Mind your Step! How Profiling Location reveals your Identity – and how you prepare for it.

Lothar Fritsch

***Abstract*—Location-based services (LBS) are services that position your mobile phone to provide some context-based service for you. Some of these services – called 'location tracking' applications - need frequent updates of the current position to decide whether a service should be initiated. Thus, internet-based systems will continuously collect and process the location in relationship to a personal context of an identified customer. This paper will present the concept of location as part of a person's identity. I will conceptualize location in information systems and relate it to concepts like privacy, geographical information systems and surveillance. The talk will present how the knowledge of a person's private life and identity can be enhanced with data mining technologies on location profiles and movement patterns. Finally, some first concepts about protecting location information to prevent or control location profiling are presented.**

***Index Terms*—communication systems privacy, location, privacy, mobility, profiling, maps, GIS.**

I. Introduction

LOCATION data at first may seem trivial. Where ever I go, there I am. I am, where I am. Location at first is a tuple of coordinates on a two-dimensional or three-dimensional grid, defining a position unambiguously, e.g. in the WGS-84 standard which defines the coordinate grid used on the planet Earth [1].

A. Location, Privacy and Identity

Some assumptions about location and identity seem trivial, too: At nighttime, my location usually is in the same place my home is in. During work days, my location is in my workplace. But there is more. Location determines belonging to social groups, from which social status can be derived. Sociologist Gary Marx defines location as a part of human being's identity in [2], which has been applied to LBS and privacy in [3], where the question of volatility and stability of profile information is raised. Sholtz assumed in [4] that there is little long-time value in profile information from an economic point of view. A risk point of view might be different, though. Formally, Gruteser and Grunwald sketched some threats in [5], where they postulate that from tracking a person's frequent nighttime location, they can guess her identity by looking it up in public phone directories. This will be discussed further down in the text.

B. Geographical Information Systems (GIS)

Combined with geographical information systems (GIS), many contexts of a place can be found out about – ranging from the neighborhood up to criminality of risk levels of natural disasters. Michael Curry states concern about this and calls for ethical standards in [6]. Recent product deployments like 'Google Earth' put these tools into the hands of the general public [7]. Ethically problematic is not only the privacy-invading character, but also the possibilities to manipulate and misinform with maps, as described in [8]. Privacy, GIS and positioning technology in combination can be even more invasive, as high lightened in [9]. Technologies involved in such privacy threats can be put in these categories:

- Position tracking technologies
- Data warehouse technologies
- Geographical data bases & information systems (GIS)
- Metadata bases with geo-coded data

Besides these factors, some a-priori knowledge about contexts of everyday life, holidays, social conventions and such can be an input to the above information systems.

C. Data Mining and its Applications

Data mining has been defined by many authors. Two definitions are presented here for clarity:

1. Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [10].;
2. Data mining is the search for relationships and global patterns that exist in large databases, but are `hidden' among the vast amounts of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and objects in the database and, if the database is a faithful mirror, of the real world registered by the database [11].

Data Mining evolved with relational data bases. Most of the original data mining work centered on processing relations and attributes in such data bases. Classic applications are fraud detection in financial or insurance matters or warehouse optimization based on customers' preferred buying patterns. Threats to privacy from data mining on customer databases have been well-discussed. An overview of privacy threats and ethical questions in customer data collection and profiling can be found in [12].

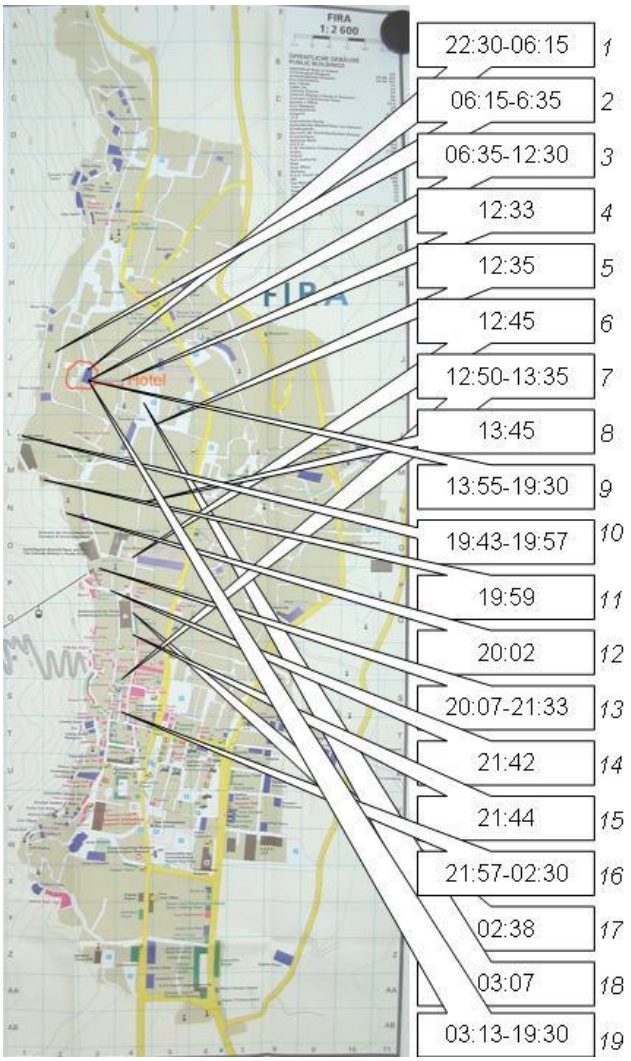

Figure 1: Map of Fira with location track and time stamps of a hypothetical person.

Figure 2: Map of Fira with location track and location marks.

Data mining technologies are now available for different kinds of geographic information. Classification of land surfaces based on satellite intelligence and mining of meta data layers of GIS are two examples. Meta data layers such as crime rate and wealth, classified by area, are in use for years to get customer scoring and to fight mail-order fraud. Land surveying with satellite data might reveal how a farmer cares for his land – or whether he uses the right amount of fertilizer. The concept is called "precision farming".

D. Profiling

Profiling a person is more than just data matching. Roger Clarke defines profiling as follows: "Profiling is a data surveillance technique which is little-understood and ill-documented, but increasingly used. It is a means of generating suspects or prospects from within a large population, and involves inferring a set of characteristics of a particular class of person from past experience, then searching data-holdings for individuals with a close fit to that set of characteristics." [13]

Thus, profiling targets at the selection of candidates out of the mass that have particular characteristics, or finding patterns over people's data that can be applied to find similar people.

As an example for the application of the techniques from sections A to D, we take a look at the location track of a hypothetical person's day on Santorini island. A compressed track is shown in **Figure 1**. For simplicity, the location recordings of the person while staying at the same place are noted as a time interval. Only selected positions are shown in **Figure 1**. For the start, the position data is time stamped (hence the time marks to the right), and then combined with GIS data containing road information and points of interest to create a map. From the road information, we can deduct with high probability the preferred roads or paths of the person, as you can spot in **Figure 2**. Additionally, by noticing that the person was using the area marked as "hotel" several times a day, one can guess that there must be the person's home. Looking at the time stamps adds more information. When you look at **Figure 1**, the information in timestamp 1 and 19 clearly indicates that the person sleeps at the hotel location. Also, by observing timestamps 3 and 9, the person spends most of the daytime at the hotel, too. One can guess that the person either is en em-

ployee of the hotel, or has some event to participate in going on. Meal times reveal that for lunch and dinner, the person left the hotel to spend time at two locations in the old town. Using the POI database on the GIS, we find the candidate restaurants. What does this tell about nutrition habits, religion, health and budget of the person? And can we buy the credit card transaction data for these restaurants from Amex to learn more? Also notice the dinner time path was on the edge of the cliff, unlike the lunchtime path. The cliff points west into the sunset. Was the person alone for the romantic view then? Timestamp 16 reveals a long stay at – thanks, GIS – an old town dancing club. Dancing obviously went on for a long time. So no company on the cliff, most likely. Or bad table manners. Or both. Timestamps 17 and 18 indicate slow progress back to the hotel, compared to the lunchtime progress. Is this due to intoxicating beverages, or due to company? Here, timestamp 19 does not help. Unfortunately, the full day at the hotel can be credited to work, hangover or company alike. Only the minibar billing on the credit card will most likely tell.

In this example, we used a time stamped location track, a GIS with some POI data, and some commonsense to create a behavioral profile of a person. We were able to guess some context, and find interfaces to other databases that will elaborate our knowledge.

II. Contextual Location Profiling

A. Spatial and temporal dimensions of location tracking

Location data can be analyzed at singular points in space or time as well as in intervals of either of them or both. The extent of analysis happens along the time axis or within geographic boundaries which I call dimensions. The respective gain or information or relating risk for the person being observed differs greatly based on the spatial and temporal dimensions. Generally, location tracking applications seem to be perceived more threatening than one-time positioning of person, as Barkhuus notes in [14]. A matrix of temporal and spatial dimensions is constructed from this in Table 1, where the matrix illustrates the kind of information that can be deducted about a person from the respective dimension category.

		Spatial dimension	
		At one point	*Within an area*
Temporal dimension	*At one moment*	**Singular:** Know about the status quo of time and space at one moment.	**Spatial snapshot:** For individuals: Makes no sense as one can be in one place at one moment only. For groups: can reveal relation-ships, cliques, collaboration.
	Within a time window	**Time-linear:** Can reveal workplace, home, social context and information about personal preferences (e.g. restaurant type).	**Two-dimensional:** Reveals shopping habits, dating habits, driving speeds and other information.

Table 1: Personal information deductible from temporal and spatial dimensions.

B. Context acquisition from temporal and spatial data

Personal information can be gained starting from knowledge about a person's identifier (not identity!) and her location track by applying temporal and spatial context to it. Temporal context is for example a person's time zone, along with time-dependant social habits, e.g. lunch breaks, Spanish siesta or night shift working. Spatial context can be derived from geographic information systems, which contain many layers of information much more specialized than mapping or navigation data. As described by Curry [6] and Monmonnier [9], existing GIS data layers include information about crime rates, property value, wealth, health, education, employment, race, pollution, noise, natural hazards and much more. Context acquisition follows an algorithmic scheme I construct in this way[1]:

1. Collect time-coded location data for some time (preferably days or weeks to operate in the two-dimensional category in Table 1).
2. Construct some temporal context of interest (e.g. private time vs. job time).
3. Check for a geographical pattern of interest in the location data track (e.g. places frequently visited, or unusual places rarely visited etc.)
4. Extract geographic coordinates along with their spatial and temporal information.
5. Query geographic information systems about locations, and extract meta data (e.g. "…is an office building").
6. Conclude from temporal context, spatial context and geographic meta data, e.g. the workplace, the home place, sports, and other personal data.

Notably, this algorithm works without knowing the person. It is enough to be able to re-identify her in the data set. It can be used for example on WiFi hot spot or mobile phone tracks that leave unique technical parameters as identifiable information. Using the algorithm above, we learn much about a - yet unknown – person's preferences and frequent behavior.

Next, the old-fashioned data bases enter the stage.

C. Data Mining, Combination and Profiling

As defined above, data mining follows relations in relational data bases to find "knowledge". Various methods from disciplines such as artificial intelligence, statistics, computational linguistics and stochastic methods are deployed on collections of data. Correlation between shopping habits and location, communication habits and location, movement and health data, social contexts and other can be mined from the data bases.

The application of profiling techniques on geographic and data base information could lead to new kinds of marketing, insurance or anti-terrorism systems. People's movement pat-

[1] The algorithm is inspired by Usama Fayyad's and Evangelos Simoudis' knowledge discovery algorithm presented in their 1995 tutorial "Knowledge Discovery and Data Mining" on the 14th International Joint Conference on Artificial Intelligence (IJCAI-95). Documentation can be found online at www-aig.jpl.nasa.gov/public/kdd95/tutorials/IJCAI95-tutorial.html

terns combined with other features could be used to segment customers, generate health insurance conditions or arrest suspects.

III. Countermeasures & Self-Protection

The protection against geo-coded data mining and profiling has three components: identity protection, camouflage and a legal and social framework for technology regulation. Each of the items will be discussed below.

A. Identity protection

A person's identity should be protected while using location-based services. If the location track is not personalized, it cannot be combined with any other data. Furthermore, the amount of data than can be accumulated about a person should be limited to avoid identity guessing from movement patterns. Identity management systems with frequent pseudonym changes and anonymous access to services provide to reach these goals. The simplest form of identity management occurs in Figure 3, where the mobile operator offers pseudonym translation services for the user before the application data traffic is forwarded into MIX cascades.

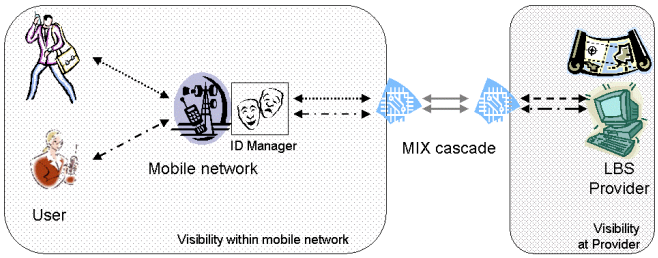

Figure 3: Minimalist Identity Management approach.

More advanced identity management approaches introduce policy management. Here, a user can set policy about location forwarding at her mobile operator, and at the same time issue anonymous credentials that identify the policy. The credential then is given to the LBS provider as a voucher Please note that naïve use of pseudonym change mechanisms can reveal all your pseudonyms used with a service though, as illustrated in Figure 4. To prevent this from happening, MIX-Zoning will be discussed later in this text (see Figure 7).

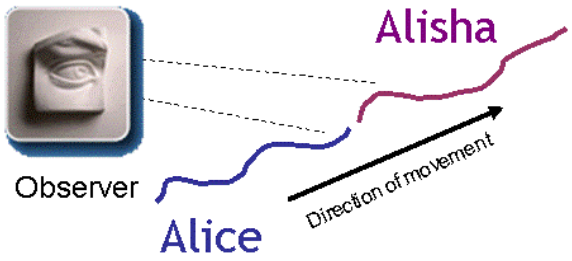

Figure 4: Naive pseudonym change reveals pseudonym connection.

B. Camouflage

The generation of less precise or even false information about identity, identifiable movement patterns and time or space disturbances will provide to protection from unauthorized geographic profiling. Early concepts have been suggested by Gruteser and Grunwald in [5]. Concepts include:

- Temporal cloaking: the time intervals for location queries are regulated to avoid micro-measurement of a user's position. The concept is illustrated in Figure 5.

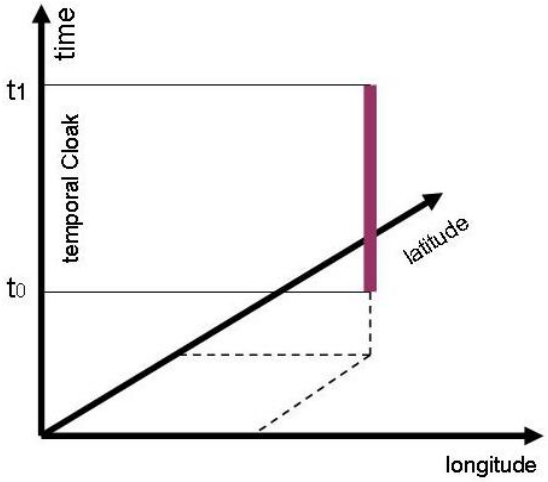

Figure 5: Temporal cloaking.

- Spatial cloaking: the precision of location information is reduced to a level tolerable by the application, but will not be delivered too precise. This intentional degradation of position precision prevents too precise information collection about a person's movements when tracking on a high-resolution level. Spatial cloaking is illustrated in Figure 6 below.

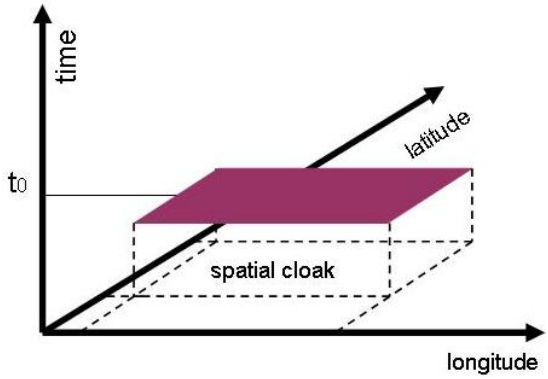

Figure 6: Spatial cloaking.

- MIX-zoning: To allow for unobservable change of pseudonyms (and solve the problem from Figure 4), a zone of unobservability is created where users can go to perform their pseudonym change. As soon as many users do this simultaneously, an anonymity set is created. The concept is inspired by Chaum's MIX [15]. An example of MIX zoning for the purpose of pseudonym change protection is shown in Figure 7.

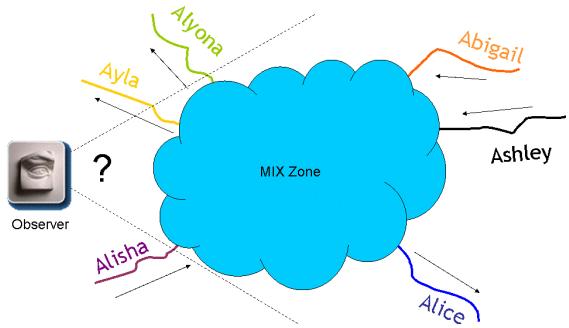

Figure 7: A MIX zone.

Temporal cloaking adds uncertainty to the point in time the position or a person was measured. The relying service using the data does receive a position datum, but only knows this is not the person's current position but from some time in the past. The usefulness of this approach is limited in terms of privacy protection. Only in contexts where a service tracks a person frequently (e.g. a pollen warning scenario, as used in [16]), but with coarse requirements on resolution and timing, temporal cloaking seems applicable. Spatial cloaking is effective in circumstances where a tracking service doesn't require high-resolution position information (e.g. for pollen warning). Here, the information is intentionally degraded to a degree where no daily routine is contained anymore.

Location dummy traffic: MIX zones are effective to protect and obfuscate pseudonym changing event. Unfortunately, MIX zones might not always have enough people in them just at the moment when they are used. To improve on this problem, the concept of dummy traffic in MIX communication can be adapted. Location track dummy traffic is performed with dummy users that are artificially generated location tracks with a certain non-compromising behaviour. The dummy pseudonyms are registered with the LBS application, and will be used for pseudonym changes.

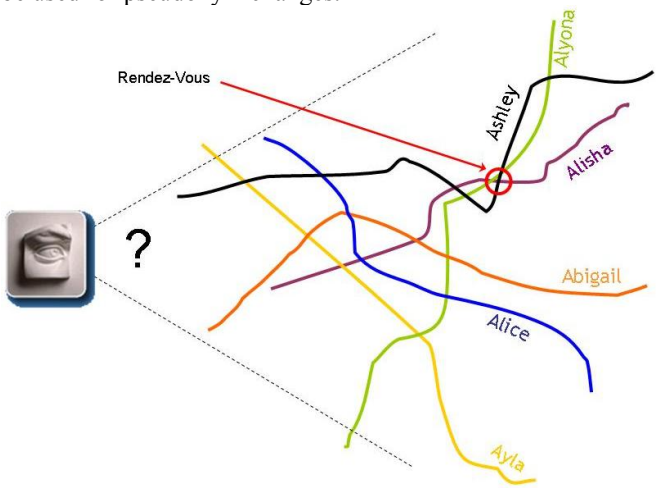

Figure 8: Location dummy traffic, Rendez-vous point.

When a user wishes to change to a different pseudonym, the dummy system ensures that some of his alternative or dummy pseudonyms will cross the user's path at a rendez-vous point, where the change will happen. The now unused pseudonym takes up a dummy life of its own, in temporary or permanent continuation of the previous path. This mechanism can take up a used pseudonym and carry it around the town virtually. The challenge here is the generation of realistic movement patterns that do not compromise the pseudonym owner by, e.g., entering the town's red light district. The application of this protection measure is restricted to LBS infrastructures that allow for injection of artificially created position data (e.g. the GPS device scenario or some special instance of the intermediary scenario described in [3]).

C. Legal and social framework

Finally, a legal and social framework is required on top of technical infrastructures. To make sense of innovations like the LBS applications with privacy protection, possibly based on trusted platforms that can enforce system constraints, a legal and social environment for the technologies must be created. Regulation is a controversial topic, but for many fields that are problems in other parts of society, there is evidence that regulation does not only complicate markets, but also provides to creation of new markets. For an outlook on how economic theory can favour regulation of pollution problems, and possibly the privacy problem, Paul Sholtz presented some interesting examples in [17] and [4]. To enable market participants to distinguish the quality of systems and parties, a public quality certification scheme can prove useful, as Backhouse et al have found in [18].

IV. Conclusion

Summarizing the challenges posed by combined data mining and geographic data mining, applied to location tracks, infrastructures for location-based applications have to offer reliable functionality to prevent misuse of personal information. As seen in section III, the development of technical countermeasures is in a very early phase of maturity. Like research in privacy-enhancing technology, it will take time for technology to mature, develop end-user usability and make its way into the application systems. With the possible countermeasures, users will have to rely on other people's IT systems correct functioning. Thus, with the development of geographic data mining technologies and LBS, there is a strong need to research and specify ways to develop privacy-respecting infrastructures (PRI). PRI – unlike the self-protecting focus of privacy-enhancing technologies (PET) – should be integrated within a regulatory, economic and technological framework. Some ideas of possible frameworks have been explored in [16] by the deployment of PET on all nodes of a distributed LBS scenario under consideration of the existing legal framework and end-user centric research. Generally, data protection policy of the future should take trusted platform technology under consideration. This form of restrictive computing base could ensure correct, fair function of LBS systems and geographic data bases, and generate assurance with a certification system.

REFERENCES

[1] EUROCONTROL, WGS-84 Implementation Manual. Bruxelles, Belgium: 1998.
[2] G. Marx, "What's in a name?" The Information Society. vol. 15, pp. 2 1999.
[3] L. Fritsch, Economic Location-Based Services, Privacy and the Relationship to Identity, 1st FIDIS Doctoral Consortium, IST FIDIS Network of Excellence, 2005, Riezlern, Austria.
[4] P. Sholtz, "Economics of Personal Information Exchange," First Monday. vol. 9, pp. 5 2003.
[5] M. Gruteser and D. Grunwald, Anonymous usage of location-based services through spatial and temporal cloaking, First International Conference on Mobile Systems, Applications, and Services (MobiSys'03), 2003,.
[6] M. R. Curry, In plain and open view: Geographic information systems and the problem of privacy. Los Angeles, California:.
[7] Google, Google Earth Webservice and Geographic Information System. 2005.
[8] M. Monmonnier, How to lie with maps. Chicago: University of Chicago Press,1996.
[9] M. Monmonnier, Spying with Maps : Surveillance Technologies and the Future of Privacy. Chicago: University of Chicago Publishers,2004.
[10] W. Frawley, G. Piatetsky-Shapiro and C. Matheus, "Knowledge discovery in databases: An Overview," AI Magazine. vol. 13, pp. 57-70, 3 1992.
[11] M. Holsheimer and A. Siebes, Data Mining: The search for knowledge in data bases. Amsterdam: 1991.
[12] E. R. Foxman and P. Kilcoyne, "Information Technology, Marketing Practice and Consumer Privacy: Ethical issues," Journal of Public Policy & Marketing. vol. 12, pp. 106-119, 1 1993.
[13] R. Clarke, "Profiling: A Hidden Challenge to the Regulation of Data Surveillance," Journal of Law and Information Science . vol. 4, pp. 2 1993.
[14] L. Barkhuus and A. Dey, Location Based Services for Mobile Telephony: a study of users' privacy concerns, Interact 2003, 2003, Zürich.
[15] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," Communications of the ACM. vol. 4, pp. 2 1981.
[16] T. Koelsch, L. Fritsch, M. Kohlweiss and D. Kesdogan, "Privacy for Profitable Location Based Services," in Proceedings of the Security in Pervasive Computing Workshop (SPC) 2005, vol. 3450, Boppard: Springer, 2005, pp. 164-179.
[17] P. Sholtz, "Transaction Costs and the Social Costs of Online Privacy," First Monday. vol. 6, pp. 5 2001.
[18] J. Backhouse, C. Hsu, J. C. Tseng and J. Baptista, "A question of trust," Communications of the ACM. vol. 48, pp. 87-91, 9 2005.

Lothar Fritsch was born in Germany in 1970. He completed his diploma degree in Computer Science with IT security and cryptography focus in 1999 at the Universität des Saarlandes in Saarbrücken, Germany. From 1995 to 1996 he studied Computer Science and Journalism at the University of Missouri in Columbia, Missouri, USA.
He worked as a Product Manager for IT security solutions at fun communications GmbH in Karlsruhe, Germany from 1999 to 2002. There, he developed and marketed certified security solutions and e-signature platforms.
Since 2002, he is a researcher at Johann Wolfgang Goethe-Universität in Frankfurt, Germany. At the chair for Mobile Commerce and Multilateral Security, he researches privacy-respecting M-Commerce for business models. He additionally is a member of Germany's Gesellschaft für Informatik (GI), and the Volcanic Hazards Docmentation and Logistics Research (VHDL).