Privacy-aware Access Control with Generalization Boundaries

Min Li [1] **Hua Wang** [1] **Ashley Plank** [1]

[1] Department of Mathematics and Computing
University of Southern Queensland, Australia
Email: {limin, wang, plank}@usq.edu.au

Abstract

Privacy is today an important concern for both data providers and data users. Data generalization can provide significant protection of an individual's privacy, which means the data value can be replaced by a less specific but semantically consistent value and the personal information can be collected in a generalized form. However, over-generalized data may render data of little value. A key question is whether or not a certain generalization strategy provides a sufficient level of privacy and usability?

In this paper, we introduce a new approach, called privacy-aware generalization boundaries, which can satisfy the requirements of both data providers and data users. We propose a privacy-aware access control model related to a retention period. Formal definitions of authorization actions and rules are presented. Further, we discuss how to manage a valid access process and analysis the access control policy. Finally, we extend our model to support highly complex privacy-related policies by taking into account features of obligations and conditions.

1 Introduction

Privacy is the right of individuals to determine for themselves when, how, and to what extent private information is communicated to others. Privacy concerns are fueled by an ever increasing list of privacy violations, ranging from privacy accidents to illegal actions. Many people are aware that giving personally identifiable information (PII) to organizations may result in the data being used in ways the person never intended.

While current information technology enables people to carry out their business virtually at any time in any place, it also provides the capability to store various types of information the users reveal during their activities. The use of innovative knowledge extraction techniques combined with advanced data integration and correlation techniques makes it possible to automatically extract a large body of information from available databases and from a large variety of information repositories available on the web (Dong et al. 2005, Sarawagi & Bhamidipaty 2002). Privacy issues are further exacerbated by the Internet which makes it easy for new data to be automatically collected and added to databases (Sandhu et al. 1996, Westin 1998, 1999).

As privacy awareness increases, individuals are becoming more reluctant to carry out business and transactions online, and many enterprises are losing a considerable amount of potential profits. Also, enterprises that collect information about individuals are in effect obligated to keep the collected information private and must strictly control the use of such information. Thus, information stored in the databases of an enterprise is not only a valuable property of the enterprise, but also a costly responsibility. By demonstrating good privacy practices, many enterprises try to utilize information analysis and knowledge extraction to provide better services to individuals without violating individual privacy. As privacy becomes a major concern for both customers and enterprises, many privacy protecting access control models have been proposed (Adam & Worthmann 1989, Agrawal et al. 2002, Ashley et al. 2002, LeFevre et al. 2004). Changes in the landscape of legislation around the world, and growing consumer attention to the issue have changed attitudes towards security and privacy concerns for database systems. This matches with a substantial body of research on approaches for managing the negotiation of personal information among customers and enterprises (Tumer et al. 2003, Agrawal et al. 2003, Seamons et al. 2001).

At the basis of every solution for the exchange between enterprises and customers, there is the principle of transparency. Transparency means that when enterprises store data about customers they should disclose to customers which data is collected and how it is used; i.e., for what purpose data is maintained. Starting from the landmark proposals for Hippocratic databases (Agrawal et al. 2002), most privacy-aware technologies use purpose as a central concept around which privacy protection is built. Byun and Bertino (Byun & Bertino 2006) proposed a model based on a typical life-cycle of data concerning individuals. Each data item is generalized and stored according to a multilevel organization, where each level corresponds to a specific privacy level. When individuals release their personal information, they specify permissible usages of each of their data items and a level of privacy for each usage.

The use of data generalization[1] can significantly increase the comfort level of data providers; i.e., the personal information can be collected in a generalized form. For example, many individuals may not be comfortable with their birthdays being used. Suppose now that the enterprise promises its customers that this information will be used only in a generalized form. This assurance will surely comfort many customers. Although more information can be utilized by employing data generalization techniques, the ability to limit the level of allowed generalization could be valuable. For example, when the address information

[1] Data generalization refers to techniques that "replace a value with a less specific but semantically consistent value."

related to Australia states is used for some specific data analysis tasks, the states should be the maximal allowed generalization values. Therefore, the address information generalized beyond the state could be useless.

A key question is: how can we determine whether or not a certain generalization strategy provides a sufficient level of privacy and usability?

To answer this question, we need metrics that methodologically measure the privacy and usability of generalized data. Such metrics are necessary to devise generalization techniques that satisfy the requirements of both data providers and data users. Further, privacy enhancing access control models should be able to utilize more information by employing data generalization techniques.

In this paper, we devise the generalization boundary technique to maximize the privacy and information utilization, which satisfies the requirements of both data providers and data users. Moreover, we propose a privacy-aware access control model related to a retention period. Compared with traditional access model, we focus on retention period and generalization level to provide a much finer level of control since the access control decision is based on the question of "how much information can be allowed for a certain user", rather than "is information allowed for a certain user or not". Further, we present efficient authorization and access functions to manage the process of accessing. Finally, we extend our model by taking obligations and conditions into account to provide full support for expressing highly complex privacy-related policies.

The remainder of the paper is structured as follows. In Section 2, we describe the motivation of the paper. We present a privacy-aware generalization strategy while specifying the generalization boundaries in Section 3 and illustrate the privacy-aware access control model in Section 4. In Section 5, we propose efficient authorization and access functions to manage the process of accessing. We extend our model to provide full support for expressing highly complex privacy-related policies, taking into account features like obligations and conditions in Section 6. We provide a brief survey of related work in Section 7 and conclude the paper in Section 8.

2 Motivation

Following (Byun & Bertino 2006), privacy level, types of data and possible data usages (i.e., purposes) are defined in Table 1. During the data collection phase, a data provider submits his/her privacy requirements, which specify permissible usages of each data item and a level of privacy for each usage. For instance, a data provider[2] may select *Low* on *Address* for *Admin*; that is, he/she does not have any privacy concern over the address information when it is used for the purpose of administration. Thus, the address information can be used for the administrative purpose without any modification. However, the data provider may select $High$ on *Address* for $Marketing$. This indicates that he/she has great concerns about privacy of the address information when it is used for the purpose of marketing; thus, the address information should be used only in a sufficiently generalized form for the marketing purpose.

In addition to storing the specified privacy requirements, the actual data items are preprocessed in the following way before being stored. Each data item is generalized and stored according to a multilevel organization, where each level corresponds to a specific privacy level. Intuitively, data for a higher privacy level requires a higher degree of generalization. For instance, the address data is stored in three levels: entire address for *Low*, city and state for $Medium$ and state for $High$.

Table 2 illustrates some fractional records and privacy requirements stored in a conceptual database relation. Note that each data item is stored in three different generalization levels, $Low, Medium, High$, each of which corresponds to a particular privacy level. Intuitively, data for a higher privacy level requires a higher degree of generalization. Admin and Marketing are metadata columns storing the set of privacy levels of data for *Admin* and $Marketing$ purposes, respectively. For instance, {M, H, H} in Marketing indicates that for the $Marketing$ purpose the privacy level of $Name$ is $Medium$ while the privacy levels of *Address* and $Income$ are both $High$.

Along with the data collection, access to the data is strictly governed by the data provider's requirements. However, different people may have different feelings about their information being used for some purposes. For instance, some consumers may feel that it is acceptable to disclose their purchase history or browsing habits in return for better services; others may feel that revealing such information violates their privacy. These differences in individuals suggest that access control models should be able to maximize information utility, which may be neglected by data providers although wanted by data users. For example, if a data provider selects {M, M, M} on $Name$, *Address*, $Income$ for $Delivery$ purpose, the information obtained by the data user is shown in Table 3. However, the information will be useless for the data user who wants to fulfill the delivery purpose because full name and address are necessary information for delivery purpose. Further, this selection may increase the chance of disclosure of the unnecessary information $Income$ since the more people who know, the more likely it would be disclosed.

We believe that a new generation of privacy-aware access control models should maximize information usability by exploiting the nature of information privacy. In order to balance privacy and utility, it is necessary to devise generalization strategies that satisfy the requirements of both data providers and data users. In this paper, we propose the privacy-aware access control model with generalization boundaries, which maximizes the individual privacy and private information utility. In particular, we

- Develop the privacy-aware access control model by taking retention period into account, since personal information shall be retained only as long as necessary for fulllment of the purpose.

- Discuss how to management a valid access and we propose an efficient access control policy to identify relevant issues based on our proposed privacy-aware access control model

- Extend the privacy-aware access control model to provide full support for expressing highly complex privacy-related policies, taking into account features like obligations and conditions.

3 Privacy with generalization boundaries

Generalization that consists in replacing the actual value of the attribute with a less specific, more general value is faithful to the original (Sweeney 2002). For example, the name 'Carol Jones' can be generalized to a less specific value 'C. Jones' or further generalized

[2]Data provider refers to the subject to whom the stored data is related.

Term	Description	Example
Privacy level	Level of privacy required by data provider	Low, Medium, High
Data item	Types of data being collected (i.e. attributes)	Name, Address, Income
Data usage type	Types of potential data usage (i.e. purpose)	Marketing, Admin, Delivery

Table 1: Privacy level, data type and data usage type

Name		Address		Income		Admin	Marketing	Delivery
L	Alice Park	L	123 First St.,Seattle,WA	L	45,000			
M	A. Park	M	Seattle,WA	M	40K-60K	{L,M,H}	{M,H,H}	{M,M,M}
H	A.P.	H	WA	H	Under 100K			

Table 2: Privacy information and Metadata

Name	Address	Income	Delivery
A. Park	Seattle,WA	40K-60K	{M,M,M}

Table 3: Private information for *Delivery* purpose

to 'C.J.'. Initially, this technique was used for categorical attributes and employed predefined domain and value generalization hierarchies. Generalization was extended to numerical attributes either by using predefined hierarchies (Iyengar 2002) or a hierarchy-free model (LeFevre et al. 2006).

For each categorical attribute, a *domain generalization hierarchy* is associated. The values from different domains of this hierarchy are represented in a tree-like structure, called*a value generalization hierarchy*.

There are several ways to perform generalization. Generalization that maps all values of initial data to a more general domain in its domain generalization hierarchy is called *full-domain generalization* (LeFevre et al. 2006, Samarati 2001). Generalization can also map an attribute's values to different domains in its domain generalization hierarchy, each value being replaced by the same generalized value in the entire dataset. The least restrictive generalization, called *cell level generalization* (Lunacek et al. 2006), extends Iyengar model (Iyengar 2002) by allowing the same value to be mapped to different generalized values in distinct tuples. In this paper, we adopt cell level generalization.

By using data generalization, data providers can specify their privacy requirements using a privacy level for each data item. Data for a higher privacy level requires a higher degree of generalization, i.e., each privacy level is accompanied with a generalization level. Therefore, we assume that the generalization level is equal to the privacy level in this paper. The maximal generalization level is to generalize a data value to $*$, denoted as ML. For simplicity of discussion, we only consider the generalization levels: low L, medium M, high H and ML. Figure 1 illustrates the generalization level and value generalization hierarchy for the attribute *Address*.

In order to specify a generalization boundary, we introduce the concept of a maximum allowed generalization level that is associated with each data item. This concept is used to express to what extent the data user thinks the data item could be generalized, such that the resulted generalized data item would still be useful. Limiting the level of generalization for the data item is necessary for various usage of the data. For instance, when data related to Australian states is used for some specific analysis tasks, the data user will select the level corresponding to the states as the maximal allowed generalization level. Address information generalized beyond the Australia state level could be useless. In this case, the only solution would be to ask the data provider to make a decreased level of generalization until the generalized data satisfies the maximum allowed generalization level requirement (i.e., no address is generalized further than the Australian state).

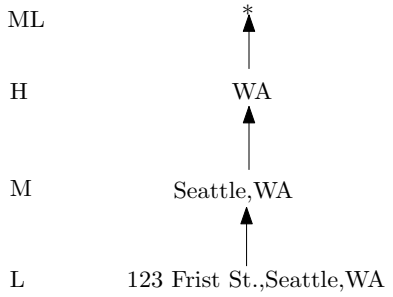

Figure 1: Example of *Address* generalization

Definition 1 ***(Maximum allowed generalization level)*** *Let D be the set of data items and P be the set of purposes, then for each data item $d \in D$ and purpose $p \in P$, the* ***maximum allowed generalization level*** *of d under purpose p, denoted by $MAGLel(d, p)$, satisfies:*

- *the data item d is permitted to be generalized only up to $MAGLel(d, p)$.*

For example, if D = {Name, Address, Income}, P = {Admin, Marking, Delivery}, then we can define the maximum allowed generalization level of $Name$ under purpose $Delivery$, $MAGLel(Name, Delivery)$ = L.

Note that the maximum allowed generalization level of the data could be different to different purposes. For example, the maximum allowed generalization level of *address* could be *Low* for *Delivery* purpose; whereas, it may be *high* for $Marketing$ purpose. Usually, for a certain purpose, the data user only has generalization restrictions for some necessary data items, e.g., there should be restrictions on $Name$ and *Address* for *Delivery* purpose but no restrictions on $Income$. If for a particular data item there are no any restrictions in respect to its generalization, then the maximal generalization level ML is specified for the usage of this data. In this case, the requirement of providing sufficient privacy and usability is satisfied by the following description.

Definition 2 ***(Privacy-aware generalization boundaries)*** *Let P be the set of access purposes and D be the set of data items, then for each purpose $p \in P$, the set $N_p \subseteq D$ denotes all necessary data to fulfill the purpose p. The* ***privacy-aware generalization boundaries*** *for p satisfies:*

- *for $\forall d \in N_p$, the data d is permitted to be generalized only up to $MAGLel(d, p)$;*

Name		Address		Income		Delivery
L	Alice Park	L	123 First St.,Seattle,WA	L	45,000	{MAGLel(Name, Delivery),
M	A. Park	M	Seattle,WA	M	40K-60K	MAGLel(Address, Delivery),
H	A.P.	H	WA	H	Under 100K	ML}
ML	*	ML	*	ML	*	

Table 4: Generalization boundaries for *Delivery* purpose

Name	Address	Income	Delivery
Alice Park	123 First St., Seattle,WA	*	{L,L,ML}

Table 5: Ideal information for *Delivery* purpose

- *for* $\forall d \notin N_p$ *and* $d \in D$*, the data* d *is permitted to be generalized up to* ML*.*

For instance, if D = {Name, Address, Income}, P = {Admin, Marking, Delivery}, and since the full name and address are necessary to fulfill the *Delivery* purpose, so $N_{Delivery}$ = {Name, Address}. Table 4 shows the example of privacy-aware generalization boundaries for the *Delivery* purpose. Because of $Name, Address \in N_{Delivery}$, the generalizations on $Name$ and $Address$ are only permitted up to $MAGLel(Name, Delivery)$ and $MAGLel(Address, Delivery)$ (i.e., *Low* and *Low*), respectively. On the other hand, for the *Income*, there is no any requirements with respect to its generalization, since $Income \notin N_{Delivery}$, so the maximal generalization level ML is specified for the usage of *Income*. According to the permissible usage of each data item, a data user can obtain the information shown in Table 5. The proposed generalization boundary strategy balances privacy and usability, satisfying the requirements of both data providers and data users.

4 Privacy-aware Access Control Model

Access control technology can be used as a starting point for managing PII data in a trustworthy fashion. But there are important principles of privacy management that require us to extend the traditional view of access control. Traditionally, the access decision is always binary; i.e., a data access is either 'allowed' or 'denied'. In this section, we present a specialized access control model that exploits the subtle nature of information privacy to maximize information utility with privacy guarantees.

4.1 Preliminaries

As a prerequisite to our model, supported elements have to be clearly defined and clarified.

- *Data Users*: Data users are individuals who access or receive data. Data users are required in a privacy context, as privacy policies will depend on the relationship between the individual requesting data and in the individual whom the data is related to. For example, one type of data users might be *physician* while another might be *primary care physician*. We denote U as the set of data users in this paper.
- *Privilege*: Some privacy policies make distinctions about who can perform activities based on the action being performed. For example, a policy might state that anyone in the company can *create* a customer record, but that only certain data users are allowed to *read* that record. We denote $Priv$ as the set of privileges.
- *Purposes*: Data access requests are made for a specific purpose or purposes. This represents how the data is going to be used by the recipient. For example, the data may be used for *Marketing* or *Delivery* purposes. We denote P as the set of purposes.
- *Generalization Level*: Generalization level refers to to what extent the data items have been generalized. For example, a *Low* generalization level on *Address* means the address information can be used without any modification. We denote GL as the set of private levels, which consists of L, M, H, and ML.
- *Retention Period*: Retention period refers to how long the information is stored. For example, if the retention period for $Name$ is one month, it means the name information can only be retained for one month. We use time intervals to describe retention period, e.g., [12/02/2008, 12/03/2008]. We denote T as the set of time intervals.

4.2 Authorizations

Since the access to the data items is strictly governed by the data provider's requirements, authorizations to access a data item in specific generalization level are usually required prior to the access. In addition to the traditional access factors: data items, data users and privileges, all the authorizations in this paper are extended to include the specific purpose and the generalization level of each data item.

Definition 3 *A generalized authorization is a 5-tuple* $(u, d, priv, p, gl)$*, where* $u \in U$*,* $d \in D$*,* $priv \in Priv$*,* $p \in P$*,* $gl \in GL$*.*

As previously mentioned, D is the set of data items. The tuple $(u, d, priv, p, gl)$ states that the data user u has been authorized to perform $priv$ on the data item d under generalization level gl for purpose p. For example, the tuple $(Tom, address, access, delivery, L)$ denotes that Tom was authorized with privilege *access* of the customer's *address* at *Low* generalization level for the *delivery* purpose.

Moreover, personal information shall be retained only as long as necessary for the fulfillment of the purpose for which it has been collected. If a certain data item was collected for a set of purposes, it is kept for the limited retention period of the purpose. We refer to an authorization together with its usage time as a temporal generalized authorization. A time interval is also associated with each authorization, imposing lower and upper bounds to the potential usage.

Definition 4 *A temporal generalized authorization is a 6-tuple* $(t, u, d, priv, p, gl)$*, where* $t \in T$*,* $u \in U$*,* $d \in D$*,* $priv \in Priv$*,* $p \in P$*,* $gl \in GL$*.*

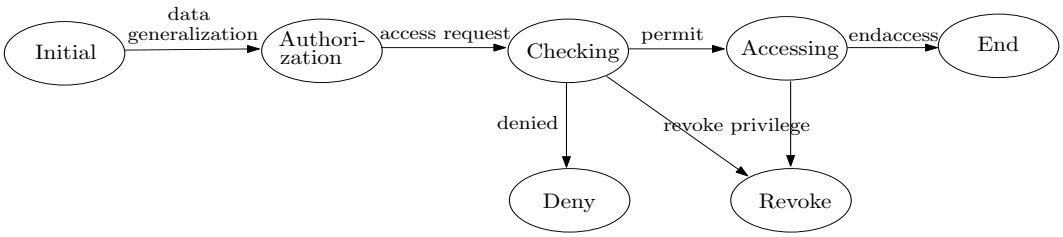

Figure 2: The state transition of privacy-aware access control actions

A tuple $([t_a, t_b], u, d, priv, p, gl)$ states that the data user u has been authorized to perform $priv$ on the data item d in the generalization level gl for the purpose p in the time interval $[t_a, t_b]$. We denote AU as the set of temporal generalized authorizations and $\sigma_{au}(*)$ as the function used to extract the element(s) $*$ in an authorization $au \in AU$.

A temporal generalized authorization $au = ([12/06/2008, 10/08/2008], Tom, income, read, admin, M)$, it means that, between Jun. 12, 2008 and Aug. 10, 2008, Tom was authorized the privilege to $read$ the customer's $income$ at the generalization level $Medium$ for the $admin$ purpose. Here, $\sigma_{au}(t)$ refers to the time interval [12/06/2008, 10/08/2008] and $\sigma_{au}(u, d, priv, p)$ returns to the tuple (Tom, income, read, admin).

4.3 Authorization rules

In this section, authorization rules are introduced to organize authorizations. The operations in the authorizations are to grant/revoke privileges to/from data users. Privileges are revoked in the following two situations:

(1) Revocation by time interval if the time interval of authorizations is expired.

(2) Revocation by generalization level if the data item is accessed in the wrong generalization level.

Further, an administrator of the system can make a forced revocation decision. For example, if a security administrator notices that a data user often sends many access requests without using services, the administrator may take actions, such as revoking his authorization, on this data user to prevent denial of service (DoS).

Based on this policy, three different attributes are required to meet these authorizations:

- *The time interval.* This includes the start time and end time for which access is permitted. At the end time, the privilege for using data items is revoked.
- *The valid period.* Access to a data item can be permitted only during the valid period of usage.
- *Generalization level.* The data item can only be accessed under the authorized generalization level.

The state transition of privacy-aware access control actions is given in Figure 2. The states and actions in Figure 2 are explained bellow.

(1) Initial: the initial state of the Metadata.

(2) Data generalization: replacing a data value with a less specific but semantically consistent value.

(3) Authorization: granting privileges of service to data users if data users meet authorization requirements of the system.

(4) Access request: a user request to access digital objects.

(5) Checking: checking the valid period of the authorization and the generalization level.

(6) Permitted and denied: if the time interval is not expired during the valid period, an access to data items is permitted, otherwise, denied.

(7) Accessing: during this state, data users are accessing data items. During the state of accessing, the accessed generalization level of the data item needs to be checked.

(8) Revoke privilege and endaccess: if the data item is accessed in a wrong generalization level, the system will revoke the privileges.

(9) Deny, Revoke and End: three final states. Deny is the sate of refusing to access without revoking privileges. Revoke is the state after the action of revoke privileges, while End is the state after the action of endaccess.

From the analysis of states and actions in privacy-aware access control, it is obvious that an access is not a simple action, but consists of a sequence of actions and active tasks.

5 Access control policy

After each data is granted with authorizations according to different purposes, an access request is needed to access the data items. In this paper, we assume that each access request is associated with an access time and a specific purpose. It is not trivial for a system to correctly infer the purpose of a query as the system must correctly deduce the actual intention of database users.

Definition 5 (Access request) *An access request is a 5-tuple $(t, u, d, priv, p)$ where $t \in T$ is the time when the access is requested, $u \in U$ is the data user who requires the access, $d \in D$ is the data item to be accessed, $priv \in Priv$ is a privilege exercised on the data, and $p \in P$ is the purpose for which the data is going to be used.*

The tuple $([t_a, t_b], u, d, priv, p)$ states that the data user u requests to perform $priv$ on the data item d for purpose p in the time interval $[t_a, t_b]$. We denote R as the set of access requests and for an access request $r \in R$, $\sigma_r(*)$ refers to the element(s) $*$ in an access request r.

For example, the access purpose $r = ([10/07/2008, 20/07/2008], Tom, income, read, admin)$ means that between Jul. 10, 2008 and Jul. 20, 2008, Tom requests to $read$ the customer's $income$ information for the $admin$ purpose. Here, $\sigma_r(t)$ refers to the time interval [10/07/2008, 20/07/2008].

As far as an authorization is concerned, the first step is to judge whether the authorization is valid or not. This is checked by the valid authorization function.

Definition 6 (Valid authorization function) *The valid authorization function is used to judge whether the current authorization au is valid. It can be expressed as follows:*

$$G(r) = \begin{cases} au & \text{if } \sigma_r(u,d,priv,p) = \sigma_{au}(u,d,priv,p) \\ & \text{and } \sigma_r(t) \subseteq \sigma_{au}(t) \\ \phi & \text{others} \end{cases}$$

where $au \in AU$, r is an access request and $\sigma_r(*)$ represents the element $*$ in r. $G(r)$ returns an authorization tuple. Except for checking the data user to perform the privileges on data items for the same purpose, the specific period of the current access request should also be checked, whether it is valid or not according to the period constraint of an authorization. When it is ϕ, the authorization is illegal.

However, a valid authorization is not enough for an access request. A valid authorized access request is a request for which an authorization exists in the current AU, which is checked by the following valid access function.

In order to express the definition of a valid access function conveniently, a useful expression is given for the relationship between the data item and the generalization level. If the data item $d \in D$ is accessed in the *Low* generalization level under the access request r, it is denoted by $\gamma_{(d,r)}(gl) = L$.

Definition 7 (Valid access function) *The valid access function is used to judge whether the access request is valid according to the current AU. It can be expressed as follows:*

$$F(r) = \begin{cases} true & \exists G(r), (\gamma_{(d,r)}(gl) = \sigma_{G(r)}(gl)) \\ false & others \end{cases}$$

where r is an access request. If $F(r)$ is true, the access is valid. Otherwise, it is invalid.

After a data user submits an access request r and $F(r)$ is true, the user is permitted to access the data. During the following process of accessing, there are three kinds of situations that should be considered. If a requested authorization tuple is a non-time authorization, the authorization au is revoked. If it is a temporal authorization, when the time exceeds the retention time, the au is illegal. If the data item being accessed is not in the same generalization level, access is rejected. The implementation of the access control policy is described in Table 6.

Algorithm: Access control(AU, r)

Input: an access request r and the set of current temporal generalized authorizations AU

$au = G(r)$; /**use the valid authorization to return a set function of authorization tuple, and then judge whether the authorization is valid**/

If $(au = \phi)$
then
Return false; /**This authorization does not exist**/

if $(\sigma_r(t) \notin \sigma_{au}(t))$
then
Return false; /**Illegal Authorization**/;

$k = F(r)$; /**use the valid access function to return a boolean value, with which to judge whether the access is valid.**/

if $(k = false)$
then
Return false; /**Illegal Access**/;

if $(\gamma_{(d,r)}(gl) \notin \sigma_{G(r)}(gl))$
then
Return false; /**The access is rejected'**/

Table 6: Implementation of Access Control Policy

6 Obligations and conditions

Traditional access control, such as Mandatory Access Control (MAC), Discretionary Access Control (DAC), and Role Based Access Control(RBAC) (Denning 1976, Sweeney 2002, Sandhu et al. 1996, Ferraiolo et al. 2001), are not designed to enforce privacy policies and barely meet privacy protection requirements, particularly, obligations and conditions. In these access control model, once the access or deny ruling has been given and properly logged, nothing more is needed from the access control function. But sometimes privacy laws and policies require that additional processes be started when an access of PII is made. As noted earlier, we refer to these activities as obligations which are incurred by the PII usage.

Some examples of such obligations might be:

- The state tax authority can use your financial records for the purposes of an audit, but only if the data subject of the audit is notified.
- A research company may use a data subjects human genome information for research purposes, but only if it pays the data subject 100 AUD per year.
- A minor's e-mail address may be used to communicate with the minor, but only if written consent is obtained from the legal guardian within 30 days.

Generally, obligations are associated with some action request; i.e., a subject promises to fulfill some obligations sometime in order to perform a specific action on some objects now. There are cases in which specific obligations are only associated with some special objects in the policies without reference to an action. However, a corresponding action can still be identified practically because usually the action making these objects special is the action causing these obligations.

Some obligations may be conditional; that is, conditional obligations are only required to be fulfilled if some related condition is true. Conditions typically include environmental or system-oriented decision factors. Examples are time of day and system load. They can also include the security status of the system, such as normal, high alert, under attack, etc. Conditions are not under direct control of individual subjects. Conditions, or prerequisites to be met before any action can be executed, are critical in some cases.

7 Related Work

To date, several approaches have been reported that deal with various aspects of the problem of high-assurance privacy systems.

The W3Cs Platform for Privacy Preference (P3P) (WWW) allows web sites to encode their privacy practice, such as what information is collected, who can access the data for what purposes, and how long the data will be stored by the sites, in a machine-readable format. P3P enabled browsers can read this privacy policy automatically and compare it to the consumers set of privacy preferences which are specified in a privacy preference language such as a P3P preference exchange language (APPEL) (WWW), also designed by the W3C. Even though P3P provides a standard means for enterprises to make privacy promises to their users, P3P does not provide

any mechanism to ensure that these promises are consistent with the internal data processing. By contrast, the work in our paper not only provides an effective generalization strategy to maximize the data privacy and usability but also provides more significant work on how to manage the a valid access process.

The concept of Hippocratic databases that incorporates privacy protection within relational database systems was introduced by Agrawal et al.(Agrawal et al. 2002). The proposed architecture uses privacy metadata, which consist of privacy policies and privacy authorizations stored in two tables. Byun et al. presented a comprehensive approach for privacy preserving access control based on the notion of purpose (Byun et al. 2004, 2005). In the model, purpose information associated with a given data element specifies the intended use of the data element, and the model allows multiple purposes to be associated with each data element. The granularity of data labeling is discussed in detail in (Byun et al. 2004), and a systematic approach to implement the notion of access purposes, using roles and role-attributes is presented in (Byun et al. 2005).

Although all these models do protect privacy of data providers, they are very rigid and do not provide ways to maximize the utilization of private information. Specifically, in those models access decision is always binary; i.e., a data access is either allowed or denied as in most conventional access control models. Different from previous models, the novelty of our model is that our approach can provide a much finer level of control as the access control decision is based on the question of "how much information can be allowed for a certain user", rather than "is information allowed for a certain user or not". In other words, every piece of information is classified into different generalization levels and every user is assigned an authorization to access the private information.

Previous work on multilevel secure relational databases (Jajodia & Sandhu 1991, Sandhu & Chen 1998) also provides many valuable insights for designing a fine-grained secure data model. In a multilevel relational database system, every piece of information is classified into a security level, and every user is assigned a security clearance. Based on this access class, the system ensures that each user gains access to only the data for which he has proper clearance, according to the basic restrictions. Byun and Bertino (Byun & Bertino 2006) proposed a new class of access control systems based on the notion of *micro-view*, which applied the idea of views at the level of the atomic components of tuples to an attribute value. However, the model in (Byun & Bertino 2006) is not to be considered a complete solution rather to show some of capabilities. Some technical challenges raised by their model have been solved in our paper. One of the challenges is to design metrics for data privacy and data usability, we solve this challenge by introducing the privacy-aware generalization boundary technique, which can maximize the privacy and utility for both data providers and data users. Another challenge is about applicability to genera-purpose access control, we solve it by providing a complete access control model with the implement of its access control policy. We further extend our proposed model by taking obligations and conditions into acconunt.

8 Conclusions and Future work

In this paper, we present a privacy-aware access control model with generalization boundaries. The devised generalization boundary technique can satisfy the requirement of both data providers and data users. Both privacy and usability of data items can be achieved when the data item is generalized by using this technique. Moreover, our approach can provide a much finer level of control as the access control decision is based on the question of "how much information can be allowed for a certain user", rather than "is information allowed for a certain user or not". The privacy-aware access control model we presented in this paper provides an example for multilevel secure relational databases.

Our proposed model provides efficient generalization strategies for privacy preserving access control systems, but much more work still remains to be done. The future work includes devising a high level language in which privacy specifications can be expressed precisely. We also plan to extend our model to cope with complex query processing. We will introduce the queries with join, sub-queries or aggregations into our model. These are challenging problems, but they are vital elements of privacy protection.

Acknowledgement

We would like to thank anonymous reviewers for their useful comments on this paper. This research is supported by Australian Research Council (ARC) grant DP0663414.

References

Adam, N. R. & Worthmann, J. C. (1989), Security-control methods for statistical databases: a comparative study. CSUR, 21(4):515-556, 1989.

Agrawal, R. & Kiernan, J. & Srikant, R. & Xu, Y. (2002), Hippocratic databases. *In: Proceedings of the 28th International Conference on Very Large Databases* (VLDB) (2002)

Agrawal, R. & Evmievski, A. & Srikant, R. (2003), Information sharing across private databases. *In Proc. of the 2003 ACM SIGMOD Int. Conf. on Management of Data.* ACM Press, 2003.

Ashley, P. & Powers, C.S. & Schunter, M. (2002), Privacy promises, access control, and privacy management. *In: Third International Symposium on Electronic Commerce* (2002)

Byun, J. W. & Bertino, E. (2006), Micro-views, or on how to protect privacy while enhancing data usability: concepts and challenges. *SIGMOD Record* 35(1): 9-13 (2006).

Byun, J. W. & Bertino, E. & Li, N. (2004), Purpose based access control for privacy protection in relational database systems. *Technical Report* 2004-52, Purdue University, 2004.

Byun, J. W. & Bertino, E. & Li, N. (2005), Purpose based access control of complex data for privacy protection. *In Symposium on Access Control Model And Technologies* (SACMAT), 2005.

Dong, X. & Madhavan, J. & Nemes, E. (2005), Reference reconciliation in complex information spaces. *In ACM International Conference on Management of Data* (SIGMOD), 2005.

Denning, D.E. (1976), A lattice model of secure information flow, Communications of the ACM, vol. 19, no. 5, 1976.

Iyengar, V. (2002), Transforming Data to Satisfy Privacy Constraints, *In Proc. of the ACM SIGKDD* (2002), pp. 279-288.

Jajodia, S. & Sandhu, R. (1991), Toward a multilevel secure relational data model. *In: ACM International Conference on Management of Dat*a (SIGMOD) pp. 50-59. ACM Press, New York (1991).

LeFevre, K. & Agrawal, R. & Ercegovac, V. & Ramakrishnan, R. & Xu, Y. & DeWitt, D. (2004), Disclosure in hippocratic databases. *In: The 30th International Conference on Very Large Databases* (VLDB) (2004)

LeFevre, K. & DeWitt, D. & Ramakrishnan, R. (2006), Mondrian Multidimensional K-Anonymity, *in Proc. of the IEEE ICDE* (2006), pp.25.

Lunacek, M. & Whitley, D. & Ray, I. (2006), A Crossover Operator for the k-Anonymity Problem, *in Proc. of the GECCO* (2006) pp. 1713-1720.

Samarati, P. (2001), Protecting Respondents Identities in Microdata Release, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 13, No. 6 (2001), pp. 1010-1027.

Seamons, K. & Winslett, M. & Yu, T. (2001), Limiting the Disclosure of Access Control Policies during Automated Trust Negotiation. *In Proc. of NDSS01*, pp. 109-125. IEEE Press, 2001.

Sandhu, R. & Chen, F. (1998), The multilevel relational data model. *ACM Trans. Inf. Syst. Secu.* 1(1), pp.93-132 (1998)

Sandhu, R. & Coyne, E. & Feinstein, H. & Youman, C. (1996), Role Based Access Control Models, *IEEE Computer*, 29, (2), pp.38-47, 1996.

Ferraiolo, D.F. & Sandhu, R. & Gavrila, S. & Kuhn, D.R. & Chandramouli, R. (2001), Proposed nist standard for role-based access control. *ACM Trans. Inf. Syst. Secur.*, 4(3):224-274, 2001.

Sarawagi, S. & Bhamidipaty, A. (2002), Interactive deduplication using active learning. *In ACM International conference on Knowledge discovery and data mining* (SIGKDD), 2002.

Sweeney, L. (2002), Achieving k-Anonymity Privacy Protection Using Generalization and Suppression, *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems*, Vol. 10, No. 5 (2002), pp. 571-588.

Tumer, A. & Dogac, A. & Toroslu, H. (2003), A Semantic based Privacy Framework for Web Services. *In Proc. of ESSW03*, 2003.

Westin, A. (1998), E-commerce and privacy: What net users want. *Technical report*, Louis Harris & Associates, June 1998.

Westin, A. (1999), Freebies and privacy: What net users think. *Technical report*, Opinion Research Corporation, July 1999.

World Wide Web Consortium (W3C). A P3P Preference Exchange Language 1.0 (APPEL 1.0). Available at `www.w3.org/TR/P3P-preferences`.

World Wide Web Consortium (W3C). Platform for Privacy Preferences (P3P). Available at `www.w3.org/P3P`.