Composition and Generalization of Context Data for Privacy Preservation*

Linda Pareschi[1] Daniele Riboni[1] Alessandra Agostini[2] Claudio Bettini[1]

[1]*DICo, University of Milano*
{*pareschi,riboni,bettini*}*@dico.unimi.it*

[2]*DISCo, University of Milano-Bicocca*
alessandra.agostini@disco.unimib.it

Abstract

This paper presents preliminary results on anonymization and obfuscation techniques to preserve users' privacy in context-aware service provisioning. The techniques are based on generalizing request parameters as well as the context data provided to the application. Local context semantic aggregation is used to improve the quality of service that can be achieved while preserving privacy. The paper also shows how the software architecture of the CARE middleware can be extended to implement the proposed techniques.

1 Introduction

Privacy has been recognized as a major issue for the provision of context-aware services. Indeed, server-side adaptation implies the communication of private information – such as user's location, personal interests, and current activity – to possibly untrusted service providers. This issue is further complicated in pervasive environments, which are characterized by the presence of ubiquitous sensing systems like positioning infrastructures, cameras, microphones, environmental and body-worn sensors.

Research in the field of privacy preservation in pervasive computing has mainly concentrated on techniques for anonymous communication [2], access control and obfuscation [8, 14], dummy requests [6], or on a combination of such techniques. Each of the proposed techniques provides an effective privacy solution for a specific scenario; however, based on the experience we have acquired while working on a framework for context-awareness [4] and on privacy in location-based services (LBS) [3], we argue that a satisfactory comprehensive solution for privacy in pervasive computing is still missing.

Obviously, the definition of a comprehensive framework for privacy in pervasive environments is a long-term goal. As a first step in this direction, in this paper we investigate the use of *anonymity* [13] in combination with other techniques based on obfuscation. While anonymity has been proposed for privacy protection in LBS, its use in generic frameworks for context-awareness is complicated by the fact that the set of context data (that we call *context dimensions*) to be considered is wide, and not restricted to solely location. As a matter of fact, the more context dimensions we consider in anonymization, the more is the risk that context data become too general to provide the service at an acceptable quality level. In this paper we concentrate on this issue, and propose mechanisms for reducing the number of context data dimensions involved in the anonymization process by composing raw context data (e.g., data directly acquired from sensors) into complex context data (e.g., activities). These mechanisms are based on user-side context reasoning and context data aggregation and reasoning on a trusted privacy module. Our proposed solutions – illustrated by means of a motivating pervasive computing scenario – are integrated into a framework for context-awareness.

The paper is structured as follows: In Section 2 we provide preliminary information about privacy protection; In Section 3 we present our motivating scenario; In Section 4 we illustrate our proposed architecture; Section 5 concludes the paper.

2 Preliminaries on privacy

A privacy threat is generally intended as the possibility that an adversary reconstructs a *sensitive association (SA)*, i.e. an association between the user's identity and some of her *private information (PI)*. In order to prevent the release of *SA* it is possible to modify

*This work was partially supported by Italian MUR grant (InterLink project N.II04C0EC1D).

the released data in order to increase the uncertainty about the user's identity or about the private information. The uncertainty on the user identity is called anonymity: it has been introduced for data base systems [13] and then adapted to LBS services (like, e.g., in [3]). Roughly speaking, the rationale of anonymity is to make the actual issuer of a request indistinguishable in a set, called *anonymity set*, of potential issuers. The cardinality of the anonymity set determines the degree k of anonymity achieved for a given request.

Some data in a request for context-aware services can increase the ability of the adversary of inferring *SA* when joined with *external knowledge* he can access (e.g., positioning systems, telephone books). These categories of data are called *quasi-identifier (QI)* [13]. Clearly, which elements of a request act as *QI* strongly depends on the external knowledge available to the adversary. We point out that *PI* are not always included in a request; however, they could be inferred through data issued with the request. Even if the private information are not contained in the request, we will consider *PI* even those data useful for inferring them.

Most k-anonymity techniques are based on the generalization/suppression of *QI* data (e.g., the exact user's coordinates are generalized to an area), and on the replacement of the user's unique identifier with a *null* value or with a *pseudoID*. Hence, before arriving to the service provider, each request is transformed into a generalized request in which identity information and *QI* components have been appropriately transformed for guaranteeing a certain degree of anonymity.

3 A pervasive computing scenario

As a motivating scenario, we consider the pervasive system of a gym (called *PerGym*) providing personalized services on the basis of sensitive context data. We describe two services provided by the system, and highlight the resulting privacy threats. In particular, we show that simply anonymizing requests by substituting the user's unique ID with a *pseudoID* may not be sufficient for protecting the user's privacy.

In the *PerGym* scenario, users of the gym carry a mobile device (e.g., their smartphone, PDA, or a smart watch provided by the gym) that collects context data from environmental and body-worn sensors to continuously monitor data such as user's position (acquired through a user-side indoor positioning system), used equipments (through RFID), emotional status, physical activity, and physiological parameters. Part of these data are communicated by users to the gym service provider included in requests to obtain personalized services.

Suppose that the *PerGym* is considered untrusted by its users (hence, from a privacy perspective, the gym system is considered a possible adversary). The *PerGym* can continuously monitor users' positions through a server-side positioning system. Since it knows users' identity and position, and the gym map, it is anytime aware of who is using a given equipment.

Example 1 *The personalized services provided by the* PerGym *include a* Virtual-trainer *service suggesting the next exercise based on personal (gender, age) and physiological data, and on the position of users in the gym. Since physiological data are particularly sensitive (as they can reveal important details about a person's health status), they are considered* PI *in this scenario. Since we assume that the gym infrastructure is aware of users' identities and position,* QI *data for this service are personal data and users' location. In this case, the use of pseudoID in service requests is not sufficient to guarantee privacy. Indeed, the* PerGym *can easily reconstruct the* SA *by matching the location and personal data included in the request with its knowledge of users' identity and position in the gym.*

Example 2 *A further service – called* Virtual-DJ *– provided by the gym system suggests music that users can listen on their portable player on the basis of activity (e.g., exercising, resting) and music preferences. Because music preferences are dynamically set on the client device considering the user's mood, they can be used by an adversary for deriving the emotional status of the user herself, which is considered* PI*. Since the* PerGym *is continuously aware of the activities of users in the gym, user activity is* QI *for this service. Even if in service requests the user's identity is replaced by a pseudoID, the* PerGym *can try to reconstruct the* SA *by matching the activity specified in the request with its knowledge of users' activities in the gym. With this attack, the* PerGym *can reduce the set of the potential issuers to the one of users currently performing the activity specified in the request.*

4 Proposed architecture

In this section we illustrate the overall architecture we propose for protecting users' privacy in pervasive environments.

Data flow The proposed architecture is an extension of the CARE middleware [4] for context-awareness. CARE supports the acquisition of context data from different sources, the reasoning with this data based on distributed policies, and the reconciliation of possibly conflicting information. In order to protect users'

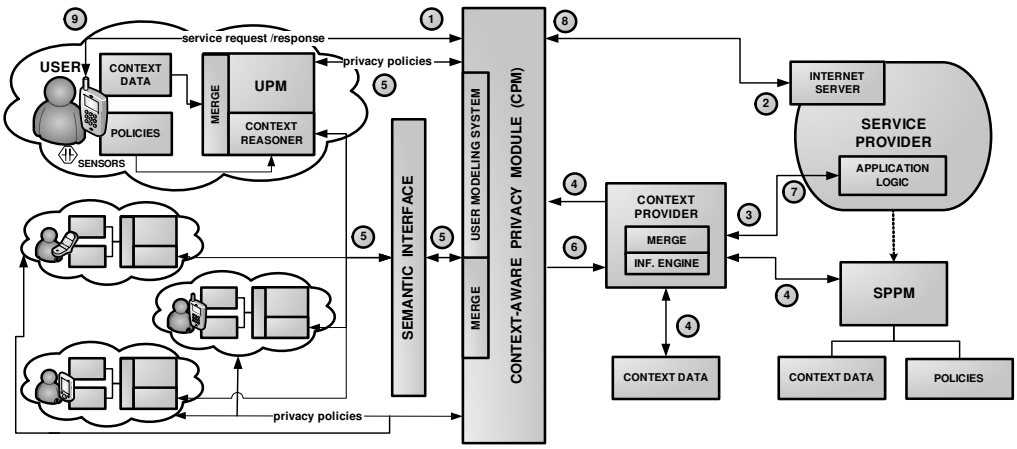

Figure 1. Overall architecture

privacy, requests are sent through an encrypted channel to a *context-aware privacy module* (CPM) in charge of enforcing users' privacy policies (see Figure 1). Context data are kept up-to-date on the CPM by periodical updates through an encrypted channel.

The CPM acts as an intermediary between the user trusted domain (left-hand side) and the rest of the world (right-hand side). The user trusted domain includes her devices and her USER PROFILE MANAGER (UPM). The UPM manages user policies and context data explicitly provided by the user or acquired from sensors in the proximity of the user. Moreover, it can perform reasoning for deriving complex context data (e.g., current activity) on the basis of raw ones (e.g., data directly acquired from sensors). Depending on the capabilities of the user's device, part of the UPM functions can be executed on the device itself (e.g., derivation of the physical activity on the basis of body-worn sensors, like proposed in [10]).

Since context data can be provided by heterogeneous entities, our architecture includes a semantic framework – called SEMANTIC INTERFACE – for integrating (either raw or complex) context data provided by different sources. The definition of such a framework is a particularly challenging research issue of semantic data integration [11]. Currently, the SEMANTIC INTERFACE module simply maps context data and associated values provided by user profile managers into profiles represented through the extended CC/PP language proposed in [1]; mapping is based on instructions manually defined by domain experts. As a future work we plan to investigate approaches based on more expressive representation languages, like the ones proposed in [5].

Numbers in Figure 1 represent the data flow upon a user request. Each user request is filtered by the CPM (1), which transforms the user ID into a *pseudoID* – used to identify the user request and to perform authentication – and removes any *QI* before forwarding the request (2) to the service provider. Then (3), the service provider asks for the context data it needs for adapting the service to a central module called CONTEXT PROVIDER, which forwards the context data request to the CPM (4), to its local profile manager (SPPM) and to its local context data sources. The CPM retrieves user's privacy policies and distributed context data from the user trusted domain through the SEMANTIC INTERFACE. Then, it merges the received context data solving possible conflicts, possibly updates the user's stereotype, and finally generalizes and obfuscates context data included in the request according to user's privacy policies (5). Then (6), it sends the resulting request to the CONTEXT PROVIDER, which merges context data in the request with data retrieved from the SPPM and external context sources, and evaluates service provider policies, thus obtaining the aggregated context data that are communicated to the application logic (7). Finally (8), the application logic adapts the service and communicates the service response to the CPM, which forwards it to the user (9).

User-side context reasoning In a pervasive computing environment the execution of context reasoning at the user side is useful for both performance and privacy reasons. For instance, consider the *PerGym* scenario (Section 3). The derivation of the data describing the physical activity of the user involves the statistical analysis of data acquired from body-worn sensors like accelerometers, microphones, and other sensors [10]. Executing for each user the reasoning tasks on a central server would be highly inefficient, both for network and for computational consumption. Moreover, in several cases the data used in the reasoning task may involve

the user's privacy (e.g., the user's calendar can be used for deriving her current activity). In these cases, the user-side execution of reasoning tasks would avoid the disclosure of sensitive information.

In our proposed architecture, the UPM associated to each user, includes a CONTEXT REASONER module that is in charge of deriving complex context data on the basis of raw ones. In our current implementation, the UPM can perform ontological reasoning (e.g., for deriving the current activity on the basis of user's calendar and location [1]). However, the UPM can be easily extended to perform statistical reasoning for deriving the user's emotional status (like, e.g., in [9]) and physical activity (like, e.g., in [7]). Moreover, since some context readings can be redundant or contradictory (e.g., location data provided by different positioning systems), the UPM includes a MERGE module devoted to solve conflicts between context data, as described in [4].

Reduction of context data dimensions with *stereotypes* As outlined before, any reference to the user's personal data are removed from the request by the CPM; hence, the service provider is no longer able to customize the service according to data such as the user age, gender, and formal education.

In order to address this issue, we propose to extend the pseudoID approach by the use of *stereotype* [12] hierarchies for composing context data such as demographic data and personalities into synthetic abstractions. Practically, stereotypes group similar users in a singleton on the basis of certain commonalities (e.g., gender, job). The kind and number of personal data needed to build stereotypes strongly depends on the service provided to the user. For instance, the Virtual-trainer service can usefully exploit stereotypes based on age, fitness, and level of sedentariness (e.g., "Young-InGoodForm-Active"). On the other hand, such kind of stereotypes will be pointless for the Virtual-DJ.

In our solution (the semantic of) stereotypes should be shared between the (untrusted) service provider and the (trusted) USER MODELING SYSTEM (UMS) of the CPM. In particular, the service provider is in charge of defining stereotypes on the basis of its application domain and its services. On the other side, the UMS – adopting these stereotypes – will be able to hide/generalize some user's sensitive data. In order to contribute to our proposed defense techniques, the stereotype derivation by the UMS should consider which data in the request act as *QI*. For instance, since in our scenario we assume that the external knowledge of the *PerGym* includes users' gender and age (provided at the registration time), gender and age are *QI*. Hence, the UMS should compose this data into a single stereotype (e.g., "Young-Lady"); other data in the request that are not *QI* do not need to be composed into a specific stereotype (i.e., they can be safely communicated by service requests). Moreover, a stereotype can be generalized to its ancestors in the hierarchy when the achieved degree of anonymity is insufficient.

Context-aware privacy module The main task of the CPM consists in transforming requests in order to make the actual issuer indistinguishable in a set of other potential issuers. A first step for anonymizing a request consists in substituting the identity information with a *null* value or with a pseudoID (when necessary for session management). As shown in Examples 1 and 2, this is not sufficient for preserving privacy, and more sophisticated techniques based on data generalization are required. The generalization of *QI* must be performed according to the semantics and representation of context data: if the data is represented by numerical values (e.g. location coordinates, age) they are generalized to an interval including the original value. Differently, if the data is represented as the element of a taxonomy T, its generalized value corresponds to one of its ancestors on the same hierarchy structure T. Moreover, when many elements of a request act as *QI* the generalization process must consider all the dimensions corresponding to each single data.

Obviously, the generalization of some context data in requests could be counterproductive for service adaptation; hence, in order to provide a good trade-off between quality of service and degree of privacy we apply, if necessary, a lower degree of *QI* generalization combined with techniques for obfuscating the *PI*.

With respect to our scenario, we assume that the external knowledge available to the *PerGym* is: *a)* users' location, activity, and identities (including gender and age); *b)* the gym map; *c)* the stereotypes hierarchy and semantics. According to this assumption, the CPM performs a multi-dimensional generalization of location and personal data (aggregated into a stereotype) when the user asks for the Virtual-trainer service; differently, it generalizes the user's activity and obfuscates the *PI* when the issuer asks for the Virtual-DJ service, as shown in the following examples.

Example 3 *Suppose that Alice asks to the Virtual-trainer service for the next exercise when in location l_1 situated in room R_3. Her request r contains her identity, physiological data Phy-D, position, gender and age: $r = \langle Alice, Phy\text{-}D, l_1, female, 27\rangle$. Suppose that in the same room of Alice there are three other persons; the following table represents the adversary knowledge acquired by matching data retrieved from the position-*

ing system with users' demographic data obtained at registration time:

ID	*location*	*gender*	*age*
Alice	l_1	*female*	*27*
Bob	l_2	*male*	*26*
Jane	l_3	*female*	*30*
Lucy	l_4	*female*	*22*

The CPM *transforms request r into r' by:* a) *substituting her identity with a pseudo-id u_1;* b) *composing her personal data into the stereotype $Young\text{-}Lady$ (i.e., female having age between* 21 *and* 30*);* c) *generalizing her exact location l_1 with room R_3: $r' = \langle u_1, Phy\text{-}D, R_3, Young\text{-}Lady \rangle$. Hence, the resulting degree of anonymity is 3 since also Jane and Lucy match the information issued with the request.*

Example 4 *Alice prefers to listen "Hip-Hop" when she is "happy", "House" music when she is angry, and "Blues" when she is tired; her preferences are stored and evaluated by the* UPM*. Alice requires the Virtual-DJ service while coffee-breaking; her* UPM *infers, reasoning with data acquired from body-worn sensors, that at the request time she is "happy". Hence, the request contains "Hip-Hop" as music preference and "coffee-breaking" as current activity. Suppose that the adversary knows that Alice is the only person in the gym having coffee-break at that time of the day. Hence, in addition to substituting her identity with a pseudoID, the* CPM *generalizes her precise activity into a more general value "resting", which is an ancestor of "coffee-breaking" in the activity hierarchy. Also suppose that the degree of anonymity achieved is not yet sufficient for guaranteeing the desired level of privacy because the adversary knows that there are only* 2 *persons in the gym having a rest. Therefore, the* CPM *obfuscates the* PI *in the request, generalizing the value of the music preference to "Soul" (that is ancestor of "Hip-Hop" in the music categories hierarchy).*

5 Conclusion and future work

In this paper we investigated the use of anonymization and obfuscation techniques for privacy preservation in context-aware pervasive environments. Some of the most challenging research issues we plan to investigate include *a)* an extension of our privacy protection techniques to support multidimensional k-anonymity for context-aware services; *b)* the definition of a comprehensive measure of the trade-off between privacy and quality of service; *c)* the representation of privacy policies, taxonomies and stereotypes. Moreover, we plan to study an extension to support the *dynamic case*, i.e., when an adversary is able to reconstruct the sensitive association by means of requests issued by the same user in different time intervals.

References

[1] A. Agostini, C. Bettini, and D. Riboni. Loosely Coupling Ontological Reasoning with an Efficient Middleware for Context-awareness. In *Proc. of MobiQuitous05*, pages 175–182. IEEE Computer Society, 2005.

[2] J. Al-Muhtadi, R. H. Campbell, A. Kapadia, M. D. Mickunas, and S. Yi. Routing Through the Mist: Privacy Preserving Communication in Ubiquitous Computing Environments. In *Proc. of ICDCS02*, pages 74–83. IEEE Computer Society, 2002.

[3] C. Bettini, S. Mascetti, and X. S. Wang. Privacy protection through anonymity in location-based services. To appear in *Handbook of Database Security: Applications and Trends*, Springer, 2007.

[4] C. Bettini and D. Riboni. Profile Aggregation and Policy Evaluation for Adaptive Internet Services. In *Proc. of MobiQuitous04*, pages 290–298. IEEE Computer Society, 2004.

[5] E. Bouillet, M. Feblowitz, Z. Liu, A. Ranganathan, A. Riabov, and F. Ye. A Semantics-Based Middleware for Utilizing Heterogeneous Sensor Networks. In *Proc. of DCOSS07*, volume 4549 of *LNCS*, pages 174–188. Springer, 2007.

[6] H. S. Cheng, D. Zhang, and J. G. Tan. Protection of Privacy in Pervasive Computing Environments. In *Proc. of ITCC05*, pages 242–247. IEEE Computer Society, 2005.

[7] B. Clarkson, A. Pentland, and K. Mase. Recognizing User Context via Wearable Sensors. In *Proc. of ISWC00*, pages 69–75. IEEE Computer Society, 2000.

[8] F. Gandon and N. M. Sadeh. A Semantic E-Wallet to Reconcile Privacy and Context Awareness. In *Proc. of ISWC03*, volume 2870 of *LNCS*, pages 385–401. Springer, 2003.

[9] K. H. Kim, S. W. Bang, and S. R. Kim. Emotion Recognition System Using Short-term Monitoring of Physiological Signals. *Medical and Biological Engineering and Computing*, 42(3):419–427, 2004.

[10] J. Lester, T. Choudhury, and G. Borriello. A Practical Approach to Recognizing Physical Activities. In *Proc. of PERVASIVE06*, volume 3968 of *LNCS*, pages 1–16. Springer, 2006.

[11] E. Rahm and P. A. Bernstein. A Survey of Approaches to Automatic Schema Matching. *VLDB Journal*, 10(4):334–350, 2001.

[12] E. Rich. User Modeling via Stereotypes. *Cognitive Science*, 3(4):329–354, 1979.

[13] P. Samarati. Protecting Respondents' Identities in Microdata Release. *IEEE Trans. on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[14] R. Wishart, K. Henricksen, and J. Indulska. Context obfuscation for privacy via ontological descriptions. In *Proc. of First Int. Workshop on Location- and Context-Awareness (LoCA)*, volume 3479 of *LNCS*, pages 276–288. Springer, 2005.