# Dagstuhl Report on Geographic Privacy-Aware Knowledge Discovery and Delivery

Bart Kuijpers (Hasselt University - Diepenbeek, BE)
Dino Pedreschi (University of Pisa, IT)
Yucel Saygin (Sabanci University - Istanbul, TR)
Stefano Spaccapietra (EPFL - Lausanne, CH)

## 1 Introduction

The Dagstuhl-Seminar on Geographic Privacy-Aware Knowledge Discovery and Delivery was held during 16 - 21 November, 2008, with 37 participants registered from various countries from Europe, as well as other parts of the world such as United States, Canada, Argentina, and Brazil. Issues in the newly emerging area of geographic knowledge discovery with a privacy perspective were discussed in a week to consolidate some of the research questions. The Dagstuhl program included plenary sessions and special interest group meetings which continued even late in the evening with heated discussions. The plenary sessions were dedicated for the talks of some of the participants covering a variety of issues in geographic knowledge discovery and delivery. The reports on special interest group meetings (SIG) were also presented and discussed during the plenary sessions.

The topics of the talks presented during the plenary sessions could be summarized as:

- Consolidation of the notion of trajectories
- Semantic aspects of trajectories
- Warehousing of trajectories
- Mining of trajectories
- Applications
- Privacy and legal aspects

The special interest groups were formed to discuss in detail the (1) Definition and Semantics of Trajectories, (2) Applications of Anonymized Geographic Data. After each SIG meeting, a plenary meeting was organized to harmonize the results of the discussions and to share them with other interested researchers. A summary session was held for the SIG meetings to



Dagstuhl Seminar Proceedings 08471
Geographic Privacy-Aware Knowledge Discovery and Delivery
http://drops.dagstuhl.de/opus/volltexte/2009/2010

discuss open questions even basic ones regarding what a trajectory is and how do we define the specific area of knowledge discovery on geographic data, and application areas of privacy in trajectory data publication and analysis.

Abstracts of the talks during the plenary sessions and SIG meetings are provided at the end of this report. The issues discussed during the SIG meetings are summarized in the following subsections.

### 1.1 Privacy SIG: Applications of Anonymized Data

In this working group privacy issues in trajectories were discussed around the popular topic of anonymity. The aim was to find some killer applications for the anonymized data so that anonymization techniques could be adopted and widely used. The questions thrown into the discussion arena were the following:

- What are the Killer Applications?
- What information is needed to support these applications?
- What structure of data best supports these applications?

The psychological as well as the economical aspects of privacy were discussed during the meetings. The relationship between the psychological vs quantifiable privacy risks were questioned in the context of geo-referenced data. The rich background information and other sources which could be linked to the location data were pointed out. The need to move from the needed information content to the anonymization techniques to support it was proposed as the main research direction of anonymization. The Economic aspects of privacy questioning wheter privacy could be treated as a "good" and if we can sell or buy privacy was discussed.

The applications of geo-referenced data were categorized into groups:

- health
- traffic and transportation
- edutainment
- public safety emergency response
- recommender systems

The types of geo-spatial self-published data in the context of the above applications were identified as:

- Geo-referenced media such as photos in Flickr and similar sites which are tagged with spatial information.
- Blogs - including textual location Calendar - event location
- Personal GPS tracks obtained during activities like running, walking, or hiking which could be combined with health data
- SMS-based microblogging like in VANET

The typical uses of the geo-referenced data were listed as:

- Direct geo-marketing
- Social science
- Investigation, public safety, and law enforcement
- Event detection
- Dating
- Recommender systems
- Information dissemination - geocast (instead of broadcast)
- Targeted health information

The possible risks or misuses of the collected georeferenced data were identified as:

- Geospam
- Harassement or stalking
- Profiling: Insurance risks
- Suppressing political dissent
- Archived data - changing standards
- Travel restrictions

- Dissemination of misinformation

The possible precautions against the misuse of the above data were identified as:

- Educating the society about the risks, and demonstrate the inference capabilities.
- Regulations
- Technology for scrubbing,generalization, cloaking for anonymity. Enabling multiple virtual identities and authentication through virtual ID.
- Continuous and dynamic Risk Assessment adaptive to change with increasing information

## 1.2 SIG for Trajectories: Definition and Semantics

The first meeting of this SIG was on the basics of trajectories. Two views of movement which are traffic or trajectory oriented are discussed. An important communication or terminology problem appeared during this first SIG meeting to identify the transition from the data to patterns. Data Mining Query Language (DMQL) was considered to be a transition from the data world to the pattern world. It was stated that trajectory mining needs a language with the capability to express spatio-temporal constraints over sequences and sets of objects.

The second meeting of this SIG was on the semantic aspects. The main question was weather semantics of movement is anything that goes beyond raw spatio-temporal positioning. Multiple Semantic Layers were identified which are

- Movement only, no semantics: sequence of ST points
- Trajectories as meaningful movement segments (meaningful: from the application viewpoint)
- Trajectories as sequences of moves between "places" (geo-places or anything else that changes over time/space)
- Trajectories attached to objects (persons, cars, birds, etc.)

The third SIG meeting was on aggregation and warehousing issues of trajectories. Conceptual models(EER) DW model, spatial hierarchies, spatial OLAP, implementation, temporal issues were discussed.

# 2 Abstracts of the talks during the plenary sessions and SIG meetings

## 2.1 Challenges for Defining and Measuring Location Privacy and Anonymity by Christopher W. Clifton , Purdue University

Privacy laws, including the EU guidelines for location-based services, generally do not apply to anonymous data. This leaves anonymized data open to a variety of (beneficial) uses, without privacy restriction. Unfortunately, "made anonymous" (in the words of 2002/58/EC, which explicitly discusses location-based services) is not as clear as it sounds. I will point out some of the issues and challenges, both general and specific to location-based data, and describe $\delta$-presence, a risk-based measure (developed with Mehmet Ercan Nergiz and Maurizio Atzori) that provides a meaningful way to measure anonymity with respect to real-world privacy concerns. The goal is to lead into discussion of appropriate measures for anonymity of location data, an open challenge.

## 2.2 Speed-based Clustering for Discovering Interesting Places in Trajectories by Luis Otavio Alvares , UFRGS - Porto Alegre

Because of the large amount of trajectory data produced by mobile devices, there is an increasing need for mechanisms to extract knowledge from this data. Most existing works have focused on the geometric properties of trajectories, but recently emerged the concept of semantic trajectories, in which the background geographic information is integrated to trajectory sample points. In this new concept, trajectories are observed as a set of stops and moves, where stops are the most important parts of the trajectory. Stops and moves have been computed by testing the intersections of trajectories with a set of geographic objects given by the user. In this paper we present an alternative solution with the capability of finding interesting places that are not expected by the user. The proposed solution is a spatio-temporal clustering method, based on speed, to work with single trajectories. We compare the two different approaches with experiments on real data and show that the computation of stops using the concept of speed can be interesting for several applications.

### 2.3 PROBE: an Obfuscation System for the Protection of Sensitive Locations by Maria Luisa Damiani , Universit di Milano

The widespread adoption of location-based services (LBS) raises increasing concerns for the protection of personal lo- cation information. A common strategy, referred to as obfuscation, to protect location privacy is based on forwarding the LSB provider a coarse user location instead of the actual user location. Conventional approaches, based on such technique, are however based only on geometric methods and therefore are unable to assure privacy when the adversary has semantic knowledge about the reference spatial context. This paper provides a comprehensive solution to this problem. Our solution presents a novel approach that obfuscates the user location by taking into account the background knowledge about the reference space.

### 2.4 SECONDO: A Database Management System for Moving Objects by Ralf Hartmut Gting , FernUniversitt in Hagen

We present brief introductions to a data model to represent histories of moving objects, based on spatio-temporal data types, and to SECONDO, an extensible DBMS platform. Secondo's kernel system is extensible by algebra modules, offering data types and operations. The optimizer and the graphical user interface are also extensible, e.g. by translation rules and cost functions, or by specialized viewers, to accomodate the new types and operations. A considerable part of the data model for moving objects mentioned has been implemented in Secondo. The main part of the talk will be a demonstration of Secondo managing moving objects.

### 2.5 Generalization of trajectories in agreement with traffic distribution by Christine Krner , Fraunhofer IAIS - St. Augustin

The number of available GPS-trajectories is often very small compared to all possible routes within a city, and many streets will not be covered by the data sample. Yet, an application may require a mobility model for a whole city. We have developed a statistical simulation model, which scatters mobility locally, but preserves the trajectory on global scale. The model combines generalized trajectories with traffic frequencies and is thus able to reproduce the distribution of traffic routes. The first part introduces our method. In

the second part we would like to enter into discussion about the relationship of the model to anonymization in privacy preserving data mining.

### 2.6 Location Privacy in Medical Research Databases by Bradley Malin , Vanderbilt University

To support large-scale biomedical research projects, organizations need to share person-specific clinical and genomic sequences without violating the privacy of their data subjects. In the past, organizations protected subjects identities by removing identifiers, such as name and Social Security Number; however, our investigations illustrate that such de-identified data can be re-identified to named individuals using simple automated methods. In particular, location-visit patterns can uniquely identify research participants. In this presentation, I will review how such patterns can be automatically constructed and exploited, using examples from real world medical research databases. I will then illustrate how formal anonymity protection models can be devised and applied to provably prevent such patterns from being leveraged by arbitrary researchers, as well as insiders with background knowledge. This presentation will conclude with a discussion on the challenges to applying these technologies in the real world.

### 2.7 Towards a Geometric Interpretation of Double-Cross Matrix-based Similarity of Polylines by Bart Moelans , Hasselt University - Diepenbeek

One of the formalisms to qualitatively describe polylines in the plane are double-cross matrices. In a double-cross matrix the relative position of any two line segments in a polyline is described with respect to a double cross based on their start points. Two polylines are called DC-similar if their double-cross matrices are identical. Although double-cross matrices have been widely applied, a geometric interpretation of the similarity they express is still lacking. In this paper, we provide a first step in the geometric interpretation of this qualitative definition of similarity. In particular, we give an effective characterization of what DC-similarity means for polylines that are drawn on a grid. We also provide algorithms that, given a DC-matrix, check whether it is realizable by a polyline on a grid and that construct, if possible, in quadratic time example polylines that satisfy this matrix. We also describe algorithms to reconstruct polylines, satisfying a given double-cross matrix, in the two-dimensional plane, that is, not necessarily on a grid.

### 2.8 Space, time and prisms by Walied Othman , Hasselt University - Diepenbeek

Hagerstrand modelled moving objects with space-time prisms. Objects that bound people's movements between time-stamped locations given speed limit constraints. In this talk we outline how we adapted the model to fit road networks, where all edges can have different speed limits. Furthermore, we outline how we then proceed to compute the alibi query in this setting, which determines if two moving objects could have met or not. Lastly, we expand our model to support uncertain, and possibly disconnected in time and space, anchorpoints, instead of certain time-stamped locations and introduce the uncertain prism, where every point can be linked to a likelihood, the fraction of space-time prisms that actually contain that point together with their respective probabilities.

### 2.9 Pattern-Preserving k-Anonymization of Sequences and its Application to Mobility Data Mining by Dino Pedreschi , Universit di Pisa

Sequential pattern mining is a major research field in knowledge discovery and data mining. Thanks to the increasing availability of transaction data, it is now possible to provide new and improved services based on users' and customers' behavior. However, this puts the citizen's privacy at risk. Thus, it is important to develop new privacy-preserving data mining techniques that do not alter the analysis results significantly. In this paper we propose a new approach for anonymizing sequential data by hiding infrequent, and thus potentially sensible, subsequences. Our approach guarantees that the disclosed data are $k$-anonymous and preserve the quality of extracted patterns. An application to a real-world moving object database is presented, which shows the effectiveness of our approach also in complex contexts.

### 2.10 Trajectory Compression under Network Constraints by Nikos Pelekis, University of Piraeus

The wide usage of location aware devices, such as GPS-enabled cellphones or PDAs, generates vast volumes of spatiotemporal streams modeling movement of objects, which eventually bring on management challenges, such as storage and querying. Therefore, compression techniques are inevitable also in the field of Moving Object Databases. Existing work on this problem is relatively limited and mainly driven by line simplification and data sequence

compression techniques. Moreover, due to the (unavoidable) erroneous measurements from GPS devices, the problem of matching the trajectory recordings on a traffic network has gain the attention of the research community. So far, the proposed compression techniques are not designed for network constrained moving objects, while map matching algorithms do not consider compression issues. In this paper, we propose (for the first time, to the best of our knowledge) solutions tackling the combined, compression and map matching, problem, the efficiency of which is demonstrated through an extensive experimental evaluation using synthetic and real trajectory datasets.

### 2.11 On the semantic interpretation of trajectory patterns by Chiara Renso, ISTI-CNR - Pisa

We present an ontology-based methodology for the post-processing of mined trajectory patterns to enable movement understanding and analysis. The objective is to provide a model for the semantic interpretation of trajectory patterns computed by mining algorithms. This has been achieved by means of a semantic enrichment process, where raw trajectories are enhanced with semantic information and integrated with application domain knowledge encoded in an ontology. The reasoning mechanisms provided by the chosen ontology formalism is exploited to accomplish a further semantic enrichment step that gives a possible interpretation of discovered patterns in respect to the application domain. A sketch of the realized system is given, along with some examples to demonstrate the usefulness and feasibility of the approach.

### 2.12 Temporal Support of Regular Expressions in Sequential Pattern Mining by Alejandro Vaisman , University of Buenos Aires

Classic algorithms for sequential pattern discovery, return all frequent sequences present in a database. Since, in general, only a few ones are interesting from a user's point of view, languages based on regular expressions (RE) have been proposed to restrict frequent sequences to the ones that satisfy user-specified constraints. Although the support of a sequence is computed as the number of data-sequences satisfying a pattern with respect to the total number of data-sequences in the database, once regular expressions come into play, new approaches to the concept of support are needed. For example, users may be interested in computing the support of the RE as a whole, in addition to the one of a particular pattern. As a simple example, the

expression $(A|B).C$ is satisfied by sequences like A.C or B.C. Even though the semantics of this RE suggests that both of them are equally interesting to the user, if neither of them verifies a minimum support (although together they do), they would not be retrieved. Also, when the items are frequently updated, the traditional way of counting support in sequential pattern mining may lead to incorrect (or, at least incomplete), conclusions. For example, if we are looking for the support of the sequence A.B, where A and B are two items such that A was created after B, all sequences in the database that were completed before A was created, can never produce a match. Therefore, accounting for them would underestimate the support of the sequence A.B. The problem gets more involved if we are interested in categorical sequential patterns. In light of the above, in this paper we propose to revise the classic notion of support in sequential pattern mining, introducing the concept of temporal support of regular expressions, intuitively defined as the number of sequences satisfying a target pattern, out of the total number of sequences that could have possibly matched such pattern, where the pattern is defined as a RE over complex items (i.e., not only item identifiers, but also attributes and functions). We present and discuss a theoretical framework for these novel notion of support.

### 2.13 Semantic Trajectory Data Mining: a User Driven Approach by Vania Bogorny , UFRGS - Porto Alegre

Trajectories left behind cars, humans, birds or any other moving object are a new kind of data which can be very useful in decision making process in several application domains. These data, however, are normally available as sample points, and therefore have very little or no semantics. The analysis and knowledge extraction from trajectory sample points is very difficult from the user's point of view, and there is an emerging need for new data models, manipulation techniques, and tools to extract meaningful patterns from these data. In this paper we propose a new methodology for knowledge discovery from trajectories. We propose through a semantic trajectory data mining query language several functionalities to select, preprocess, and transform trajectory sample points into semantic trajectories at higher abstraction levels, in order to allow the user to extract meaningful, understandable, and useful patterns from trajectories. We claim that meaningful patterns can only be extracted from trajectories if the background geographical information is considered. Therefore we build the proposed methodology considering both moving object data and geographic informa-

tion. The proposed language has been implemented in a toolkit in order to provide a first software prototype for trajectory knowledge discovery.

### 2.14 Legal Issues in Privacy of Location Data by Chris Clifton, Purdue University

This mini-tutorial presents a sampling of privacy regulations in the context of location data. We briefly present both regulations of the text, as well as a selection of opinions, interpretations, and case law.

### 2.15 Location Prediction by means of Trajectory Pattern Mining by Fosca Giannotti , ISTI-CNR - Pisa

The increasing offer of location based services together with the increasing availability of mobility data is leading to an increasing interest in the analysis of movement. In this paper, we propose a method to predict the next location of a moving object, which uses the movement patterns previously extracted by the Trajectory Pattern paradigm which yields movement patterns as sequences of frequently visited regions with typical travel times. Several proposals in the literature try to predict the next location of a moving object based only on the movement history of the object itself. On the contrary, this work introduces a method for learning a predictor using the movements of all moving objects in a certain area. The predictor is built on top of the spatio-temporal sequences extracted by Trajectory Pattern in the form of T-patterns Tree, which represents the set of all typical movement behaviors in the selected area. Therefore, the prediction of the next location of a new moving object consists in finding on the T-patterns Tree the path that best matches the current sequence of movements of that object. Three different best matching methods to classify a new moving object have been proposed and their impact on the quality of classification has been extensively studied.

### 2.16 Enriching trajectory data and trajectory pattern semantics with background knowledge by Jose Macedo , EPFL - Lausanne

Many real world applications today are built on analyses of movement and related features. Examples of such applications include transportation management, urban planning, tourism services, and animal migration monitoring, just to name a few. Recent database modeling and management research

prototypes have the capability to store and manipulate movement data in terms of point or region geometries that evolve over time (moving point or moving and deforming region). This captures the spatio-temporal trace left by a moving object, but ignores its links with non-geometric information that enable a semantic interpretation of the movement of moving objects. The concept of trajectory has been introduced to express a more semantic understanding of movement, taking it closer to the perception of applications. This presentation describes a framework for a semantics-oriented structuring, modeling and querying of trajectory data. The framework relies on the definition of trajectory-related ontologies, addressing domain-independent and application-specific geometric and semantic facets. This framework was implemented in tool called ATHENA that allows to query and visualize trajectory data using application domain concepts.

### 2.17 Mining Traffic Patterns by Gerasimos Marketos , University of Piraeus

Modern sensing devices allow the collection of traffic data which can be analyzed using novel data mining techniques in order to extract useful traffic patterns. In this paper, we undertake the problem of analyzing traffic in a road network so as to discover time-focused traffic relationships like propagation, split, merge, sink and source among the road segments. A graph based modeling of the network traffic is presented which provides insights on the flow of movements within the network. We exploit this graph and utilize appropriate similarity measures in order to discover patterns in specific time periods. First experimental results illustrate the applicability and usefulness of our approach.

### 2.18 Map matching vehicle data using semantics of road networking and uncertainty by Bart Moelans, Hasselt University - Diepenbeek

I present a simple but efficient way to do a map matching of vehicle GPS data on a road network. The idea is to use the bead model introduced by Kuijpers and Othman for moving object data, to limit the scope where a vehicle could have been. Followed by a Dijkstra algorithm to reconstruct the original road using weights (higher weight is higher probability) and semantics of the road network (one way, tunnel, maximum speed, ...)

## 2.19 Design Issues and Solutions for a Trajectory Data Warehouse by Alessandra Raffaet , Universit Ca' Foscari di Venezia

We discuss how data warehousing technology can be used to store aggregate information about trajectories of mobile objects, and to perform OLAP operations over them. To this end, we define a data cube with spatial and temporal dimensions, discretized according to a hierarchy of regular grids. We analyse some measures of interest related to trajectories, such as the number of distinct trajectories in a cell or starting from a cell, the distance covered by the trajectories in a cell, the average and maximum speed and the average acceleration of the trajectories in the cell, and the frequent patterns obtained by a data mining process on trajectories. We focus on some specialised algorithms to transform data, and load the measures in the base cells. Such stored values are used, along with suitable aggregate functions, to compute the roll-up operations. The main issues derive, in this case, from the characteristics of input data, i.e., trajectory observations of mobile objects, which are usually produced at different rates, and arrive in streams in an unpredictable and unbounded way. Finally, we also discuss some use cases that would benefit from such a framework, in particular in the domain of supervision systems to monitor road traffic (or movements of individuals) in a given geographical area.

## 2.20 A Probabilistic Framework for Building Privacy-Preserving Synopses of Multi-dimensional Data by Domenico Sacca , University of Calabria

The problem of summarizing multi-dimensional data into lossy synopses supporting the estimation of aggregate range queries has been deeply investigated in the last three decades, and several summarization techniques have been proposed, based on different approaches, such as histograms, wavelets and sampling. The aim of most of the works in this area was to effectively summarize data in order to enable fast answers of range queries to be retrieved from the summary data, trading off the efficiency of query evaluation with the accuracy of query estimates. In this talk, the use of summarization is investigated in a more specific context, where privacy issues are taken into account. Thus, we study the problem of constructing privacy-preserving synopses, which can be exploited to efficiently support different analysis tasks, with no risk for privacy issues. To this aim, we in-

troduce a probabilistic framework modeling point-query estimates retrieved from data synopses as random variables, which enables the evaluation of the quality of the estimates of sensitive data which can be obtained by a user owning the summary data. Based on this framework, we devise a technique for constructing histogram-based synopses of multi-dimensional data which provide as much accurate as possible answers for a given workload of safe queries, while preventing high-quality estimates of sensitive information from being extracted. Issues concerned with geographic nature of data will be discussed as well.

### 2.21 Safety and Privacy in Mobile Services by Qianhong Wu, Universitat Rovira i Virgili -Tarragona

We report new safety and privacy issues in mobile services including vehicle ad-hoc networks, location based services and RFID. In addition to the general security concerns, we identify some subtle ones, e.g., some security or privacy risks from the employed security techniques in existing solutions, which have yet not been noticed so far. Motivated by the observations, new solutions are proposed to meet these safety and privacy requirements with rational performance.