

Mobility Data Mining and Privacy

Edited by

Christopher W. Clifton¹, Bart Kuijpers², Katharina Morik³, and Yucel Saygin⁴

1 Purdue University, US, clifton@cs.purdue.edu

2 Hasselt University – Diepenbeek, BE, bart.kuijpers@uhasselt.be

3 TU Dortmund, DE

4 Sabanci University – Istanbul, TR

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 12331 “Mobility Data Mining and Privacy”. Mobility data mining aims to extract knowledge from movement behaviour of people, but this data also poses novel privacy risks. This seminar gathered a multidisciplinary team for a conversation on how to balance the value in mining mobility data with privacy issues. The seminar focused on four key issues: Privacy in vehicular data, in cellular data, context-dependent privacy, and use of location uncertainty to provide privacy.

Seminar 12.–17. August, 2012 – www.dagstuhl.de/12331

1998 ACM Subject Classification K.4.1 Public Policy Issues: Privacy

Keywords and phrases Privacy, Mobility, Cellular, Vehicular Data

Digital Object Identifier 10.4230/DagRep.2.8.16

1 Executive Summary

Chris Clifton

Bart Kuijpers

License  Creative Commons BY-NC-ND 3.0 Unported license
© Chris Clifton and Bart Kuijpers

Mobility Data Mining and Privacy aimed to stimulate the emergence of a new research community to address mobility data mining together with privacy issues. Mobility data mining aims to extract knowledge from movement behaviour of people. This is an interdisciplinary research area combining a variety of disciplines such as data mining, geography, visualization, data/knowledge representation, and transforming them into a new context of mobility while considering privacy which is the social aspect of this area. The high societal impact of this topic is mainly due to the two related facets of its area of interest, i.e., people’s movement behaviour, and the associated privacy implications. Privacy is often associated with the negative impact of technology, especially with recent scandals in the US such as AOL’s data release which had a lot of media coverage. The contribution of *Mobility Data Mining and Privacy* is to turn this negative impact into positive impact by investigating how privacy technology can be integrated into mobility data mining. This is a challenging task which also imposes a high risk, since nobody knows what kinds of privacy threats exist due to mobility data and how such data can be linked to other data sources.

The seminar looked closely at two application areas: Vehicular data and cellular data. Further discussions covered two specific new general approaches to protecting location privacy: context-dependent privacy, and location uncertainty as a means to protect privacy. In each



Except where otherwise noted, content of this report is licensed under a Creative Commons BY-NC-ND 3.0 Unported license

Mobility Data Mining and Privacy, *Dagstuhl Reports*, Vol. 2, Issue 8, pp. 16–53

Editors: Christopher W. Clifton, Bart Kuijpers, Katharina Morik, and Yucel Saygin



DAGSTUHL
REPORTS

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

of these areas, new ideas were developed; further information is given in the working group reports.

The seminar emphasized discussion of issues and collaborative development of solutions – the majority of the time was divided between working group breakout sessions followed by report-back and general discussion sessions. While the working group reports were written by subgroups, the contents reflect discussions involving all 22 participants of the seminar.

The seminar concluded that there are numerous challenges to be addressed in mobility data mining and privacy. These challenges require investigation on both technical and policy levels. Of particular importance is educating stakeholders from various communities on the issues and potential solutions.

2 Table of Contents

Executive Summary

<i>Chris Clifton and Bart Kuijpers</i>	16
--	----

Overview of Talks

Dynamic privacy adaptation in ubiquitous computing <i>Florian Schaub</i>	19
Privacy-Aware Spatio-Temporal Queries on Unreliable Data Sources <i>Erik Buchmann</i>	20
Privacy-preserving sharing of sensitive semantic locations under road-network constraints <i>Maria-Luisa Damiani</i>	20
Methods of Analysis of Episodic Movement Data <i>Thomas Liebig</i>	20
Privacy-preserving Distributed Monitoring of Visit Quantities <i>Christine Körner</i>	21
A visual analytics framework for spatio-temporal analysis and modelling <i>Gennady Andrienko</i>	22
Tutorial: Privacy Law <i>Nilgun Basalp</i>	22
“Movie night”: Tutorial on Differential Privacy <i>Christine Task (video of talk by non-participant)</i>	22

Working Groups

Working Group: Cellular Data <i>Gennady Andrienko, Aris Gkoulalas-Divanis, Marco Gruteser, Christine Körner, Thomas Liebig, Klaus Rechert, and Michael Marhöfer</i>	23
Working Group: Vehicular Data <i>Glenn Geers, Marco Gruteser, Michael Marhoefer, Christian Wietfeld, Claudia Sánta, Olaf Spinczyk, and Ouri Wolfson</i>	32
Working Group: Context-dependent Privacy in Mobile Systems <i>Florian Schaub, Maria Luisa Damiani, and Bradley Malin</i>	36
Working Group: Privacy through Uncertainty in Location-Based Services <i>Nilgün Basalp, Joachim Biskup, Erik Buchmann, Chris Clifton, Bart Kuijpers, Walied Othman, and Erkay Savas</i>	44

Open Problems

What we learned	51
New Discoveries	51
What needs to be done	52
Future plans	52


Participants	53
-------------------------------	----

3 Overview of Talks

We tried to maximize the time spent in discussions, which limited the time available for talks. However, we did have a few talks, primarily from young researchers. We also had some short tutorial talks given as the need arose based on the direction of the ongoing discussions. Titles and brief abstracts are given below.

3.1 Dynamic privacy adaptation in ubiquitous computing

Florian Schaub, (Universität Ulm, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Florian Schaub

Ubiquitous and pervasive computing is characterized by the integration of computing aspects into the physical environment. Physical and virtual worlds start to merge as physical artifacts gain digital sensing, processing, and communication capabilities. This development introduces a number of privacy challenges. Physical boundaries lose their meaning in terms of privacy demarcation. Furthermore, the tight integration with the physical world necessitates the consideration not only of observation and information privacy aspects but also disturbances of the user's physical environment [1].


By viewing privacy as a dynamic regulation process and developing respective privacy mechanisms, we aim to support users in ubiquitous computing environments in gaining privacy awareness, making informed privacy decisions, and controlling their privacy effectively. The presentation outlined our dynamic privacy adaptation process for ubiquitous computing that supports privacy regulation based on the user's current context [2]. Furthermore, our work on a higher level privacy context model [3] has been presented. The proposed privacy context model captures privacy-relevant context features and facilitates the detection of privacy-relevant context changes in the user's physical and virtual environment. When privacy-relevant context changes occur, an adaptive privacy system can dynamically adapt to the changed situation by reasoning about the context change and learned privacy preferences of an individual user. Individualized privacy recommendations can be offered to the user or sharing behavior can be automatically adjusted to help the user maintain a desired level of privacy.

References

- 1 Bastian Könings, Florian Schaub, "Territorial Privacy in Ubiquitous Computing", Proc. 8th Int. Conf. on Wireless On-demand Network Systems and Services (WONS '11), IEEE 2011 DOI: 10.1109/WONS.2011.5720177
- 2 Florian Schaub, Bastian Könings, Michael Weber, Frank Kargl, "Towards Context Adaptive Privacy Decisions in Ubiquitous Computing", Proc. 10th Int. Conf. on Pervasive Computing and Communications (PerCom '12), Work in Progress, IEEE 2012 DOI: 10.1109/PerComW.2012.6197521
- 3 Florian Schaub, Bastian Könings, Stefan Dietzel, Michael Weber, Frank Kargl, "Privacy Context Model for Dynamic Privacy Adaptation in Ubiquitous Computing", Proc. 6th Int. Workshop on Context-Awareness for Self-Managing Systems (Casemans '12), Ubicomp '12 workshop, ACM 2012

3.2 Privacy-Aware Spatio-Temporal Queries on Unreliable Data Sources


Erik Buchmann (KIT – Karlsruhe Institute of Technology, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Erik Buchmann

A declarative spatio-temporal query processor is an important building block for many kinds of location-based applications. Such applications often apply methods to obfuscate, anonymize or delete certain spatio-temporal information for privacy reasons. However, the information that some data has been modified is privacy-relevant as well. This talk is about hiding the difference between spatio-temporal data that has been modified for privacy reasons, and unreliable information (e.g., missing values or sensors with a low precision), on the semantics level of a query processor. In particular, we evaluate spatio-temporal predicate sequences like Enter (an object was outside of a region first, then on the border, then inside) to true, false, maybe. This allows a wide range of data analyses without making restrictive assumptions on the data quality or the privacy methods used.

3.3 Privacy-preserving sharing of sensitive semantic locations under road-network constraints

Maria-Luisa Damiani (University of Milano, IT)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Maria-Luisa Damiani

This talk illustrates recent research on the protection of sensitive positions in real time trajectories under road network constraints

3.4 Methods of Analysis of Episodic Movement Data

Thomas Liebig (Fraunhofer IAIS – St. Augustin, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Thomas Liebig

Analysis of people's movements represented by continuous sequences of spatio-temporal data tuples received lots of attention within the last years. Focus of the studies was mostly GPS data recorded on a constant sample rate. However, creation of intelligent location aware models and environments requires also reliable localization in indoor environments as well as in mixed indoor outdoor scenarios. In these cases, signal loss makes usage of GPS infeasible, therefore other recording technologies evolved.

Our approach is analysis of episodic movement data. This data contains some uncertainties among time (continuity), space (accuracy) and number of recorded objects (coverage). Prominent examples of episodic movement data are spatio-temporal activity logs, cell based tracking data and billing records. To give one detailed example, Bluetooth tracking monitors presence of mobile phones and intercoms within the sensors footprints. Usage of multiple sensors provides flows among the sensors.

Most existing data mining algorithms use interpolation and therefore are infeasible for this kind of data. For example, speed and movement direction cannot be derived from episodic data; trajectories may not be depicted as a continuous line; and densities cannot be computed.

Though this data is infeasible for individual movement or path analysis, it bares lots of information on group movement. Our approach is to aggregate movement in order to overcome the uncertainties. Deriving number of objects for spatio-temporal compartments and transitions among them gives interesting insights on spatio-temporal behavior of moving objects. As a next step to support analysts, we propose clustering of the spatio-temporal presence and flow situations. This work focuses as well on creation of a descriptive probability model for the movement based on Spatial Bayesian Networks.

We present our methods on real world data sets collected during a football game in Nîmes, France in June 2011 and another one in Dusseldorf, Germany 2012. Episodic movement data is quite frequent and more methods for its analysis are needed. To facilitate method development and exchange of ideas, we are willing to share the collected data and our findings.

3.5 Privacy-preserving Distributed Monitoring of Visit Quantities


Christine Körner (Fraunhofer IAIS – St. Augustin, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Christine Körner

The organization and planning of services (e.g. shopping facilities, infrastructure) requires up-to-date knowledge about the usage behavior of customers. Especially quantitative information about the number of customers and their frequency of visiting is important. In this paper we present a framework which enables the collection of quantitative visit information for arbitrary sets of locations in a distributed and privacy-preserving way. While trajectory analysis is typically performed on a central database requiring the transmission of sensitive personal movement information, the main principle of our approach is the local processing of movement data. Only aggregated statistics are transmitted anonymously to a central coordinator, which generates the global statistics. In this presentation we introduce our approach including the methodical background that enables distributed data processing as well as the architecture of the framework. We further discuss our approach with respect to potential privacy attacks as well as its application in practice. We have implemented the local processing mechanism on an Android mobile phone in order to ensure the feasibility of our approach.

3.6 A visual analytics framework for spatio-temporal analysis and modelling

Gennady Andrienko (Fraunhofer IAIS – St. Augustin, DE)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Gennady Andrienko


Main reference N. Andrienko, G. Andrienko, “A Visual Analytics Framework for Spatio-temporal Analysis and Modelling,” *Data Mining and Knowledge Discovery*, 2013 (accepted).

URL <http://dx.doi.org/10.1007/s10618-012-0285-7>

To support analysis and modelling of large amounts of spatio-temporal data having the form of spatially referenced time series (TS) of numeric values, we combine interactive visual techniques with computational methods from machine learning and statistics. Clustering methods and interactive techniques are used to group TS by similarity. Statistical methods for TS modelling are then applied to representative TS derived from the groups of similar TS. The framework includes interactive visual interfaces to a library of modelling methods supporting the selection of a suitable method, adjustment of model parameters, and evaluation of the models obtained. The models can be externally stored, communicated, and used for prediction and in further computational analyses. From the visual analytics perspective, the framework suggests a way to externalize spatio-temporal patterns emerging in the mind of the analyst as a result of interactive visual analysis: the patterns are represented in the form of computer-processable and reusable models. From the statistical analysis perspective, the framework demonstrates how TS analysis and modelling can be supported by interactive visual interfaces, particularly, in a case of numerous TS that are hard to analyse individually. From the application perspective, the framework suggests a way to analyse large numbers of spatial TS with the use of well-established statistical methods for TS analysis.

3.7 Tutorial: Privacy Law


Nilgun Basalp (Istanbul Bilgi University, TR)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Nilgun Basalp

This session is a brief tutorial on the EC 95/46 privacy directive, with a focus on issues affecting mobile data.

3.8 “Movie night”: Tutorial on Differential Privacy

Christine Task (video of talk by non-participant)

License  Creative Commons BY-NC-ND 3.0 Unported license
© Christine Task (video of talk by non-participant)

Differential Privacy is a relatively new approach to privacy protection, based on adding sufficient noise to query results on data to hide information of any single individual. This tutorial, given as a CERIAS seminar at Purdue in April, is an introduction to Differential Privacy targeted to a broad audience. Several of the seminar participants gathered for a “movie night” to watch a video of this presentation.




The presenter, Christine Task, is a Ph.D. student at Purdue University working with Chris Clifton. She has a B.S. in theoretical math from Ohio State and M.S. in computer science from Indiana University.

http://www.cerias.purdue.edu/news_and_events/events/security_seminar/details/index/j9cvs3as2h1qds1jrdqfdc3hu8

4 Working Groups

4.1 Working Group: Cellular Data

Gennady Andrienko, Aris Gkoulalas-Divanis, Marco Gruteser, Christine Körner, Thomas Liebig, Klaus Rechert, and Michael Marhöfer

License    Creative Commons BY-NC-ND 3.0 Unported license
© Gennady Andrienko, Aris Gkoulalas-Divanis, Marco Gruteser, Christine Körner, Thomas Liebig, Klaus Rechert, and Michael Marhöfer

4.1.1 Introduction

The phones we carry around as we go about our daily lives do not only provide a convenient way to communicate and access information, but also pose privacy risks by collecting data about our movements and habits. For example, they can record when we get up in the morning, when we leave our homes, whether we violate speed limits, how much time we spend at work, how much we exercise, whom we meet, and where we spend the night. The places we visit allow inferences about not just one, but many potentially sensitive subjects: health, sexual orientation, finances or creditworthiness, religion, and political opinions. For many, such inferences can be embarrassing, even if they are untrue and simply misinterpretations of the data. For some, this movement data can even pose a danger of physical harm, such as in stalking cases.

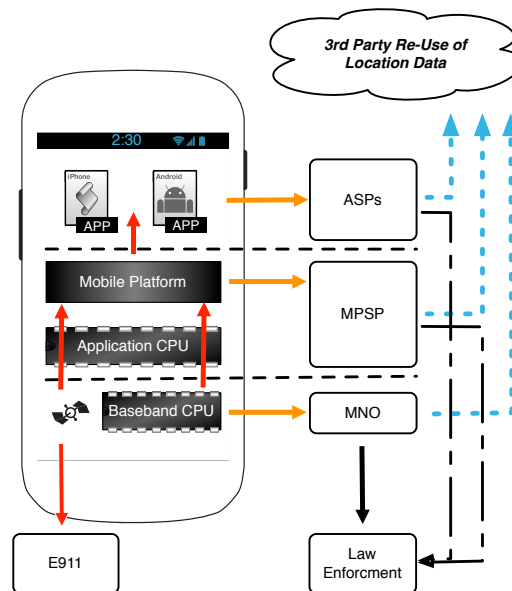
These risks have been amplified by the emergence of smartphones and the app economy over the last few years. We have witnessed a fundamental shift in mobility data collection and processing from a selected group of tightly regulated cellular operators to a complex web of app providers and Internet companies. This new ecosystem of mobility data collectors relies on a more sophisticated mix of positioning technologies to acquire increasingly precise mobility data. In addition, smartphones also carry a much richer set of sensors and input devices, which allow collection of a diverse set of other data types in combination with the mobility data. Many of these types of data were previously unavailable. While individual aspects of these changes have been highlighted in a number of articles as well as in a string of well-publicized privacy scandals, the overall structure of current mobility data streams remains confusing.

The goal of the Dagstuhl cellular data working group was to survey this new mobility data ecosystem and to discuss the implications of this broader shift. Section 4.1.2 gives an overview about the types of data collected by mobile network operators (MNO), mobile platform service providers (MPSP) and application service providers (ASP). In Section 4.1.3 we discuss privacy threats and risks that arise from the data collection. We conclude our work with a section on the manifold implications of this rapidly evolving mobility data ecosystem. We find that it is difficult to understand the data flows, apparently even for the service providers and operators themselves [5, 13, 18]. There appears to be a much greater need for transparency, perhaps supported by technical solutions that monitor and raise awareness

of such data collection. We find that location data is increasingly flowing across national borders, which raises questions about the effectiveness of current regulatory protections. We also find that applications are accessing a richer set of sensors, which allows cross-referencing and linking of data in ways that are not yet fully understood.

4.1.2 Location Data Collection

We distinguish three groups of data collectors (observers): mobile network operators (MNO), mobile platform service providers (MPSP), and application service providers (ASP). While for the first group of observers (MNOs), location data is generated and collected primarily due to technical reasons, i.e. efficient signaling, in the case of MPSP and ASPs location information is usually generated and collected to support positioning, mapping, and advertising services and to provide various kinds of location based services. Figure 1 provides a schematic overview on location data generated by mobile phones but also highlights the specific components and building blocks of mobile phones which are controlled by the different entities. Furthermore, available location data originating from the aforementioned layers may be re-used to support various new (third-party) businesses. Typically the data is then anonymized or aggregated in some way before being transferred to third parties.



■ **Figure 1** A schematic overview on a today's smartphone, its essential building blocks and their controllers illustrating the generation and information flow of location data.

Most of the data collected by MNO, MPSP and ASP can be referred as 'Episodical Movement Data': data about spatial positions of moving objects where the time intervals between the measurements may be quite large and therefore the intermediate positions cannot be reliably reconstructed by means of interpolation, map matching, or other methods. Such data can also be called 'temporally sparse'; however, this term is not very accurate since the temporal resolution of the data may greatly vary and occasionally be quite fine.

4.1.2.1 Collection and Usage of Mobile Telephony Network Data

As an example for mobile telephony networks we discuss the widely deployed GSM infrastructure, as its successors UMTS (3G) and LTE (4G) have a significantly smaller coverage and share most of its principal characteristics. A typical GSM network is structured into cells, each served by a single base transceiver station (BTS). Larger cell-compounds are called location areas. To establish a connection to the mobile station (MS) e.g. in the case of an incoming connection request, the network has to know if the MS is still available and in which location area it is currently located. To cope with subscriber mobility the location update procedure was introduced. Either periodically or when changing the location area, a location update is triggered. The time lapse between periodic location updates is defined by the network and varies between infrastructure providers.

Additionally, the infrastructure's radio subsystem measures the distance of phones to the serving cell to compensate for the signal propagation delay between the MS and BTS. The timing advance (TA) value (8-bit value) is used to split the cell radius into virtual rings. In the case of GSM these rings have a size of roughly 550 m in diameter. The TA is regularly updated and is sent by the serving infrastructure to each mobile phone.

For billing purposes so-called call data records (CDR) are generated. This datum usually consists of the cell-ID where a call has been started (either incoming or outgoing), the cell where a call has been terminated, start time, duration, ID of the caller and the phone number called. A typical GSM cell size ranges from a few hundred meters in diameter to a maximum size of 35 km. In a typical network setup a cell is further divided into three sectors. In that case also the sector ID is available, and the sector ID is also part of a call record. CDRs are usually not available in real-time. However, MNOs store CDRs for a certain time span, either because of legal requirement (e.g. EU data retention directive [9]) or accounting purposes.

Mobile telephony networks and their physical characteristics are able to help locating mobile phone users in the case of an emergency and may be a valuable tool for search and rescue (SAR) [21]. For instance, [6] analyzed post-disaster populations displacement using SIM-card movements in order to improve the allocation of relief supplies.

Furthermore, location information gathered through mobile telephony networks is now a standard tool for crime prosecution and is used by the EC Data Retention Directive with the aim of reducing the risk of terror and organized crime [9]. As an example, law enforcement officials seized CDRs over a 48 hour timespan resulting in 896,072 individual records containing 257,858 call numbers after a demonstration in Dresden, Germany, went violent [20]. Further, the police of North Rhine-Westphalia issued 225,784 active location determinations on 2,644 different subjects in 778 preliminary proceedings in 2010 [23]. While in principle law enforcement could also collect location- and movement-data from MPSP and ASPs, difficulties arise if such data is stored outside of the respective jurisdiction.

Additionally, commercial services are based on the availability of live mobility patterns of larger groups. (e.g. for traffic monitoring or location-aware advertising [19]). Thus, location information of network subscribers might be passed on to third parties. Usually, subscribers are neither aware of the extent of their information disclosure (just by carrying a switched-on mobile phone), nor of how the collected data is used and by whom. Even sporadic disclosure of location data, e.g. through periodic location updates, are able to disclose a users frequently visited places (i.e. preferences) in an accuracy similar to continuous location data after 10-14 days [25].

4.1.2.2 Collection and Usage of Data Through MPSPs and ASPs

Over the past years, researchers and journalists have started to analyze apps and mobile operating systems w.r.t. the collection of personal data [12, 16, 7, 1, 26, 27, 5]. The analyses show that sensitive information is accessed and transmitted. There are mainly three reasons for MPSP to collect location information: positioning, map services, and advertising.

Even though a mobile phone may not be equipped with GPS, a position may be obtained by approximate location determination based on mobile telephony infrastructure or WiFi. The sensor data is sent to external services and external information sources are used to improve (i.e. speed-up) the determination of the user's current location. For instance, in Spring 2011 it was found that Apple's iPhone generates and stores a user's location history, more specifically, data records correlating visible WiFi access-points or mobile telephony cell-ids with the device's GPS location on the user's phone. Moreover, the recorded data-sets are frequently synchronized with the platform provider. Presumably, this data is used by MPSPs to improve database-based, alternative location determination techniques for situations where GNSS or similar techniques are not available or not operational.

By aggregating location information of many users, such information could improve or enable new kinds of services. For instance, Google Mobile Maps makes use of user contributed data (with the user's consent) to determine and visualize the current traffic situation.

Finally, mobile advertising is one of the fastest growing advertising media, doubling its yearly revenue over the next years by a prediction of [10]. The availability of smartphones in combination with comprehensive and affordable mobile broadband communication has given rise to this new generation of advertising media, which allows to deliver up-to-date information in a context-aware and personalized manner. However, personal information as a user's current location and personal preferences are prerequisite for a tailored advertisement delivery. Consequently, MPSPs and ASPs are interested to profile users and personal data is disclosed in an unprecedented manner to various (unknown) commercial entities which poses serious privacy risks.

In order to perform dedicated tasks apps, also access other data such as the user's contacts, calendar and bookmarks as well as sensors readings (e.g. camera, microphone). If these apps have access to the Internet, they are potentially able to disclose this information and are a serious thread to user privacy [16]. Most often, advertisement libraries (e.g., as part of an app) require access to the phone information and location API [12] in order to obtain the phone's IMEI number and geographic position. For instance, Apple Siri records, stores and transmits any spoken request to Apple's cloud-based services where it is processed through speech recognition software, is analyzed to be understood, and is subsequently serviced. The computed result of each request is communicated back to the user. Additionally, to fully support inferencing from context, Siri is "expected to have knowledge of users' contact lists, relationships, messaging accounts, media (songs, playlists, etc) and more"¹, including location data to provide the context of the request, which are communicated to Apple's data center. As an example of Siri's use of location data, users are able to geo-tag familiar locations (such as their home or work) and set a reminder when they visit these locations. Moreover, user location data is used to enable Siri to support requests for finding the nearest place of interest (e.g., restaurant) or to report the local weather.

¹ <http://privacycast.com/siri-privacy-and-data-collection-retention/>, Online, Version of 9/6/2012

4.1.3 Privacy Threats and Risks

From a business perspective, mobility data with sufficiently precise location estimation are often valuable data enabling various location-based services; from the perspective of privacy advocates, such insights are often deemed a privacy threat or a privacy risk. Location privacy risks can arise if a third-party acquires a data tuple (user ID, location), which proves that an identifiable user has visited a certain location. In most cases, the datum will be a triple that also includes a time field describing when the user was present at this location. Note that in theory there are no location privacy risks if the user cannot be identified or if the location cannot be inferred from the data. In practice, however, it is difficult to determine when identification and such inferences are possible.

4.1.3.1 Collection of Location Information with Assigned User ID

Collecting location information along with a user id is the most trivial case of observing personal movement information, as long as the location of the user is estimated with sufficient accuracy for providing the intended LBS. In case the location is not yet precise enough, various techniques (e.g. fusion of several raw location data from various sensors) allow for improving the accuracy.

Additionally, ASP may have direct access to a variety of publicly available spatial and temporal data such as geographical space and inherent properties of different locations and parts of the space (e.g. street vs. park), or various objects existing or occurring in space and time: static spatial objects (having particular constant positions in space), events (having particular positions in time) and moving objects (changing their spatial positions over time). Such information either exists in explicit form in public databases like OSM, WikiMapia or in ASP's data centers, or can be extracted from publicly available data by means of event detection or situation similarity assessment [3][4]. Combining such information with positions and identities of users allow deep semantic understanding of their habits, contacts, and lifestyle.

4.1.3.2 Collection of Anonymous Location Information

When location data is collected without any obvious user identifiers, privacy risks are reduced and such seemingly anonymous data is usually exempted from privacy regulations. It is, however, often possible to re-identify the user based on quasi-identifying data that has been collected. Therefore, the aforementioned risks can apply even to such anonymous data.

The degree of difficulty in re-identifying anonymous data depends on the exact details of the data collection and anonymization scheme and the adversaries access to background information. Consider the following examples:

Re-identifying individual samples. Individual location records can be re-identified through observation re-identification [22]. The adversary knows that user Alice was the only user in location (area) l at time t , perhaps because the adversary has seen the person at this location or because records from another source prove it. If the adversary now finds an anonymous datum (l, t) in the collected mobility data, the adversary can infer that this datum could only have been collected from Alice and has re-identified the data. In this trivial example, there is actually no privacy risk from this re-identification because the adversary knew a priori that Alice was at location l at time t , so the adversary has not learned anything new. There are, however, three important variants of this trivial case that can pose privacy risks. First, the anonymous datum may contain a more precise location l' or a more precise time t' than the adversary knew about a priori. In this case, the adversary learns this more

precise information. Second, the adversary may not know that Alice was at l but simply know that Alice is the only user who has access to location l . In this latter case, also referred to as restricted space identification, the adversary would learn when Alice was actually present at this location. Third, the anonymous datum may contain additional fields with potentially sensitive information that the adversary did not know before. Note, however, that such additional information can also make the re-identification task easier.

Re-identifying time-series location data. Re-identification can also become substantially easier when location data is repeatedly collected and time-series location traces are available. We refer to time-series location traces, rather than individual location samples when it is clear which set of location samples was collected from the same user (even though the identity of the user is not known). For example, the location data may be stored in separate files for each user or a pseudonym may be used to link multiple records to the same user.

Empirical research [11] has further observed that the pair (home location, work location) is often already identifying a unique user. A recent empirical study [29] explains various approaches for re-identification of a user. Another paper has analyzed the consequences of increasingly strong re-identification methods to privacy law and its interpretation [24]. Further re-identification methods for location data rely on various inference and data mining techniques.

4.1.3.3 Collection of Data without Location

Even in absence of actual location readings provided by positioning devices, location disclosures may occur by means of other modern technologies. Recent work by Han, et al. [17] demonstrated that the complete trajectory of a user can be revealed with a 200 m accuracy by using accelerometer readings, even when no initial location information is known. What is even more alarming is that accelerometers, typically installed in modern smartphones, are usually not secured against third-party applications, which can easily obtain such readings without requiring any special privileges. Acceleration information can thus be transmitted to external servers and be used to disclose user location even if all localization mechanisms of the mobile device are disabled.

Furthermore, several privacy vulnerabilities may be exposed through the various resource types that are typically supported and communicated by modern mobile phone applications. Hornyack, et al. [16] examined several popular Android applications which require both internet access and access to sensitive data, such as location, contacts, camera, microphone, etc. for their operation. Their examination showed that almost 34% of the top 1100 popular Android applications required access to location data, while almost 10% of the applications required access to the user contacts. As can be anticipated, access of third-party applications to such sensitive data sources may lead to both user re-identification as well as sensitive information disclosure attacks, unless privacy enabling technology is in place.

4.1.4 Implications

Potentially sensitive location data from the use of smartphones is now flowing to a largely inscrutable ecosystem of international app and mobile platform providers, often without knowledge of the data subject. This represents a fundamental shift from the traditional mobile phone system, where location data was primarily stored at more tightly regulated cellular carriers that operated within national borders.

A large number of apps customize the presented information or their functionality based on user location. Examples of such apps include local weather information, location-based

reminders, maps and navigation, restaurant rating, and friend finders. Such apps often transmit the user location to a server, where it may be stored for a longer duration.

It is particularly noteworthy, however, that mobile advertisers and platform providers have emerged as an additional entity that aggregates massive sets of location records obtained from user interactions with a variety of apps. When apps request location information, the user location can also be disclosed to the mobile platform service provider as part of the wireless positioning service function. Even apps that do not need any location information to function, often reveal the user location to mobile advertisers. The information collected by these advertising and mobile providers is arguably more precise than the call data records stored by cellular carriers, since it is often obtained via WiFi positioning or the GPS. In addition, privacy notices by app providers often neglect to disclose such background data flows [1]. While the diversity of location-based apps has been foreseen by mobile privacy research to some extent—for example, research on spatial cloaking [14] has sought to provide privacy-preserving mechanisms for sharing location data with a large number of apps—this aggregation of data at mobile platform providers was less expected. In essence, this development is for economic reasons. Personal location information has become a tradable good: users provide personal information for targeted advertising in exchange for free services (quite similar to web-based advertising models). The advertising revenue generated out of such data, finances the operation of the service provider. Because of this implicit bargain between users and service providers, there is little incentive to curb data flows or adopt stronger technical privacy protections as long as it is not demanded by users or regulators.

We suspect, however, that many users are not fully aware of this implicit bargain. Therefore, we believe that it is most important from a privacy perspective to create awareness of these data flows among users, which is not incidentally the very first core principle of the fair information practice principles [28]. It is well understood that lengthy privacy disclosures, if they exist for smartphone apps, are not very effective at reaching the majority of users and even the recent media attention regarding smartphone privacy² does not appear to have found a sufficiently wide audience as our workshop discussions suggest. Raising awareness and empowering users to make informed decisions about their privacy will require novel approaches, user-interfaces, and tools.

When using smartphones, users should not only be aware of *what* data they are revealing to third-parties and how frequently it is revealed but also should be able to understand the potential risks of sharing such data. For instance, users/subscribers in the EU are currently entitled to get a full copy of their personal data stored by a commercial entity³ but such voluminous datasets can currently only be analyzed by experts. Even then, it will be difficult to judge what sensitive information can be learned from this dataset when it is linked with other data about the same person or when it is analyzed by a human expert with powerful visual analysis tools [2]. Was the precision of a location record sufficient to determine the building that a user has entered? Is it possible to reconstruct the path a users has taken between two location records? How easily can one infer habits or health of a person based on the location records collected from smartphones?

As another example, some service providers claim to collect location data only in anonymous form. The methods for re-identification, however, have evolved quickly. When can 'anonymized' time-series location data really qualify as data that is not personally identifiable information and remain outside most current privacy regulations? Finally, even

² For instance, <http://blogs.wsj.com/wtk-mobile/> Retrieved 2012/10/18.

³ For example, an Austrian student requested all personal data from Facebook and received a CD [15]

non-georeferenced data provided by the sensors embedded in a smartphone (camera, accelerometer, microphone, etc.), as well as the files stored in the internal memory (photos, music, playlists), allow extracting knowledge about a person's location and mobility. Overall, it appears necessary to investigate what associations can be established and what inferences can be made by a human when the data is considered in context, and how such information can be conveyed to users of services.

Users should also be able to learn in which countries their data is stored or processed, since this can have important implications for the applicable legal privacy framework. While the European Union has achieved some degree of harmonization of privacy standards for exported data from its citizens through the safe harbor provisions [8], differences still exist, for example, with respect to law enforcement access to user data. We believe that providing transparency of cross-border data flows would lead to a more meaningful public discussion of data protection policies. For example, when data is handled by multi-national corporations, should data subjects be given a choice where their data is processed and stored?

We hope that the research community will help address these questions and will interface with data protection authorities and policy experts to actively define privacy for this mobility data ecosystem.

References


- 1 Mobile apps for kids: Current privacy disclosures are disappointing. Technical report, Federal Trade Commission, 2012. http://www.ftc.gov/os/2012/02/120216mobile_apps_kids.pdf.
- 2 G. Andrienko and N. Andrienko. Privacy issues in geospatial visual analytics. In Georg Gartner, Felix Ortag, William Cartwright, Georg Gartner, Liqiu Meng, and Michael P. Peterson, editors, *Advances in Location-Based Services*, Lecture Notes in Geoinformation and Cartography, pages 239–246. Springer Berlin Heidelberg, 2012.
- 3 G. L. Andrienko, N. V. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel. From movement tracks through events to places: Extracting and characterizing significant places from mobility data. In *IEEE VAST*, pages 161–170. IEEE, 2011.
- 4 G. L. Andrienko, N. V. Andrienko, M. Mladenov, M. Mock, and C. Pölitiz. Identifying place histories from activity traces with an eye to parameter impact. *IEEE Trans. Vis. Comput. Graph.*, 18(5):675–688, 2012.
- 5 C. Arthur. iphone keeps record of everywhere you go. *The Guardian*, 2011. <http://www.guardian.co.uk/technology/2011/apr/20/iphone-tracking-prompts-privacy-fears>.
- 6 L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. von Schreeb. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in haiti. *PLoS Med*, 8(8):e1001083, 08 2011.
- 7 W. Enck, P. Gilbert, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth. Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. In *Proceedings of the 9th USENIX conference on Operating systems design and implementation*, OSDI'10, pages 1–6, Berkeley, CA, USA, 2010. USENIX Association.
- 8 European Parliament and European Council. Directive 95/46/ec on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Communities*, (L281), 1995.
- 9 Council European Parliament. Directive 2006/24/ec of the european parliament and of the council of 15 march 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks and amending directive 2002/58/ec. *Official Journal of the European Union*, L 105:54 – 63, 2006.

- 10 Gartner, Inc. Gartner says worldwide mobile advertising revenue forecast to reach \$3.3 billion in 2011, 2011. <http://www.gartner.com/it/page.jsp?id=1726614>.
- 11 Philippe Golle and Kurt Partridge. On the anonymity of home/work location pairs. In Hideyuki Tokuda, Michael Beigl, Adrian Friday, A. Brush, and Yoshito Tobe, editors, *Pervasive Computing*, volume 5538 of *Lecture Notes in Computer Science*, pages 390–397. Springer Berlin / Heidelberg, 2009.
- 12 M. C. Grace, W. Zhou, X. Jiang, and A.-R. Sadeghi. Unsafe exposure analysis of mobile in-app advertisements. In *Proceedings of the fifth ACM conference on Security and Privacy in Wireless and Mobile Networks*, WISEC '12, pages 101–112, New York, NY, USA, 2012. ACM.
- 13 A. Greenberg. Phone 'rootkit' maker carrier iq may have violated wiretap law in millions of cases. *Forbes*, 2011. <http://www.forbes.com/sites/andygreenberg/2011/11/30/phone-rootkit-carrier-iq-may-have-violated-wiretap-law-in-millions-of-cases/>. Retrieved 2012/10/18.
- 14 M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st international conference on Mobile systems, applications and services*, MobiSys '03, pages 31–42, New York, NY, USA, 2003. ACM.
- 15 K. Hill. *Forbes*, 2012. <http://www.forbes.com/sites/kashmirhill/2012/02/07/the-austrian-thorn-in-facebooks-side/>. Retrieved 2012/10/18.
- 16 P. Hornyack, S. Han, J. Jung, S. Schechter, and D. Wetherall. These aren't the droids you're looking for: retrofitting android to protect data from imperious applications. In *Proceedings of the 18th ACM conference on Computer and communications security*, CCS '11, pages 639–652, New York, NY, USA, 2011. ACM.
- 17 H. Jun, E. Owusu, L. T. Nguyen, A. Perrig, and J. Zhang. Accomplice: Location inference using accelerometers on smartphones. In *Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on*, pages 1–9, 2012.
- 18 D. Kravets. An intentional mistake: The anatomy of google's wi-fi sniffing debacle. *Wired*, 2011. <http://www.wired.com/threatlevel/2012/05/google-wifi-fcc-investigation/>. Retrieved 2012/10/18.
- 19 J. Krumm. Ubiquitous advertising: The killer application for the 21st century. *Pervasive Computing, IEEE*, 10(1):66–73, jan.-march 2011.
- 20 S. Landtag. Drucksache 5/6787. Sächsischer Landtag 5. Wahlperiode, 2011.
- 21 C. Ling, M. Loschonsky, and L. M. Reindl. Characterization of delay spread for mobile radio communications under collapsed buildings. In *IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, pages 329–334, Sept. 2010.
- 22 C. Y. T. Ma, D. K. Y. Yau, N. K. Yip, and N. S. V. Rao. Privacy vulnerability of published anonymous mobility traces. In *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, MobiCom '10, pages 185–196, New York, NY, USA, 2010. ACM.
- 23 Ministerium für Inneres und Kommunales NRW. Funkzellenauswertung (FZA) und Versenden "Stiller SMS" zur Kriminalitätsbekämpfung. MMD 15/3300, 2011.
- 24 P. Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, Vol. 57, p. 1701, 2010, 2009.
- 25 K. Rechert, K. Meier, B. Greschbach, D. Wehrle, and D. von Suchodoletz. Assessing location privacy in mobile communication networks. In H. Li X. Lai, J. Zhou, editor, *ISC 11*, LNCS 7001, pages 309–324. Springer, Heidelberg, 2011.
- 26 E. Smith. iphone applications & privacy issues: An analysis of application transmission of iphone unique device identifiers (udids). Technical report, PSKL, 2010. <http://www.pskl.us/wp/wp-content/uploads/2010/09/iPhone-Applications-Privacy-Issues.pdf>.

- 27 S. Thurm and I. Yukari Kane. Your apps are watching you. *The Wall Street Journal*, 2010. <http://online.wsj.com/article/SB10001424052748704694004576020083703574602.html>.
- 28 W. H. Ware. Records, computers, and the rights of citizens. Secretary's Advisory Committee on Automated Personal Data Systems, Department of Health, Education and Welfare, Washington, D.C., 1973.
- 29 H. Zang and J. Bolot. Anonymization of location data does not work: a large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, MobiCom '11, pages 145–156, New York, NY, USA, 2011. ACM.

4.2 Working Group: Vehicular Data

Glenn Geers, Marco Gruteser, Michael Marhoefer, Christian Wietfeld, Claudia Santa, Olaf Spinczyk, and Ouri Wolfson

License  Creative Commons BY-NC-ND 3.0 Unported license
 © Glenn Geers, Marco Gruteser, Michael Marhoefer, Christian Wietfeld, Claudia Santa, Olaf Spinczyk, and Ouri Wolfson

4.2.1 Vehicular Applications (An Introduction)

Currently there is much research in development of vehicular applications.

► **Example 1.** Imagine the year 2025 it is Saturday before Christmas and you want to go to a shopping mall to get some Christmas gifts for your friends. You go to your garage, enter your destination in the navigation device and start your e-car. Your current location, destination, preferable routes and maximum accepted arrival time is transferred to a central server and the fastest way is returned back by taking into account all other navigation requests. You follow the advised route through the city and the green phase of the traffic lights are optimized in your direction, because everybody else also wants to go to get the last gifts. Your car automatically adapts to the optimal speed of the green wave. To do so your position and type of car is send to the traffic light server every second. During your drive a friend is calling you and despite of the complete fusion of vehicle and mobile phone, you lose concentration for a short while and cannot see the person on the pedestrian crossing. You feel a jerky braking movements triggered by the pupil warning system. After some time you finally reach the shopping mall. Your car guides you to a free parking space next to the wine shop because it knows that your last bills showed a preference for wine. You have to cross the whole mall because this time you only need computer games for the children. After having found all gifts you walk back to your car and while leaving the parking place you are automatically charged based on the duration of your stay. The parking time is send to the shopping mall to analyze the buying behavior of customers. Late in the evening after your exhaustive shopping trip you finally arrive back home and your driving data including length of the trip, speed and driving behavior is send to your insurance to settle your account.

This vision of the future involves a lot of benefits in terms of safety, comfort and efficiency, for the driver as well as for society as a whole: accidents are prevented, parking fees are charge automatically, minimum stops at traffic lights are needed, no searching for parking spaces is needed and insurance fees are paid according to your individual risks. All these applications are currently under development in academia and/or industry.

Even though these applications offer a clear benefit, it is obvious that the involved systems gather a lot of individual data, which needs to be properly protected against privacy violations. It is therefore necessary to design the systems in a way, which protects the privacy,

not only to comply with legal regulations, but also to ensure, that users will accept these systems. The following paragraphs will show that vehicular applications may be implemented in different ways and users may decide to make a trade-off between utility, cost and privacy risks.

4.2.2 Glossary

Identifier : temporary device name

Identity : personally linked name

Data criticality : level of perceived damage due to data revelation (level of detail, amount of detail)

Identity revelation risk : describes the risk and effort to match an identifier to an identity

4.2.3 Expectations on Privacy & Privacy Risks

The expectations of users towards vehicular data analysis regarding privacy differ from other Internet applications due to several reasons:

- Driving a vehicle with a number plate takes place in public space.
- The operation of a vehicle requires to follow a set of well-established rules, which are enforced by police, with the help of electronic systems (e.g. cameras). It is a well established fact, that a driver/owner of a vehicle can be identified by officials can via the license plate, especially in case of traffic rule violations, but also in other legally relevant cases (e.g. bank robbery).

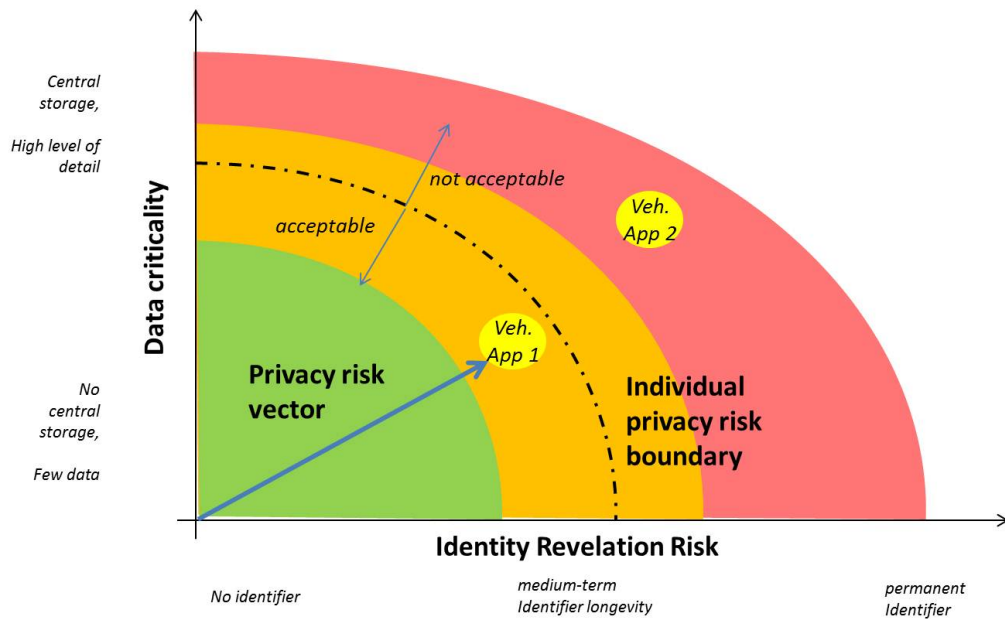
On the other hand, existing observation and enforcement is rather sparse and local. Names of persons are only revealed in case of violations and following certain regulations. Therefore anybody who does not violate the rules can move anonymously and it requires a very high technical and organizational effort to produce movement tracks of individuals.

With the introduction of ITS, many user benefits are associated, in particular fewer accidents, more comfort and less traffic jams. Nevertheless, with Intelligent Transport Systems (ITS) technical options are potentially introduced for deriving movement tracks with much lower effort. In ITS, machine-readable **identifiers** are used to address the devices (in particular for the communication devices inside vehicles), which collect vehicular data. These identifiers can be eventually matched with the **identity** of a person, i.e. a personally linked name.

This might lead for example to actual or perceived privacy risks, as the vehicular data may be used

- for traffic law enforcement (e.g. speed violations, traffic accidents)
- for general law suits (employments, divorce, etc.)
- to derive embarrassing interpretation of location tracks (actual or wrongly interpreted visits of certain doctors, places, red light districts).
- to derive transport-specific personal information about driving styles, vehicle usage profiles, etc.
- to reveal general personal information about movement pattern, relationships, etc.

Therefore during the design and implementation of vehicular-related applications, privacy risks need to be assessed. Even so the actual risks might be limited compared to other technical systems impacting the privacy (such as mobile phone usage), the user acceptance of ITS applications also strongly relies on appropriate privacy protection mechanisms.



■ **Figure 2** Assessing privacy risk.

For the systematic privacy risk assessment and risk management of applications involving vehicular data, we introduce two dimensions of risks:

First, the **identity revelation risk** describes the risk and effort to match an identifier to an identity. As a second dimension, the **criticality of the data** is assessed. The data criticality indicates the level of perceived damage due to data revelation (e.g. level of detail, amount of detail).

In the following section, the methodology of the proposed risk assessment is introduced in more detail.

4.2.4 Privacy Assessment

In order to achieve better transparency of the privacy related risks that are caused by concrete implementations of vehicular applications we propose a simple scheme that provides a privacy risk vector. A risk vector fulfills two purposes:

- It supports application developers by helping them to understand the privacy implications of the system. They can use this information to reduce the privacy risk for their future users and, thus, make their product more attractive.
- It allows users, i.e. car drivers, to assess the risk of a specific applications without having to look into the details of the implementation. This transparency helps them to manually, semi-automatically, or even fully-automatically choose which applications they want to use.

Figure 2 illustrates the meaning of the risk vector. The two-dimensional characteristic of the diagram provides a quick overview on where the privacy risk comes from: It can be the revelation of your identity, the risk of making critical data available, or a combination of both. The length of the vector can be taken a very simple indicator for the overall risk caused by using an application.

■ **Table 1** Identity Revelation Risk Quantification.

Characteristic	Contribution to Value
no identifier used	none
short/medium/long-term identifier	+ / ++ / +++
(true) identity used	+++

■ **Table 2** Data Criticality Quantification.

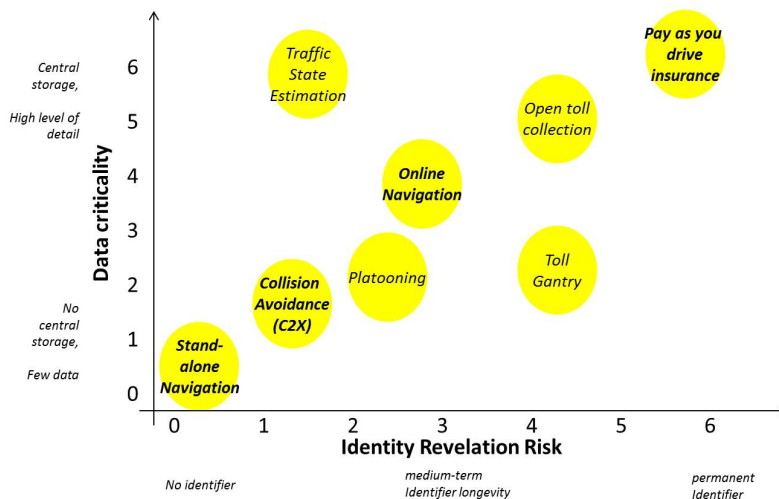
Characteristic	Contribution to Value
no storage of data / local storage / centralized storage	none / + / +++
size (level of detail) of data records	none / + / ++ / +++

The color scheme is intended to transport privacy expert knowledge to software developers and application users and intentionally kept simple: The well-known green, yellow, red known from traffic lights. Each user could, of course, define his/her own individual privacy risk boundary in this diagram, or maybe several context-dependent boundaries. Given this boundary a software solution could even automatically accept or reject the privacy risk induced by an application.

Quantification of the coordinates in the diagram is difficult but strongly desired, because it is a prerequisite for the calculation of a risk vector length. Therefore, we have analyzed a set of known vehicular applications with respect to the identity revelation risk and data criticality. Tables 1 and 2 show a proposal for mapping application characteristics to values:

Based on these tables the two coordinates can be calculated by summing up the values ('+'), resulting in a value from 0 to 6 on both axes.

4.2.5 Case Studies



4.2.6 Additional Material and Ideas

What about criminal activity? What judicial/policy controls are required?

Is there a privacy algebra? Does the direction of the privacy vector have meaning?

If users are given full control (i.e., free version with ads or user pays for no ads) what will be the effect on app development?

Do transport apps on their own pose a privacy issue or is cross-linking with other data the worry?

4.2.7 Discussion / Conclusions


Potential title of workshop paper “Assessing and managing privacy risks during design, implementation and operation of vehicular applications”

References

- 1 Daniel Ashbrook and Thad Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, October 2003.
- 2 Caitlin D. Cottrill. Approaches to privacy preservation in intelligent transportation systems and vehicle-infrastructure integration initiative. *Transportation Research Record: Journal of the Transportation Research Board*, 2129:9–15.
- 3 M. Gruteser and Xuan Liu. Protecting privacy, in continuous location-tracking applications. *Security Privacy, IEEE*, 2(2):28 – 34, mar-apr 2004.
- 4 B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Enhancing security and privacy in traffic-monitoring systems. *Pervasive Computing, IEEE*, 5(4):38 –46, oct.-dec. 2006.
- 5 D. Reid. An algorithm for tracking multiple targets. *Automatic Control, IEEE Transactions on*, 24(6):843 – 854, dec 1979.
- 6 R. Shokri, G. Theodorakopoulos, J. Le Boudec, and J. Hubaux. Quantifying location privacy. In *Security and Privacy (SP), 2011 IEEE Symposium on*, pages 247 –262, may 2011.
- 7 Hairuo Xie, L. Kulik, and E. Tanin. Privacy-aware traffic monitoring. *Intelligent Transportation Systems, IEEE Transactions on*, 11(1):61 –70, march 2010.

4.3 Working Group: Context-dependent Privacy in Mobile Systems

Florian Schaub, Maria Luisa Damiani, and Bradley Malin

License  Creative Commons BY-NC-ND 3.0 Unported license
© Florian Schaub, Maria Luisa Damiani, and Bradley Malin

4.3.1 Introduction

Mobile systems, and smartphones especially, are becoming ever more present in our society. A smartphone can gather a significant quantity information about its owner’s movements and behavior. Such *mobility data* can be communicated and shared with a wide range of parties (e.g., mobile applications, location-based services, mobile operators, mobile phone vendors and providers of mobile application platforms). The accumulation of mobility data enables the mining of mobility patterns, with the goals of patterns at the level of an individual and general trends. At the same time, there are concerns that the collection, storage, and processing of such data may violate the privacy of the individuals to whom the data corresponds.

Traditionally, privacy of such personal information has been achieved through definitional and methodological approaches. From a definitional perspective, access control is commonly applied to specify policies regarding what entities are permitted access to which information under a range of specified purposes and restrictions. From a methodological perspective,

data obfuscation and perturbation manipulate the data itself prior to its flow through an information system. Definitional and methodological aspects are not mutually exclusive and can be combined when the information flow is predefined or governed by policies.

To date, such privacy protection systems have often been *one size fits all* solutions. For example, in the case of k -anonymity [31] (i.e., the size of the group to which a piece of information corresponds), k is often fixed to a constant value for everyone in all situations. This is not always the case ([14]) and, more generally there exists a variety of solutions that enable users to configure their privacy settings. Yet, the personalization of such controls to a range of settings is a cumbersome process, which can lead to misconfiguration and dissonance between privacy expectations and their actual realization [27]. However, the context (e.g., “In what setting is the individual situated?”) provide meaningful indications about the level of required protection. While there is prior research on protecting contextual information, there are limited investigations into how such contextual information can be leveraged for enhancing privacy protection.

In this report, we discuss several perspectives on context-dependent privacy, with a focus on several application domains. Subsequently, we argue for a consolidated view of context-dependent privacy to facilitate the exchange of results between computational subcommunities. When appropriate, we outline a number of research challenges and future directions. The report closes with an outlook on ongoing and planned actions.

4.3.2 Perspectives on Context-dependent Privacy

Context-dependent privacy has been considered, to a limited extent, in certain subcommunities of computer science. We give an overview of the endeavors in three representative fields to highlight the different viewpoints and approaches taken.

4.3.2.1 Context-aware Privacy in Ubiquitous Computing

Ubiquitous computing (ubicom) centers on the vision of an interconnected world of smart devices integrated into a users’ surroundings and daily activities [33]. Some ubicom application areas are smart homes and environments, ambient assisted living, support in work environments, life logging, and everyday computing. Context awareness is a salient characteristic of ubicom applications. Applications use sensors to obtain information about the current situation of users, applications, and devices. Depending on the application, different context features are used to reason about the situation and detect user activities in order to adapt the application’s features and behavior accordingly. Dey broadly defines context as “any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves.” [11]. Commonly used context features [1] center around the *where* and *when* of the current situation and present entities (*who*). Furthermore, *what* the user is doing, as well as *why* they are doing it, are of particular interest. In addition to physically-sensed phenomena, contextual information includes information available (or inferred) about the user, such as data derived from a calendar, social network connections, and user preferences.

The connected nature and deep integration of technology with the environment in ubicom scenarios raises a number of privacy issues [21, 33, 29]. Specifically, the collection scale, often invisible data collection, and non-transparent information flow pose issues. The physicality of ubicom systems also extends the privacy scope beyond information collection to intrusions and disturbances in the user’s physical environment [20]. Furthermore, the complexity and

distributed nature of ubicomp systems makes it difficult for users to estimate the privacy implications of their actions and properly express their privacy expectations [27].

Context awareness has been utilized to improve the awareness of users about their current, as well as potential, privacy risks in ubicomp systems. For example, discovery protocols and detection mechanisms can be used to model what entities are present in a user's environment and what privacy-sensitive items they can access [21, 28, 20]. Context-aware privacy mechanisms, such as Confab [16], have been proposed to govern disclosure and granularity of specific information items. Privacy preferences are typically formalized in context-aware privacy policies. Moncrieff et al. [23] employ a context-aware privacy system in an ambient assisted living facility to dynamically govern what information is available to care givers. The information provided is dependent on the user's activities and potential hazards (e.g., a stove that is turned on and left unattended). User-centric privacy support systems use context awareness and heuristics in order to help users express their privacy preferences for specific situations [27]. Palen and Dourish [25] argue that, in order to ensure that the privacy provided by technical mechanisms matches a user's privacy expectations, privacy management needs to be viewed as a dynamic and continuous regulation process. Context awareness can enable ubicomp systems to support dynamic privacy adaptation either by 1) providing users with context-dependent individual recommendations or 2) automatically reconfiguring privacy mechanisms [28].

4.3.2.2 Context-based Location Privacy in LBS

Location-based services (LBS) comprise a wide spectrum of information services and applications (e.g., searching for business points of interest within a vicinity, sharing a personal location with the members of a social network). These applications typically rely on a client-server architecture in which clients convey their location to a LBS provider in exchange for spatially-contextualized information, returned in real time. Most current research on privacy in LBS is driven by the consideration that 1) location can reveal much about an individual and 2) LBS providers are potentially untrustworthy and, thus, can exploit location data for illegitimate purposes. In the literature, privacy requirements for LBS are often voiced from multiple perspectives, which calls for different classes of protection solutions. Historically, two broadly investigated privacy requirements are 1) identity privacy and 2) location privacy [17]. In the former case, the goal is to forestall the re-identification of seemingly anonymous users based on their location (e.g. home) [15]. In the latter case, the goal is to prevent the disclosure of detailed location or trace of a user (e.g., [35]). More recently, we emphasized the emergence of a third privacy requirement called behavioral privacy [8]. Behavioral information can be extracted from the traces of users by relating location with context. For example, context can reveal what the person may be doing (e.g., a person visiting a retail store is likely to be browsing or purchasing merchandise) or who they interact with (e.g., two people frequenting the same fitness club in the same period are likely to know each other). Preventing the extraction of behavioral information calls for techniques that are capable of recognizing geographical, temporal and social context.

Behavioral privacy is related to the concept of semantic location privacy and user-defined private places. Semantic location privacy refers to the capability of protecting semantic locations (i.e., the places in which users are located), such as a jewelry store [10]. In this scenario, a gentleman who never visits a jewelry store might be looking to surprise his girlfriend with an engagement ring (or some other token of affection) which he wishes to keep secret until the right time. The motivation behind this model is that places contribute more semantic information than the traditional coordinates (e.g., latitude and longitude)

commonly used to represent the user's location in traditional privacy-enhancing techniques. Moreover some places can be perceived more sensitive than others by users. For example, while standing on a crowded street may be considered innocuous, the revelation of one's presence in a jewelry store may represent a sensitive piece of knowledge. The protection of such information is challenging, particularly when users are tracked (i.e., their position is continuously reported to an LBS provider) and their movement is confined to road networks (i.e., a restricted set of trajectories) [34].

By contrast, the notion of a user-defined private place is motivated by the fact that many entities involved in the provision of location services are untrusted. These entities may be an LBS provider, but also third parties acting as location providers (LP), such as Google Location Service. Thus, a key question is, "To what extent can a user's location be protected against the party who computes it?" Placeprint [9] is a first attempt to protect the location information from both the LP and the LBS provider. In this system, users equipped with commodity devices can be geolocated in user-defined private places without revealing their presence to the LP. Additionally, users can specify context-based privacy rules to forestall the disclosure of private places also to LBS providers. The ultimate goal is to provide users with the capability of exercising flexible control over the disclosure of position information to LPs and LBS providers.

4.3.2.3 Context-Aware Access Control in E-Health Organizations

An individual's health status is considered to be among the most sensitive types of personal information. The collection and utilization of health information in the context of primary care (e.g., large hospitals) is not a new phenomenon. Electronic medical record (EMR) systems have been in place for over fifty years [32]. However, modern computing systems have enabled the integration of information into, and utilization of, EMRs at virtually anytime and anywhere. As a consequence, health information is now collected at all points of care; e.g., through hundreds of applications and IT systems in hospitals [24], home and remote health management systems (e.g., [2, 18, 19]), and through wearable devices (e.g., [5, 12]). The convergence of these systems is enabling mobility both in the collection of information and provision of treatment.

In many respects, the privacy issues associated with gathering and utilization of health information is not much different than other types of information. Yet, healthcare illustrates how context is related to more than just an individual and their personal preferences. Healthcare is a complex service-oriented business, composed of many interleaved processes. Consider, a patient may wish to specify which care providers can access their medical information where and for what purposes. Such specifications can be formalized in logical access control policies [22, 30]; however, the complexity of healthcare systems makes it difficult to define specific roles, as well as which care providers need access to what information and when [26].

For instance, healthcare is an inherently dynamic environment where teams of clinicians and support staff constantly interact. The notion of dynamic teams, while effective in supporting operations, leads to relationships that are significantly more complex pairwise than typically expected by access control frameworks [4]. The problem is further compounded by the fact that healthcare organizations are constantly evolving to update management protocols assimilate new employees and systems. Thus, the context of access to an individual's health information is critical to assess if the utilization is expected or not. As has been recently shown, this context may be inferred based on various features, such as the hospital service to which a patient is assigned or the location within a healthcare system the patient

is currently residing (which could be at home) [13, 36].

However, contextual features must also address the social dynamic of the organization [7, 6]. Healthcare professionals often collaborate in teams of employees with diverse skillsets. The likelihood that a particular team member will need (or choose to) access a patient's record is often dependent on the state of the patient (e.g., "Has the patient recently been admitted for treatment? Are they in a recovery room following a surgical procedure? Are they about to be discharged to an assisted living facility? Is the accessor of the patient's medical information a family member?"). Given that there are many organizational policies, most of which are not appropriately documented, it is critical to use audit logs and organizational knowledge to assess when access should be granted.

While this discussion focuses on healthcare environments, it is clear that this issue is more general. Similar issues clearly transpire in other domains, such as financial systems, intelligence environments, and just about anywhere where the relation between the subject (e.g., the patient) and the recipient (e.g., healthcare provider) is complex, dynamic, and driven by organizational behavior.

4.3.3 Discussion and Research Directions

Disparate subcommunities in computer science approach context-dependent privacy from very different angles. Context is used to facilitate dynamic and individualized adaptation of privacy mechanisms both 1) as a criterion for adjusting location and information granularity and 2) as domain semantics leveraged in access control systems. As a consequence of this diversity, proposed solutions, models, and mechanisms tend to be tailored for a specific perspective and have limited impact beyond the application domain in which they are initially introduced. However, in a preliminary analysis of problem characteristics, we noticed that, although framed differently, leveraging context for privacy protection raises similar research questions across domains. We believe that a consolidated view on context-dependent privacy could facilitate a common understanding of the associated research challenges to accelerate research and enhance the generality of results.

4.3.3.1 Shared Problem Characteristics

Despite diverse perspectives on context-dependent privacy all of the privacy mechanisms alluded to in Section 4.3.2 share common goals for a dual optimization: 1) minimize privacy risk for some information or some user and 2) maximize the utility of the information. Privacy risk can grossly be defined as the probability that, given the output of a privacy mechanism, a third party can infer the corresponding input (e.g., "To what extent can a user's exact location be ascertained from an obfuscated location?") Similarly, utility can be defined as the probability that the output can be utilized to support a specific purpose (e.g., "How similar are the results of a location-based service based on an obfuscated location and an exact location?").

Context can be invoked as an input parameter in the optimization to facilitate the determination of appropriate privacy mechanisms. Specifically, context can be used to determine the level of privacy necessary in a given situation or to perturb information in a manner that enhances utility. Additional inputs could be domain-specific constraints that restrict the abilities of individual users to determine privacy settings. Such constraints could be of an organizational or a regulatory nature.

4.3.3.2 Context Definition and Modeling

The benefits of considering context in privacy mechanisms and dynamic privacy adaptation are readily apparent. However it remains a challenge to identify which context information is relevant for privacy and how best to represent it. While there is extensive work on generic context modeling for application adaptation [3], the identification of privacy-relevant context features is often only performed for specific applications. Very few context models are specifically focused on privacy [28]. A general categorization and model for privacy-relevant context features would facilitate cross-utilization of privacy context across different domains.

When defining privacy-relevant context, it must be considered that context features and their values can depend on each other. Furthermore, context information and the data to be protected can be intertwined. For example, consider privacy mechanisms for applications that generate and process spatiotemporal data series, such as mobility tracking and vehicle-centric applications. These applications need to consider the context in which information has been generated (e.g., a traffic jam), but also ensure that other information types in the same series do not facilitate inference of the original data (e.g., rush hour).

A reasonable approach for representing privacy-relevant context is the definition of multiple context layers. Each layer could provide abstractions of context information on a different level. At the same time, each layer could represent semantics for specific domains (or provide granularity definitions for different information types). Modeling context on different levels is a common approach in context-aware systems, but the definition, management and structuring of context layers that appropriately capture generic and domain-specific privacy semantics, language, and knowledge is an open challenge. Context models must also be maintainable and scalable.

Context-dependent privacy requires reasoning about context information in order to adapt privacy mechanisms. This reasoning must extend over different context layers and be able to deal with uncertainty concerning the accuracy of context information. A related challenge is the development of formal models for context-dependent privacy that facilitate proofs of the provided privacy protection as well as the utility of information in the face of dynamic adaptation of privacy mechanisms.

4.3.3.3 Sensitivity of Context

An inherent challenge of context-dependent privacy is that the context information utilized for enhancing privacy adaptation is, in itself, privacy sensitive. This is because it contains potentially detailed information about the user's situation. Thus, privacy preserving context acquisition mechanisms are preferable to reduce privacy risk.

Beyond the sensitivity of actual context information, acting upon context can also potentially reveal information about the situation that caused the privacy adaptation. For example, consider a context-dependent location-based service that adapts location granularity based on context. When the user moves into a sensitive area, the location information becomes coarser, thereby revealing that the location is of higher sensitivity to the user than other locations. Thus, the sensitivity of a context-dependent privacy change must be considered, which requires respective mechanisms and models. One approach to address this issue could be to exploit historical context information to determine the probability that obfuscated information can be derived from previous information.

4.3.3.4 Making Privacy Mechanisms Context Aware

A further challenge is the integration of context awareness and context-dependent privacy into existing privacy and access control mechanisms. Although there is some work on context-aware access control, a challenge remains regarding how to account for context in privacy policies on a practical level.

4.3.4 Conclusions

Our investigation of context-dependent privacy in different domains identified a number of shared characteristics of privacy mechanisms utilizing context. However, there is only limited exchange and cross-utilization of results between sub-communities. Our preliminary analysis indicates that a generalized and formalized problem definition is potentially feasible. The benefits of such a model would be enhanced comparability and compatibility of approaches and results between different communities despite domain-specific requirements. We encourage the computer science community to explore the possibilities of such problem formalization in future collaborations.

Acknowledgements

We would like to thank all participants of the Dagstuhl Seminar on Mobility Data Mining and Privacy for stimulating discussions, valuable input, and feedback. We further extend our gratitude to the seminar organizers and the Dagstuhl staff for bringing us all together and making this seminar a very pleasant experience.

References


- 1 Gregory Abowd and Elizabeth Mynatt. Charting past, present, and future research in ubiquitous computing. *ACM TOCHI*, 7(1):29–58, 2000.
- 2 Hande Özgür Alemdar, Tim van Kasteren, and Cem Ersoy. Using active learning to allow activity recognition on a large scale. In *Proceedings of the 2nd International Joint Conference on Ambient Intelligence*, pages 105–114, 2011.
- 3 Claudio Bettini, Oliver Brdiczka, Karen Henriksen, Jadwiga Indulska, Daniela Nicklas, Anand Ranganathan, and Daniele Riboni. A survey of context modelling and reasoning techniques. *Pervasive and Mobile Computing*, 6(2):161–180, 2010.
- 4 B. Blobel, R. Nordberg, J. M. Davis, and P. Pharow. Modelling privilege management and access control. *Int J Med Inform*, 75(8):597–623, Aug 2006.
- 5 Shih-Lun Chen, Ho-Yin Lee, Chiung-An Chen, Hong-Yi Huang, and Ching-Hsing Luo. Wireless body sensor network with adaptive low-power design for biometrics and healthcare applications. *IEEE Systems Journal*, 3(4):398–409, 2009.
- 6 You Chen, Steve Nyemba, and Bradley Malin. Auditing medical record accesses via healthcare interaction networks. In *Proceedings of the American Medical Informatics Association Annual Symposium*, 2012. (in press).
- 7 You Chen, Steve Nyemba, and Bradley Malin. Detecting anomalous insiders in collaborative information systems. *IEEE Trans. Dependable and Secure Computing*, 9(3):332–344, 2012.
- 8 M. L. Damiani. Privacy enhancing techniques for the protection of mobility patterns in LBS: research issues and trends. In *European Data Protection: Coming of Age?* Springer, 2012. (to appear).
- 9 M. L. Damiani and M. Galbiati. Handling user-defined private contexts for location privacy in LBS. In *Proc. ACM GIS*, 2012. (to appear).
- 10 M.L. Damiani, C. Silvestri, and E. Bertino. Fine-grained cloaking of sensitive positions in location-sharing applications. *IEEE Pervasive Computing*, 10(4):64–72, 2011.

- 11 Anind K. Dey. Understanding and Using Context. *Pers Ubiquit Comput*, 5(1):4–7, 2001.
- 12 A. Dinh, D. Teng, L. Chen, S.B. Ko, Y. Shi, C. McCrosky, J. Basran, and V. del Bello-Hass. A wearable device for physical activity monitoring with built-in heart rate variability. In *Proc. 3rd Int. Conference on Bioinformatics and Biomedical Engineering (ICDBBE)*, 2009.
- 13 Daniel Fabbri, Kristen LeFevre, and David A. Hanauer. Explaining accesses to electronic health records. In *Proceedings of the 2011 ACM SIGKDD Workshop on Data Mining for Medicine and Healthcare*, pages 10–17, 2011.
- 14 Buğra Gedik and Ling Liu. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *IEEE Transactions on Mobile Computing*, 7(1):1–18, 2008.
- 15 M. Gruteser and D. Grunwald. Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In *Proc. MobiSys '03*. ACM Press, 2003.
- 16 Jason I Hong and James A Landay. An architecture for privacy-sensitive ubiquitous computing. In *Proc. MobiSys '04*. ACM Press, 2004.
- 17 C. S. Jensen, H. Lu, and M.L. Yiu. Location Privacy Techniques in Client-Server Architectures. In *Privacy in Location-Based Applications: Research Issues and Emerging Trends*. Springer-Verlag, 2009.
- 18 Shanshan Jiang, Yanchuan Cao, Sameer Iyengar, Philip Kuryloski, Roozbeh Jafari, Yuan Xue, Ruzena Bajcsy, and Stephen Wicker. Carenet: an integrated wireless sensor networking environment for remote healthcare. In *Proceedings of the ICST 3rd international conference on Body area networks*, BodyNets '08, 2008.
- 19 Andrew D. Jurik and Alfred C. Weaver. Remote medical monitoring. *IEEE Computer*, 41(4):96–99, 2008.
- 20 Bastian Könings and Florian Schaub. Territorial privacy in ubiquitous computing. In *Proc. Int. Conf. Wireless On-Demand Network Systems and Services (WONS '11)*. IEEE, 2011.
- 21 Marc Langheinrich. Privacy in Ubiquitous Computing. In John Krumm, editor, *Ubiquitous Computing Fundamentals*, chapter 3, pages 95–160. CRC Press, 2009.
- 22 Ming Li, Shucheng Yu, Kui Ren, and Wenjing Lou. Securing personal health records in cloud computing: Patient-centric and fine-grained data access control in multi-owner settings. In *Proceedings of the 6th International ICST Conference on Security and Privacy in Communication Networks*, pages 89–106, 2010.
- 23 Simon Moncrieff, Svetha Venkatesh, and Geoff West. Dynamic privacy assessment in a smart house environment using multimodal sensing. *ACM TOMCCAP*, 5(2):10, 2008.
- 24 National Research Council (US) Committee on Engaging the Computer Science Research Community in Health Care Informatics, W. Stead and H. Lin, eds. *Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions*. National Academies Press, Washington, DC, 2009.
- 25 Leysia Palen and Paul Dourish. Unpacking “privacy” for a networked world. In *Proc. Conf. on Human factors in computing systems (CHI '03)*, pages 129–136. ACM, 2003.
- 26 Lillian Røstad and Øystein Nytrø. Access control and integration of health care systems: An experience report and future challenges. In *Proceedings of the the 2nd International Conference on Availability, Reliability and Security*, pages 871–878, 2007.
- 27 Norman Sadeh, Jason Hong, Lorrie Cranor, Ian Fette, Patrick Kelley, Madhu Prabaker, and Jinghai Rao. Understanding and capturing people’s privacy policies in a mobile social networking application. *Personal and Ubiquitous Computing*, 13(6):401–412, 2009.
- 28 Florian Schaub, Bastian Könings, Stefan Dietzel, Michael Weber, and Frank Kargl. Privacy Context Model for Dynamic Privacy Adaptation in Ubiquitous Computing. In *6th Int. Workshop on Context-Awareness for Self-Managing Systems, UbiComp '12*. ACM, 2012.
- 29 Florian Schaub, Bastian Könings, Michael Weber, and Frank Kargl. Towards context adaptive privacy decisions in ubiquitous computing. In *PerCom 2012 Workshops*. IEEE, 2012.

- 30 Robert Steele and Kyongho Min. Healthpass: Fine-grained access control to portable personal health records. In *Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications*, pages 1012–1019, 2010.
- 31 Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, 10(5):557–570, 2002.
- 32 L. L. Weed. Medical records that guide and teach. *N. Engl. J. Med.*, 278(11):593–600, Mar 1968.
- 33 Mark Weiser. The computer for the 21st Century. *Sci. Am.*, 265(3):94–104, 1991.
- 34 E. Yigitoglu, M.L Damiani, O. Abul, and C. Silvestri. Privacy-preserving sharing of sensitive semantic locations under road-network constraints. In *Proc. IEEE MDM*, 2012.
- 35 Man Lung Yiu, C.S. Jensen, Xuegang Huang, and Hua Lu. SpaceTwist: Managing the Trade-Offs Among Location Privacy, Query Performance, and Query Accuracy in Mobile Services. In *proc. IEEE 24th International Conference on Data Engineering*, 2008.
- 36 Wen Zhang, Carl Gunter, David Liebovitz, Jian Tian, and Bradley Malin. Role prediction using electronic medical record system audits. In *Proceedings of the American Medical Informatics Association Annual Symposium*, pages 858–867, 2011.

4.4 Working Group: Privacy through Uncertainty in Location-Based Services

Nilgün Basalp, Joachim Biskup, Erik Buchmann, Chris Clifton, Bart Kuijpers, Walied Othman, and Erkay Savas

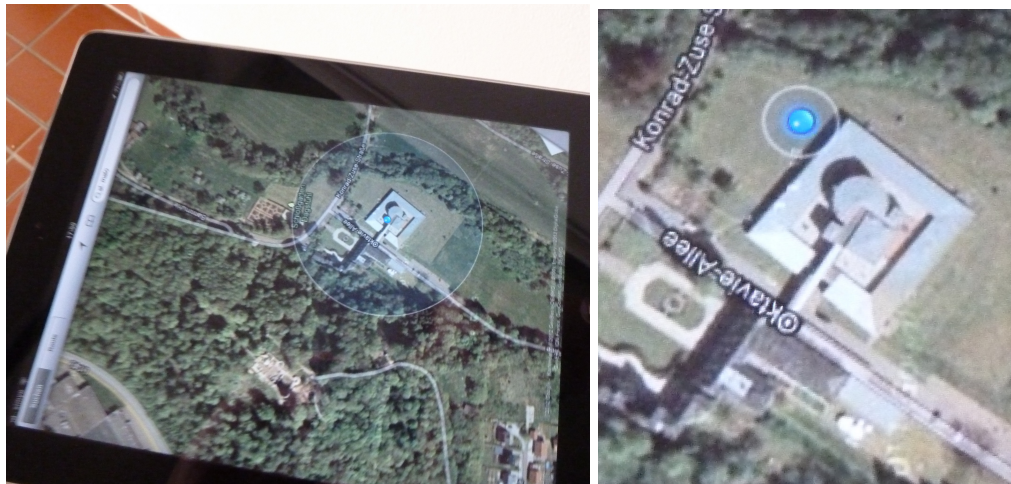
License  Creative Commons BY-NC-ND 3.0 Unported license
 © Nilgün Basalp, Joachim Biskup, Erik Buchmann, Chris Clifton, Bart Kuijpers, Walied Othman, and Erkay Savas

Location-Based Services (LBS) are becoming more prevalent. While there are many benefits, there are also real privacy risks. People are unwilling to give up the benefits – but can we reduce privacy risks without giving up on LBS entirely? This working group explored the possibility of introducing uncertainty into location information when using an LBS, so as to reduce privacy risk.

Uncertainty occurs naturally, so LBS likely to work in spite of uncertainty. For example, Figure 3 shows location determined by an Apple iPad at Schloss Dagstuhl. Initially, the location was reported with a high degree of uncertainty. Later, the uncertainty was reduced – but the location was not exactly as reported. A good LBS will have to accept that location may be uncertain, and give appropriate service in spite of that. Our question is, can we protect privacy by providing uncertain location, while still retaining good service?

4.4.1 Examples

In this section we explore some instances which have raised privacy issues in the past. One such case, which set off the privacy debate, is the case of Apple storing and collecting location data from its users' iPhones, unbeknownst to the user. The issue was uncovered on April 20th, 2011. Researchers discovered a file, consolidated.db, which contained longitudes and latitudes combined with a timestamp. This file contained locations that dated as far back as the release of iOS 4, which makes it contain a year's worth of location data, stored on the iPhone, synced (backed up) with iTunes and transmitted to Apple. All without the user's knowledge. A week later, Apple formally responded in a press release <http://www.apple.com/pr/library/2011/04/27Apple-Q-A-on-Location-Data.html> Apple



■ **Figure 3** Location uncertainty at Schloss Dagstuhl (actual location in hallway outside room N009).

maintained this data was anonymised, never shared with third parties, and its intent was to facilitate location look-ups by GPS in what is known as A-GPS, assisted GPS. Apple resolved the case with a software update that

- reduces the size of the crowd-sourced Wi-Fi hotspot and cell tower database cached on the iPhone,
- ceases backing up this cache, and
- deletes this cache entirely when Location Services is turned off.

In the time since, other apps have been uncovered to collect location (and other data) from mobile devices. These apps can be divided into two categories. The first category contains apps that collect location data but do not use it to provide a service, the second category uses it to provide a service. Ultimately, the goal is to provide techniques for users to protect themselves against tracking apps, while at the same time ensuring they can enjoy the same level of service from the apps. These two goals appear to be at odds with each other, but they do not need to be.

Currently, there are ways to protect yourself from being tracked. These do, however, require devices to be jailbroken (iOS) or rooted (Android). On iOS there is an app called ProtectMyPrivacy. It intercepts calls to a user's address book, playlists and location. If a user decided to hide any of these from an app that is requesting access, ProtectMyPrivacy jumbles the different fields for address book and playlist requests, and returns a fake, but preset by the user, location. This kind of protection serves well for apps that do not provide a service based on these records, but fails to ensure any service from apps in the second category.

There are a lot of research papers on such location-based social recommendation systems, and also some real systems. One particularly scary example (pointed out by Florian Schaub) is a “find a date” application that combines social media and location to find nearby women (who haven't necessarily said they want to be found):

<http://www.cultofmac.com/157641/this-creepy-app-isnt-just-stalking-women-without-their-knowledge-its-a-wake-up-call-about-facebook-privacy/>

Another examples is location-based keyword search: Modifying search results based on current location. While less seemingly privacy-invasive, this is also one where exact location is likely not needed.

4.4.2 Methods to induce uncertainty

The goal of this work is to increase the degree of uncertainty that comes from the location sensor in a way that the privacy of the individual is preserved. In particular, we strive for an amount of uncertainty that makes it impossible for an adversary to either (1) identify an individual from a particular location or a sequence of locations, (2) link a location or a sequence of locations to an individual or (3) connect a set of locations to a trajectory that belongs to the same (yet anonymous) individual. For that purpose, a wide number of anonymization approaches and obfuscation techniques exist. We subsume these techniques under the term "uncertainty methods". The applicability of uncertainty methods depends on the kind of data that needs to be modified. Thus, we distinguish between the induction of uncertainty to single locations and sequences of locations. While single locations are typical for one-shot queries sent to a location-based system, sequences of locations might be trajectories recorded by a smartphone with an activated GPS receiver or sequences of consecutive queries that have been sent from multiple positions to a location-based service.

The amount of uncertainty that is bound to a certain location or a trajectory depends on many factors. For example, if a pedestrian produces a location in the middle of a motorway or in a military exclusion area, an adversary might guess that this is implausible. Another example is a cyclist who has generated sequences of locations in distances that cannot be reached with the typical speed of a bicycle. We will address these issues in Section 4.4.6.

To ease our presentation, in the following we consider pairs of latitude, longitude only. However, all approaches described can be easily applied to more complex spatio-temporal settings by considering height and time as additional dimensions that are treated in the same way as latitude, longitude.

4.4.3 Obfuscation techniques

In general, obfuscation techniques can be applied by each user in isolation. One of the most intuitive techniques to increase the uncertainty of a location information is to add or multiply the latitude, longitude-record with a random numbers taken from a uniform distribution. The upper and lower bounds of the probability distribution function are a measure for the amount of uncertainty obtained.

A more sophisticated approach [1] takes the amount of uncertainty into account that has been already induced by the location sensor. In particular, the approach assumes that the correct position of a user is uniformly distributed over a circle that has been reported as center, radius by a GPS device. The radius specifies the accuracy of the location information. In this setting, uncertainty can be increased by shifting the center and enlarging or decreasing the radius.

Another way of obfuscation is the creation of a set of realistic dummies. With this approach, the user does not only sends a single position information to a location-based service, but a number of artificial dummy positions plus the real position. Accordingly, the service returns one result for each query. The client filters the set of results for that answer that corresponds to the real position. In this case, uncertainty is not defined as uncertainty towards a region (specified by a probability distribution), but as uncertainty towards which of the queried positions is the real one.

Finally, there exist approaches to replace latitude, longitude-pairs with the positions of prominent landmarks. For example, the exact position 49.530, 6.899 of a participant of a seminar in Dagstuhl could be reported as "Saarland", "Germany" or "Europe". The applicability of this kind of obfuscation depends on the format in which a location-based service requires a location information.

4.4.4 Anonymization approaches

Uncertainty by anonymization means to hide the identity and position of a user among a set of other users. Thus, anonymization requires the position information of multiple users. A popular approach is Spatio-Temporal Cloaking [2]. The approach adapts the concept of k-anonymity for geographic coordinates. For this purpose, the approach computes cliques of users that are close together, and releases minimum bounding rectangles that contain the positions of at least k different users each. Various variants of this concept exist, e.g. peer-to-peer-anonymization [2].

Mix zones are an approach to add uncertainty to spatiotemporal settings where the users are continuously observed by a service provider. The approach identifies each user by a pseudonym. Furthermore, it divides regions into mix zones and application zones. In predefined time intervals, all individuals within a mix zone have the option to choose a new pseudonym. Given the number of users in a mix zone is large enough, the service provider cannot link the movement of an individual in one application zone to the movement of the same individual in another application zone.

4.4.5 Legal status location privacy

The possibility to minimize violations of privacy can be achieved by creating uncertainty in location data while using location based services. An element in the definition of personal data in the EC Directive is that personal data indicates an identified or identifiable person. In other words the terms “identified or identifiable” focus on the conditions under which an individual should be considered as “identifiable”. In this regard the particular conditions of a specific case play an important role in this determination. Therefore the effect of uncertainty has to be addressed individually.

Location-based services in general process personal data in order to fulfill their contractual duties. The legal ground of using such information primarily is bound to the requirements of “informed consent” or “performance of a contractual duty” under EC Directive 95/46. Furthermore, a processing on a secondary basis requires the fulfillment of at least one of the exceptions under EC Directive such as the existence of a “legitimate interest” of the data processor or the existence of a “vital interest” of the data subject. In this context, the “legitimate interest of the data processor” criterion will be subject to further analysis. Especially the legal framework and – if any – case law will be pointed out.

4.4.6 Privacy analysis

One key challenge that must be addressed is how to analyze privacy. While there has been some research in this area (e.g., the talk by Maria Luisa Damiani 3.3), this is still a challenge with room for research.

We are considering the following agents: an application *provider* offering one or more dedicated services to a certain class of *clients*, which might be formed by subscription or even on an ad-hoc basis. Each of the clients might repeatedly request one of the services. Each request comes along with data about the requester’s location in order to enable the provider to return a location-dependent reaction to the requester. Without privacy-preserving measurements, the location would be determined in the best possible way, while however we are facing the possibility that there are technical failures leading to uncertainties regarding the actual precise position. In addition, each request is associated with a time stamp, which reflects real time by default. Introducing additional mechanisms for privacy-preservation enables a client to purposely generate further uncertainty about his or her actual location.

Though the time data might be blurred as well, for the sake of simplicity we leave this option open to future considerations. Accordingly, over the time the provider receives a sequences of requests, each consisting of at least the following components: location, time, kind of request.

Beyond supposedly honestly in each case actually providing the service requested, the provider may behave in a *curious* way, *analyzing* the collection of request data received so far for the sake of any secondary use. We identified three major basic kinds of analysis. Moreover we emphasized the following need of a client: in order to decide about the employment of uncertainty generating mechanisms, a requesting client has to evaluate the potentials for a successful analysis by the provider according to some metrics.

- Location-based reidentifiability:

In case that requests are coming without the respective requester's identification, the provider might aim to associate an identifier to each of the requests, at the best definitely the identifier of the actual requester; alternatively and less ambitiously, some assertion about the relationship between the request and the clients. As far as the provider succeeds in establishing a nontrivial and meaningful association, he would either learn precise personal data or obtain data that is somehow potentially personal and, thus, he would be able to compromise privacy to some extent.

Seen from the point of view of a client acting as requester, that client would be interested to estimate the extent of compromise achievable by the provider according to some suitable metric.

- Location identification and classification:

Whether by technical failures or by intentional blurring, a communicated location might differ from the actual one. Accordingly, the provider might aim to determine the actual location by some kind of reasoning, thereby strengthening his knowledge about the requester. As far as the requester is already identified, in this way the provider would obtain improved personal data regarding the requester. If the requester is so far not fully identified, more precise knowledge about the actual location might be helpful for the reidentification analysis or other analysis tasks.

Besides the pure geographical data about the location, the provider might also aim at determining the kind of social activities offered at the respective place and thus learning information of the requester's activities, again potentially leading to even more crucial personal data. Typically, social activities could be classified and denominated according to some ontology, e.g., distinguishing between shopping, medical care, entertainment, food services, sports, education, and so on, even possibly refined to subcategories and enhanced by further descriptive features.

Again, the client would like to evaluate the expected achievements, in particular in terms of the grade of success and the sensitivity of an identified location depending on its semantics, according to some metric.

- Subtrajectory linkage

While strictly speaking a client only communicates location-time points, she or he actually provides information about her or his movements over the time, i.e., about the resulting trajectory. Accordingly, the provider might aim at reconstructing the actual trajectory in an approximative way in order to learn more about the client. As before, besides the pure geographical data about the full trajectory, he might additionally be interested in the semantics of the curve in terms of a suitable ontology that extends the ontology for single locations.

Reconstructing an actual trajectory necessarily includes linking single locations as communicated and subtrajectories obtained before as belonging to the same client. This need

is clearly supported by already knowing the association of the requests with identifiers, but also conversely, if originally unknown learning this association might be facilitated by having established links before.

Again, the client would like to evaluate the expected achievements, in particular in terms of the grade of success and the sensitivity of a reconstructed trajectory depending on its semantics, according to some metric.

4.4.7 Parameters, how relevant, and how to estimate

The success of the analyses described above and in particular the quality of the results achieved depend on a wide range of parameters. Such parameters might refer to both the data provided by the requests and various kinds of background knowledge available to or even generated by the provider. In this subsection we briefly discuss some important examples.

- Plausible locations and their semantics:
Given a region in purely geographical terms, e.g., described by a circle, the provider is likely to possess one or more maps including this circle and its further environment as a priori background knowledge. Each such map provides some kind of semantics, in particular suggesting plausible actual positions within the region and a classification of many objects within the region.
- Density and properties of other clients:
Given data that indicates that some client, whether identified or not, stays within a region, the provider might also have gained the knowledge that other clients are staying near by. This knowledge might have various effects. For example, the existence of many unidentified clients within a region results in forming an anonymity class for the client considered, and this fact might support the client's privacy concerns. Conversely, having learned the semantics of the stay of sufficiently many other clients might suggest a semantics for the client considered.
- History-frequency of appearance/past revelation:
Analyses might not only be based on the data of the most recent requests but also refer to histories and their evaluations. For example, previously successfully identified and classified trajectories might be related to a recent trajectory under inspection, suggesting an identification or semantics based on similarities. Useful notions of similarity and the resulting suggestions might have been obtained by means of data mining applied to the data received before and stored in a repository.

4.4.8 QoS analysis

The degradation in quality of service from intentionally giving uncertain locations is a critical issue. While this is largely dependent on the particular application, and the implementation of that service provider, we can experimentally derive quality of service measures. By comparing the results from requesting a service with the actual location at the best anticipated resolution with the results from intentionally degraded uncertainty, we can establish how great the impact of a given uncertainty method is on the particular application.

While metrics are also application dependent, many location-based services return ranked lists. Examples include location-based keyword search, recommendation systems, closest point of interest, driving route finders, and public transit schedule systems. In all these cases, we can compare the impact of uncertainty by comparing the results with uncertain location against the results with actual location. There are standard ways to do this, such as KL-Divergence[3].

We plan to choose specific applications from Section 4.4.1, and evaluate against all relevant techniques described in Section 4.4.2 at various levels of uncertainty. Levels of uncertainty will be chosen to achieve interesting privacy points as determined in Section 4.4.6. Results will be presented by graphing the KL-Divergence across various parameter changes.

4.4.9 Secondary use utility analysis

Can we relate type of uncertainty to impact on classes of mining.

- What is the effect of the user-injected uncertainty in location on aggregate data collected on service provider side? Can we model it mathematically?
- The common business practice is that the service providers use whatever the user provides during the service usage. What law/regulations stipulate for the secondary use does not necessarily reflect current practice. We do not intend to change it. In fact, in certain circumstances the secondary process can create benefits.
- Following up on the previous point, the service provider may not be able to provide the adequate protection for the data submitted by the users which is subject to theft/compromise by other parties who will do the secondary (mis-)use.
- Can service provider learn more about personal information that intended after the aggregated data is obtained? To what extent? Can it be the case the uncertainty turns out to be not a worthwhile effort?
- Find a good example as a motivation for the case parties find secondary processing useful/beneficial/unharmful
- Experimental work can be useful to compare the data mining results extracted from aggregate data with uncertainty against those without.

4.4.10 Questions that remain to be answered

4.4.10.1 Other personal data

Location data isn't the only data exposed to location-based services.

- A query may contain some other personal information, which is necessary for the completion of service. What are the implications for privacy? Can a service provider get more information than intended?
- Does user provide his/her identity in the query? Does s/he query anonymously?
- What happens if the service requires tracking of user for a certain time interval?

4.4.10.2 Safety applications

Systems would need to be designed to bypass the uncertainty when safety of individual necessitates accurate location information

4.4.10.3 Acceptance/Feasibility

There are also issues concerning practical application of uncertainty. Are service providers able and willing to work with inaccurate data? Can the proposed method be implemented with current technology? (This seems feasible for Android, perhaps not for Apple.)

4.4.10.4 Adverse affects

False location declaration can lead to undesired situations; for instance putting a person in potentially problematic locations (e.g., in a crime scene)

References

- 1 Claudio Agostino Ardagna, Marco Cremonini, Ernesto Damiani, Sabrina De Capitani di Vimercati, and Pierangela Samarati. Location privacy protection through obfuscation-based techniques. In *Data and Applications Security XXI, 21st Annual IFIP WG 11.3 Working Conference on Data and Applications Security*, Redondo Beach, CA, USA, July 8-11 2007.
- 2 Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of First ACM/USENIX International Conference on Mobile Systems, Applications, and Services (MobiSys)*, May 2003.
- 3 Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, April 2004.

5 Open Problems

In addition to the developments and plans for future work in the working group reports, the seminar saw general lessons learned and future directions for Mobility Data Mining and Privacy.

5.1 What we learned

- Privacy issues in mobile data are real
 - Applications collect much more data than needed
 - Serious potential for harm, some evidence of actual harm
- Privacy violations may have direct societal impact
 - Not just the individual who is harmed
- Data previously highly regulated moving to lightly/unregulated businesses
- Wide variety of vocabularies used to talk about mobility, data mining, privacy
 - Poor understanding of cross-community issues by different subcommunities
- Dimensionality of uncertainty/unknowns
- Problems not clearly defined (or clearly understood even by experts)
- Age-related bias in assessment of issues

5.2 New Discoveries

- Danger of location data plus additional information gives supra-linear increase in risk
 - This linkage needs further investigation: Re-identification attacks, potential for damage, ...
- Sensors / content in smart phones lead to severely increased privacy risk
 - Lots of information
 - Easily revealed
 - Often disclosed with little awareness on part of individual
- Privacy preferences/expectations are dynamic
 - Need approaches beyond static privacy preference options
 - Location and time key components of this dynamism

- Need user studies on mobility and privacy
- Re-identification, particularly of mobile data, is too hard to prevent
 - Need broader perspectives on how to protect privacy
 - Disconnect between technology developments and law/regulation
- Technology developments are resulting in even “technology-neutral” privacy law becoming outdated

5.3 What needs to be done

- How to educate
 - Privacy technologists don’t understand law and regulation
 - Regulators don’t understand technical privacy definitions
 - Users don’t understand either
 - Application (and infrastructure) developers don’t worry about privacy
 - Curricula for privacy – not just “how to comply”
- Technology needs better guidance from law and regulation
 - And back to education – need to understand guidance
- Need to be suitably proactive
 - don’t want to let the cat out of the bag
 - Need for people to desire privacy beyond legal mandates
 - Need tools to manage (most) privacy that are general – user specifies their privacy preferences, not what they want a particular app to do.
- How do we manage cross-border data on the internet?
- Privacy research needs better understanding of economic value of data
- Venue/mechanism to better support multidisciplinary work to bring privacy into various communities

5.4 Future plans

Based on the above conclusions, the seminar participants felt that we need future study and discussion in this area. One option would be a future Dagstuhl seminar, as well as a more standard conference or workshop. More immediately, we felt that a panel would be appropriate, but this needs to be in a venue that reaches to the mobile data and location based computing community, not to the privacy community.

The study groups are pursuing further research and publication of the ideas from the seminar. However, a consolidated white paper on privacy and mobility, possibly with an accompanying tutorial, may also be a way to further raise the issues. The area also needs a study of the broader societal impact of privacy breach – social mores, economic issues, security implications – this may be a good topic to target an international grant proposal.

Participants

- Gennady Andrienko
Fraunhofer IAIS –
St. Augustin, DE
- Nilgün Basalp
Istanbul Bilgi University, TR
- Joachim Biskup
TU Dortmund, DE
- Erik Buchmann
KIT – Karlsruhe Institute of
Technology, DE
- Christopher W. Clifton
Purdue University, US
- Maria Luisa Damiani
University of Milano, IT
- Glenn Geers
NICTA – Kensington, AU
- Aris Gkoulalas-Divanis
IBM Research – Dublin, IE
- Marco Gruteser
Rutgers Univ. –
New Brunswick, US
- Christine Körner
Fraunhofer IAIS –
St. Augustin, DE
- Bart Kuijpers
Hasselt University –
Diepenbeek, BE
- Thomas Liebig
Fraunhofer IAIS –
St. Augustin, DE
- Bradley Malin
Vanderbilt University, US
- Michael Marhöfer
Oberhaching, DE
- Walied Othman
Universität Zürich, CH
- Klaus Rechert
Universität Freiburg, DE
- Claudia Santa
TU München, DE
- ErKay Savas
Sabanci University –
Istanbul, TR
- Florian Schaub
Universität Ulm, DE
- Olaf Spinczyk
TU Dortmund, DE
- Christian Wietfeld
TU Dortmund, DE
- Ouri E. Wolfson
University of Illinois –
Chicago, US

