Exploring the Relationship Between Context and Privacy

Timo Heiber and Pedro José Marrón
University of Stuttgart
Institute for Parallel and Distributed Systems (IPVS)

Abstract

Privacy is an important consideration for context-aware systems, because an individual's context contains a large amount of personal information. Moreover, the semantics of context information make it possible to infer knowledge beyond the actual information represented: For instance, even if a user is working under a pseudonym, she can be identified if a sufficient amount of location context is known about her. In this paper, we describe a generic framework to model privacy in context-aware systems that makes it possible to model inferences based on data semantics. We also present an example instance of the framework to demonstrate its practical application.

1 Introduction

Most people agree that privacy protection is an important aspect of networked and distributed applications, especially in the fields of mobile and pervasive computing. However, it is hard to agree on a common definition of privacy, for two main reasons: First, the definition depends on the highly variable preferences of individuals and socio-cultural groups. Secondly, in contrast to the related security goal of data confidentiality, privacy is not an all-or-nothing notion. It is often acceptable to divulge a limited amount of personal data, whereas it may be unacceptable if large amounts of the same type of data become known. Stajano ([15], Chapter 5.2) gives several examples for such quantitative aspects of privacy.

The problem becomes even harder when one considers the question of personal privacy with respect to context-aware applications, i.e. applications that take the context of entities into account. In pervasive computing, the most important entities are individuals. According to [4], context is information that describes the situation of an individual, which means that the question of personal privacy arises naturally: The amount of context information that is personal (such as the location of a user) or related to personal information (such as the location of a user's mobile device) could conceivably grow quite large. Additionally, someone interested in obtaining personal information (hereafter termed "adversary") would have a multitude of opportunities. Moreover, the semantics of context information can be leveraged to infer context information that is not explicitly stated in the available pieces of context information. Consider, for instance, the point that the location of a certain user's mobile device can be used to infer information about the location of that user.

Previous work on privacy in pervasive computing has studied the possibility of augmenting personal information with usage policies [9, 11]. Apart from the fact that enforcement of such policies is problematic in context-aware systems, these approaches do not explicitly consider inferences based on the semantics of context information.

Assuming a global view of the problem, there are three main questions that influence a user's degree of privacy in context-aware environments:

1. How much personal context data can be collected by an adversary?

2. What is the content of that context data?

3. How successful is the adversary in attributing that data to a particular person?

In this paper, we present a generic framework for privacy in context-aware computing systems. We focus on how to model inference-based attacks on context information within this framework. We demonstrate the feasibility of our approach by showing how to model inferences based on the context information of location, time and identifier (sometimes called primary context [4, 13]).

This paper is structured as follows: In Section 2, we present an example scenario from which we derive a generic framework for privacy in context-aware systems in Section 3. We then discuss the formalization of this generic framework in Section 4, with a concrete instance provided in Section 5. We review the related work in Section 6. In Section 7, we summarize our approach and discuss directions for future work.

2 Motivation

Consider a scenario with an abundant supply of context-based information systems: Location-based services track the locations of their customers and supply information relevant at their current location (e.g. route planning, public transport timetables etc.) while "infostations" supply information to anyone in their transmission range. User Alice uses her personal devices to communicate with such information systems. Location tracking and communication with the location-based service is done via a mobile phone network that can provide high location resolution (e.g. UMTS). Access to the infostations is gained through her WLAN-equipped PDA.

We assume that Alice needs to authenticate herself to the location-based information system (LBS) for billing purposes. A a consequence, she is known to the LBS under the identifier *Alice-LBS* (which might be a pseudonym). The PDA uses a wireless LAN adapter with the constant device ID (MAC address) *Alice_PDA*.

Now consider adversary Eve that has gained access to the information generated by the transmissions of Alice's devices (for example, a UMTS service provider that also monitors WLAN traffic at some locations). Eve could then collect two types of location traces for all users. With respect to Alice, she would obtain location traces under the ID *Alice-LBS* and also other location traces under the ID *Alice_PDA*, using the location information that comes implicitly with the WLAN transmissions.

Eve's next step would be to correlate both types of location traces in order to see whether a pair of location traces from different sources matches. In that way, two different identifiers (*Alice-LBS* and *Alice_PDA*) could be linked to the same person (Alice). Furthermore, only one success in this regard will be enough to link Alice's identifiers *Alice-LBS* and *Alice_PDA* from this point on.

The important point of this scenario is that with increasing amounts of context information, attempts to penetrate an individual's privacy will also be increasingly successful, because the adversary will be able to leverage the semantics of context data items to infer additional information that is not explicitly stated in the context information. Even if data is only stored in pseudonymous form, adversaries will often be able to link items of context data based on their content. Moreover, the problem discussed here is not restricted to a specific application scenario, but remains valid for any form of constant identifier, for instance RFID tags that can be interrogated remotely.

Note that the amount of data used by Eve in the example above is comparatively small and restricted to identifiers and spatio-temporal coordinates. This is an indication that the privacy problems will become even worse when context-aware computing is used ubiquitously and other forms of context data are taken into account.

3 A Generic Framework for Modeling Privacy in Context-Aware Systems

In this section, we describe a common framework for privacy that reflects the main factors that are relevant for any model of privacy in context-aware systems. It consists of several interrelated components that model the essential parameters for privacy in context-aware systems. These are the contents of the data, the capabilities of the adversary to obtain data, possible inferences and the actual privacy requirements, as shown in Figure 1.

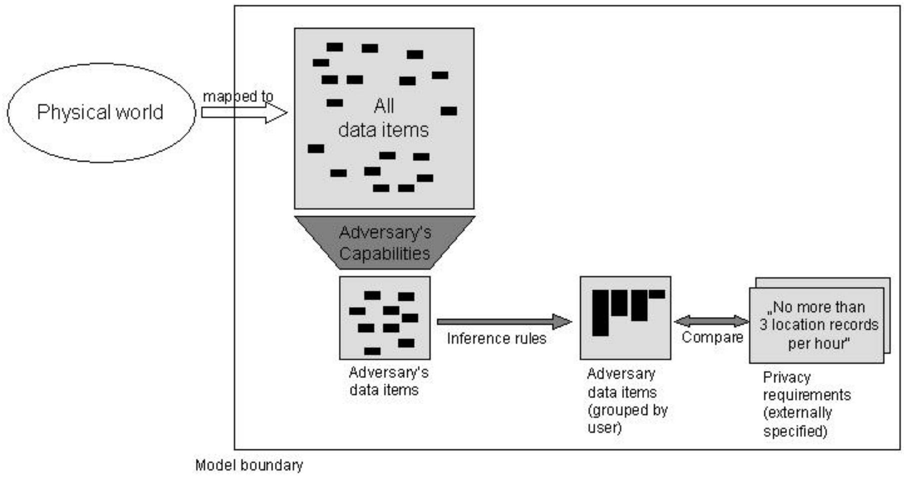

Figure 1: Generic framework for privacy

The *physical world* is external to our model boundaries and represents the events occurring in the physical world: people walking around, people using computers to access information or send messages etc.

Some of these events leave an "electronic trace", e.g. cause records of information to be stored on a computer system. We refer to each of these discrete records as a *data item*. Conceptually, we consider the *set of all data items* to be one database to which an adversary trying to violate people's privacy would like to gain access. A data item may, for instance, be created when a person sends an e-mail message, or when his or her whereabouts are recorded by a location tracking system. In the example of Section 2, data items containing Alice's location are generated at the location-based service and also due to the communication of the PDA.

A subset of all existing data items is available to an adversary (*adversary's data items*). The size and exact composition of this subset depends on the *adversary's capabilities*, which are a function of her abilities and the access control mechanisms employed to restrict access to the data items. In the "Alice" example, we assumed Eve to have access to the data stored at both the public service and the two non-public services.

The next step of the adversary is to apply *inference rules* to organize the data available to her. Organizing the data items refers to grouping the data items by user. This grouping is done by examining the contents of the data items to determine which data items have been created through the activities of the same user.

Whether the *privacy requirements* of a certain user have been violated is determined by how much and what type of information the adversary has gained access to.

Based on the description above, a model for privacy in context-aware system needs four components:

1. A *data model* that describes what kind of data items are created.
2. An *adversary model* that describes what data items the adversary can gain access to.
3. The *inference rules* that can be applied to the data by the adversary.
4. And, finally, a characterization of the *privacy requirements* for the system.

A formal characterization of these four components makes it possible to derive the knowledge that can possibly be gained by an adversary and evaluate it with respect to previously stated privacy requirements for the context-aware system.

In the following section, we provide a formalization of the first three components using structured data and predicates on this data. That way, it is possible to derive the knowledge that can possibly be gained by an adversary. This suffices to evaluate whether simple privacy requirements like "no more than n location records within a time interval of length t" hold. A system for formally stating complex privacy requirements within this model is beyond the scope of this paper and will be considered in future work.

4 Formalization of the Model

The generic formalization of the model is based on predicates that describe the capabilities and inferences of the attacker.

4.1 Generic Data Model

With respect to the data model, two questions need to be answered:

1. What information do we need to represent?

2. How can we design a flexible and extensible data model?

Let us first consider what information we need to model with respect to user privacy. Here, the most important pieces of information are identity, location and time. We refer to these as *primary context* and they suffice to model the well-known location privacy problem for context-aware environments [8].

All other context information is referred to as *secondary context*. This includes specific constant properties of an entity (such as user preferences) and its current state and activity.

Existing work [4, 13] uses the separation into primary and secondary context, but researchers disagree on what type of information belongs to which category. For example, Dey [4] considers *identity*, *location*, *time* and *activity* to be part of the primary context, whereas Rothermel et al. [13] limit it to *identity*, *location* and *time*. Since the following discussion is mostly concerned with identifying and isolating mobile entities, we use the definition in [13].

In order to achieve flexibility and extensibility, we use the following generic representation for data items: A data item is a four-tuple *(ID, Location, Time, Secondary Context)* containing the following information:

ID The identifier under which this data item was created. In the example, this field would contain Alice's customer ID or the MAC address for her PDA.

Location Spatial information about a data item. This field would for example contain Alice's location when it is stored by the location-based service.

Time Temporal information about the data item. In the example, this refers to the time of a communication or the time of an update of location data.

Secondary Context Any other information in the data item. An example for *secondary context* would be the content of Alice's communication.

We think of data items as generic records that can be further structured into subfields, depending on the actual data created in an application scenario. Using dot-notation to refer to subfields, we would, for example, model the different types of IDs by substructuring the ID field into $ID.MAC\_Address$ and $ID.Customer\_ID$. Such substructuring can also be used to introduce name spaces in order to avoid clashes between the customer IDs of several service providers. In this case, we would introduce the field $ID.Provider$ to name the service provider explicitly. A complete instance of the generic data model can be found in Section 5.1.

4.2 Generic Adversary Model

The adversary model is, in effect, a filter applied to the set of all data items. We represent the capabilities of an adversary with a generic predicate *visible_to_adversary*. This predicate is defined on data items and evaluates to true if the adversary can learn this data item. A concrete instance of this predicate can be found in the example in Section 5.2.

4.3 Generic Inference Rules

Context data items are a-priori independent of each other. However, primary context contains sufficient information relating to users that can be exploited to learn whether data items were caused by the same user. In effect, the adversary infers, based on the fields of a data item, that certain data items relate to the same person. As a result, the adversary collects of sets of data items, where all data items in the same set can be attributed to the same person. In this section, we describe the generic structure of an inference system based on primary context.

Referring back to our example again, we saw how inference allowed the adversary to link data items based on their content (in this case, user IDs). Also, correlation of spatio-temporal coordinates made it possible to link unrelated identifiers for Alice and increase the amount of knowledge about her. This means that there are two types of inference rules: Linking based on user IDs, which only requires examining the content of two single data items, and correlation of coordinates, which needs sufficient overlap in whole location traces of a user. That is, in the second case, two whole *sets* containing already linked data items must be examined in order to obtain a match.

We represent these linking strategies by two generic inference rules, one that deals with linking data items, thereby aggregating them into sets of linked data items and one that deals with linking sets of already linked data items, thereby producing even larger sets of data items. The privacy of a person degrades directly with the size of the set of data items attributable to him or her.

Formally, the generic inference model provides two inference rules, one that works on pairs of single data items and one that works on pairs of sets of data items. These rules are based on two predicates, which are instantiated according to the data model and the inference possibilities of the application scenario:

linkable The predicate *linkable* is defined on data items: Two data items are *linkable* if their respective contents warrant the conclusion that they relate to the same person. For example, two data items that contain the same unique identifier for a person could be considered to be linkable. The predicate is transitive and induces an equivalence relation on data items.

matching The predicate *matching* is defined on sets of data items: It represents those cases where correlation of two sets of linked data items leads to the conclusion that both sets relate to the same person. For example, two sets of data items are matching if they contain large numbers of matching location records from highly different locations, *and* there is no pair of data items from the different sets that record different locations for the same point in time. Note that the negation implies that this predicate is not necessarily transitive.

Using these predicates, an adversary can execute Algorithm 1 and, after that, Algorithm 2. Each of the sets obtained in this way is a representations of an actual person. This means that, for the adversary, a person is defined as a set of the data items created by that person. Section 5.3 will provide a concrete instance of these predicates.

5 Modeling the Example Scenario

In this section, we formalize the data model, adversary model and inference rules used in the example scenario of Section 2.

Algorithm 1 Generic algorithm for collecting linkable data items

function collect_items (D : **set of** data_item) : **set of** (**set of** data_item)
begin
if $D = \{\}$ **then**
  return $\{\}$
end if
{Otherwise, build a set of data items that are (transitively) linkable}
select any $d \in D$
$D := D - \{d\}$
$R := \{d\}$
for all $d' \in D$ **do**
  if $\exists d'' \in R : linkable(d', d'')$ **then**
    $D := D - \{d'\}$
    $R := R \cup \{d'\}$
  end if
end for
return $\{R\} \cup collect\_items$(D)
end

Algorithm 2 Generic algorithm for building sets attributable to the same person

function collect_sets (S : **set of** (**set of** data_item))
  : **set of** (**set of** data_item)
begin
if $\exists s_1, s_2 \in S : matching(s_1, s_2)$ **then**
  $S := S - \{s_1, s_2\}$
  $S := S \cup \{s_1 \cup s_2\}$
  return collect_sets(S)
end if
return S
end

5.1 Data Model

For the example, we need to model two types of data items, one for the WLAN communication and one for the location-based service.

5.1.1 WLAN Communication

Each transmission of a WLAN-equipped device creates a data item. In order to represent this, we use data items of the following format:

ID This field has the following subfields:

ID.MAC_Address The MAC address of the device.

ID.Technology The technology used to make transmissions. All data items caused by 802.11 wireless LAN devices will have the constant value “IEEE 802.11 MAC”.

Location The location at which the transmission occurred. This location is the area served by a certain WLAN access point.

Time The time at which the transmission occurred. If an adversary can perceive a transmission, this information will be fairly exact, since the delay between physical transmission and reception will be negligible.

5.1.2 Location-Based Service

The location records for the location-based service have the following form:

ID Again, we make use of subfields:

> **ID.Customer_ID** The customer for which this record is created.
>
> **ID.Provider** The name of the service provider.

Location The location at which the transmission occurred as determined by the location system in use.

Time The time at which this record was created. Again, it should be possible to determine this information in a fairly exact way.

For two locations l_1 and l_2, we write $l_1 \sim l_2$ if and only if l_1 and l_2 are less than 50 meters apart. We also define a comparison operator $\approx$ for times. For two times t_1 and t_2, $t_1 \approx t_2$ if and only if t_1 and t_2 are less than one minute apart. For the purpose of this discussion, we assume that location and temporal information can be determined with a good enough accuracy for theses operators.

5.2 Adversary Model

In the context of our common framework, the amount of such items the adversary can actually perceive (and the accuracy of the location information) will depend on his capabilities. A capable adversary will have good coverage of large areas.

For simplicity's sake, we assume that adversary Eve is capable of overhearing wireless LAN transmissions and has full access to location-based service X:

A data item d is *visible_to_adversary* if

$$d.ID.Technology = \text{“IEEE 802.11 MAC”}$$

or

$$d.ID.Provider = \text{“Location-based Service X”}$$

5.3 Inference Rules

The predicates *linkable* and *matching* are defined as follows:

5.3.1 Linkable

Two data items, d_1 and d_2 are *linkable* if

$$d_1.ID.Technology = d_2.ID.Technology$$

and

$$d_1.ID.MAC\_Address = d_2.ID.MAC\_Address.$$

Two data items, d_1 and d_2 are also *linkable* if

$$d_1.ID.Provider = d_2.ID.Provider.$$

and

$$d_1.ID.Customer\_ID = d_2.ID.Customer\_ID.$$

This definition makes the reasonable assumption of constant identifiers and defines linkability by the identity of the identifiers. Additionally *ID.MAC_Address* and *Provider* are used to provide name spaces for the identifiers.

5.3.2 Matching

For a parameter $k \in \mathbb{N}$, which is dependent on the accuracy with which location and time information can be captured and compared, two sets of data items D_1 and D_1 are *matching* if for some $k' \geq k$

$$\begin{aligned}\exists D_1' = \{d_{11}, \ldots, d_{1k'}\} \subseteq D_1, D_2' = \{d_{21}, \ldots, d_{2k'}\} \subseteq D_2 :\\ \forall 1 < i < k' : d_{1i}.Time \sim d_{2i}.Time \wedge\\ d_{1i}.Location \approx d_{2i}.Location\end{aligned}$$

and

$$\neg \exists d_1 \in D_1, d_2 \in D_2 : d_1.Time \sim d_2.Time \wedge d_1.Location \not\approx d_2.Location$$

where $\sim$ and $\approx$ are defined as in Section 5.1.

The predicate *matching* is defined by requiring two sets of data items to contain a sufficient ($k' \geq k$) number of items that place the user at the same location at the same time. Also, the match fails if the two sets contain data items that have the user at different locations at the same time.

Note that for this inference rule, the problem of uncertainty comes into play: For higher values of k, the confidence behind this inference becomes higher (but the adversary might then erroneously miss a correct inference). Presently, we are working on ways to incorporate uncertain inference into our model.

5.4 Remarks

A set of data items derived through application of this definition describes a person in terms of the device he or she used and the locations at which that occurred. It is noteworthy that sets derived in this way do not contain directly identifying information. However, sufficiently detailed location information would make identification of any person comparatively easy (e.g. because most people spend most of their time at home). Also, after a person has been identified once, his or her name can always be linked to the use of his or her personal device.

It is clearly possible to model a multitude of scenarios based on primary context by defining appropriate data models and inference rules. Additionally, different types of attackers can be modeled by changing the predicate *visible_to_adversary*. For instance, an adversary could be restricted to operate only at certain locations. Also, by appropriately extending the data and inference models, extensions incorporating secondary context are conceivable.

6 Related Work

Pervasive Computing scenarios [12, 18] are full of privacy issues. However, much of the current work in this field has, with some exceptions, not yet progressed much beyond the conceptual stage [16, 17].

Most of the related work discusses privacy in narrow application scenarios, instead of taking a more generic approach. For instance, location privacy in mobile environments has been extensively studied, e.g. by [5, 6, 10]. However, these approaches only consider location privacy and are mostly concerned with concrete solutions aimed at specific communication technologies. In contrast, our approach is broader in scope and more concerned with providing a privacy model from which possible solutions for general privacy problems can be derived.

The Platform for Privacy Preferences Project (P3P) [2] aims at developing machine-readable privacy policies and preferences. This approach is somewhat related to our model component for privacy conditions. An interesting issue that comes up in both P3P and our work and that is worth further investigation is how preferences can be described in an easy to understand and human-readable form and then transformed into a more formal representation. Marc Langheinrich, one of the authors of P3P has also extended the P3P concept to ubiquitous computing scenarios [11].

The Freiburg Privacy Diamond [19] is more closely related to our approach. The authors model privacy in mobile computing environments using relations between the sets of users, devices, actions and locations. The only inference rule in their model is transitive closure. As a result, the expressiveness of the model is limited. The authors also discuss the possibility of including probabilities and time in their model, although it remains unclear where the probabilities come from and the concept of time is only mentioned briefly in the paper.

The work of Snekkenes [14] discusses the question of access control policies with respect to personal information. Snekkenes presents a lattice model to describe the accuracy of information (e.g. the accuracy of location, time or identifying information). The way we represent identifying information as sets of data items relating to the same person is comparable (but not identical) to his approach. The high-level view of the privacy problem presented here does not consider the accuracy of other types of information. We plan to consider the question of accuracy of information in our further work.

Hengartner and Steenkiste [7, 8] consider access control with respect to personal information in a pervasive computing context. Their second work [7] mentions the need to model the relationships between different "pieces of information", although the paper does not yet give any details about their approach. The generic data model and inference system presented here is an attempt to provide such a model.

Inference Control [1, 3] is the common term for approaches to limit the inference capabilities of an adversary. Our framework provides a method to model the semantic inference capabilities of the adversary, in contrast to the more common syntactic approaches to inference control.

Our approach is also related to Deductive Databases and Expert Systems, however, our work is strictly focused on the field of privacy. Moreover, we are also less concerned with deducing a result from a database and more with modeling risks.

7 Conclusion and Further Work

The contributions of is paper are twofold: First, we presented a generic framework for discussing the privacy problem in context-aware systems. Secondly, we introduced the inference problem for context data by providing a generic model for the representation of context data and concepts for the modeling of inferences based on the primary context of location, time and ID.

We are presently in the process of formalizing the inference model for primary context more rigorously, using a restricted form of First-Order Logic. We are also exploring ways to represent inexact information and uncertain inference within our model.

Future work will extend our model to inferences that take secondary context into account. Note that the inclusion of the context information of *Activity* alone will open up further inference problems (not the least of them being the fact that activities will often provide implicit location information).

Additionally, we are working on methods to evaluate the accuracy of models based on our framework and to derive access control rules for personal information derived from such models.

One future problem may be the potentially high complexity of the inference predicates, especially when secondary context and uncertain inference is taken into account. Therefore, we plan on investigating the use of heuristics to limit the complexity of modeling inference capabilities.

A final line of work is the question of how to directly use the inference models (which represent the possible tools of the adversary) in order to drive access control mechanisms for context-aware systems.

References

[1] Ross Anderson. *Security Engineering*. J. Wiley and Sons, 2001.

[2] Lorrie Cranor, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, and Joseph Reagle. The Platform for Privacy Preferences 1.0 specification. W3C Recommendation, The World Wide Web Consortium, April 2002.

[3] Dorothy E. Denning. *Cryptography and Data Security*. Addison-Wesley, 1982.

[4] Anind K. Dey. Understanding and using context. *Personal Ubiquitous Comput.*, 5(1):4–7, 2001.

[5] A. Fasbender, D. Kesdogan, and O. Kubitz. Analysis of security and privacy in mobile ip. In *Proceedings of the 4th International Conference on Telecommunication Systems Modeling and Analysis*, March 1996.

[6] Christian Hauser, Alexander Leonhardi, and Paul J. Kühn. Sicherheitsaspekte in Nexus - einer Plattform für ortsbezogene Anwendungen. *Informationstechnik und Technische Informatik (it+ti)*, 44(5):268–277, oct 2002.

[7] Urs Hengartner and Peter Steenkiste. Access control to information in pervasive computing environments. In *Proceedings of the 9th Workshop on Hot Topics in Operating Systems (HotOS IX)*, Lihue, Hawaii, May 2003.

[8] Urs Hengartner and Peter Steenkiste. Protecting access to people location information. In *Proceedings of the First International Conference on Security in Pervasive Computing (SPC 2003)*, Lecture Notes in Computer Science, Boppard, Germany, March 2003. Springer-Verlag.

[9] Xiaodong Jiang and James Landay. Modeling privacy control in context-aware systems using decentralized information spaces. *IEEE Pervasive Computing*, 1(3):59–63, July-September 2002.

[10] D Kesdogan, H Federrath, A Jerichow, and A Pfitzmann. Location management strategies increasing privacy in mobile communication. In *12th International Information Security Conference*, Samos, Greece, 21–24 1996. Chapman & Hall.

[11] Marc Langheinrich. A privacy awareness system for ubiquitous computing environments. In Gaetano Borriello and Lars Erik Holmquist, editors, *UbiComp 2002: Ubiquitous Computing, 4th International Conference*, volume 2498 of *Lecture Notes in Computer Science*, page 237ff, Göteborg, Sweden, September 29 - October 1 2002. Springer-Verlag.

[12] Friedemann Mattern. The vision and technical foundations of Ubiquitous Computing. *Upgrade*, 2(5):75–84, October 2001.

[13] Kurt Rothermel, Martin Bauer, and Christian Becker. Digitale Weltmodelle – Grundlage kontextbezogener Systeme. In Friedemann Mattern, editor, *Total Vernetzt?!* Springer-Verlag, 2003.

[14] Einar Snekkenes. Concepts for personal location privacy policies. In *Proceedings of the 3rd ACM conference on Electronic Commerce*. ACM Press, 2001.

[15] Frank Stajano. *Security for Ubiquitous Computing*. J. Wiley and Sons, 2002.

[16] Ubicomp'2002. Socially-informed design of privacy-enhancing solutions in ubiquitous computing: Workshop at Ubicomp'2002, sep 2002.

[17] Ubicomp'2003. Ubicomp communities: Privacy as boundary negotiation: Workshop at Ubicomp'2003, oct 2003.

[18] Mark Weiser. Some computer science issues in ubiquitous computing. *Communications of the ACM*, 36(7):75–84, July 1993.

[19] Alf Zugenmaier, Michael Kreutzer, and Günter Müller. The Freiburg Privacy Diamond: An attacker model for a mobile computing environment. In K. Irmscher and K.-P. Fährich, editors, *Proceedings KIVS 2003*. VDE-Verlag, February 2003.