequence mining paradigm to (temporally-)annotated sequential patterns, where each transition in a sequential pattern is annotated with a typical transition time derived from the source data [4].
- *Anonymity-preserving Data Mining*: It is generally believed that data mining results do not violate the *anonymity* of the individuals recorded in the source database. In fact, data mining models and patterns represent a large number of individuals and thus conceal individual identities: this is the effect of the minimum support threshold in association rule mining. We show that this belief is ill-founded. By shifting the concept of *k-anonymity* from data to patterns, we formally characterize the notion of a threat to anonymity in the context of pattern discovery, and provide a methodology to efficiently and effectively identify all possible such threats that might arise from the disclosure of a set of extracted patterns [1].
- *Classification in Geographical Information Systems*: We explore the application of decision tree learning methods to the classification of spatial datasets, which, according to the Geographic Information System approach, are represented as stacks of layers. We propose an entropy measure, weighted on a specific spatial relation, and describe an application to the classification of geographical areas for agricultural purposes [5].

References

1. M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi. *k*-anonymous patterns. In Proc. ECML/PKDD 2005.
2. M. D'Auria, M. Nanni, and D. Pedreschi. Time-focused density-based clustering of trajectories of moving objects. In Proc. ECML/PKDD 2005 Workshop on Mining Spatio-Temporal Data (MSTD).
3. F. Giannotti, A. Mazzoni, S. Puntoni, and C. Renso. Synthetic generation of cellular network positioning data. ACM GIS'05.
4. F. Giannotti, M. Nanni, D. Pedreschi, and F. Pinelli. Mining sequences with temporal annotations. Submitted.
5. S. Rinzivillo and F. Turini. Classification in geographical information systems. In Proc. ECML/PKDD 2004.