

Limiting Privacy Disclosure in Relational Databases*

Min Li

PhD student enrolled in July, 2007
Supervised by Dr. Hua Wang and Dr. Ashley Plank

Department of Mathematics & Computing
University of Southern Queensland
Toowoomba, Queensland, Australia
Email: limin@usq.edu.au

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 [10, 16]. Privacy issues are further exacerbated by the Internet which makes it easy for new data to be automatically collected and added to databases [18, 19].

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.

Data and privacy protection have become critical issues in the development of information systems. Not only is a secure infrastructure required but also appropriate access control technology. This reflects the growing attention of customers to their personal information and the increasing number of laws, policies, and regulations that are intended to safeguard it.

2 Related Work

As privacy becomes a major concern for both customers and enterprises, many privacy protecting access control models have been proposed [1, 2, 4, 13]. 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 [17, 3, 15].

To our best knowledge, the most well known effort is the W3C's Platform for Privacy Preference (P3P) [9]. P3P allows websites to express their privacy policy in a machine readable format so that using a software agent, consumers can easily compare the published privacy policies against their privacy preferences. While P3P provides a mechanism for ensuring that users can be informed about privacy policies before they release personal information, some other approaches are proposed [5, 8, 12, 14, 20], where the notion of *purpose* plays an important role in order to capture the intended usage of information.

Together with the notion of *purpose*, current privacy legislations also define the privacy principles that an information system has to meet in order to guarantee customer's privacy [11, 2, 3, 17]. Mechanism for negotiation is presented by Tumer et al. [17]. Enterprises specify which information is mandatory for achieving a service and which is optional, while customers specify the type of access for each part of their personal information. The concept of Hippocratic databases that incorporates privacy protection within relational database systems was introduced by Agrawal et al. [2]. 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 [7, 8]. 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 [7], and a systematic approach to implement the notion of access purposes, using roles and role-attributes is presented in [8].

3 Existing Problems and Potential Topics

In this section, we provide some potential research topics regarding privacy preserving in relational databases.

3.1 Overview of Hippocratic Databases

Hippocratic databases [2] use purpose as a central concept and consider it as a special attribute occurring in every tables forming the database and associated with each piece of data stored in the database.

For example, Table 1 shows the schema of two tables, customer and order, that store the personal information collected by Mississippi. Then, for each purpose and for each data item stored in the database, we have:

*This research is supported by Australian Research Council (ARC) grant DP0663414.

table	attribute
customer	purpose, customer-id, name, address, email, fax-number, credit-card-info
order	purpose, customer-id, transaction, book-info, status

Table 1: Database schema

table	attribute
privacy-policies	purpose, table, attribute, {external-receipts}, retention
privacy-authorizations	purpose, table, attribute, {authorized-users}

Table 2: Privacy metadata schema

External-recipients: the actors to whom the data item is disclosed;

Retention-period: the period during which the data item should be maintained;

Authorized-users: the users entitled to access the data item.

Purpose, external recipients, authorized users, and retention period are stored in the database with respect to the metadata schema defined in Table 2. Specifically, the above information is split into separate tables: external-recipients and retention period are in the *privacy-policies table*, while authorized-users in the *privacy-authorizations table*. The purpose is stored in both of them. The privacy-policies table contains the privacy policies of the enterprise, while privacy-authorizations table contains the access control policies that implement the privacy policy and represents the actual disclosure of information. In particular, privacy-authorizations tables are derived from privacy-policies tables by instantiating each external recipient with the corresponding users. Therefore, Hippocratic database systems define one privacy-authorizations table for each privacy-policies table, and these tables represent what information is actually disclosed.

Hippocratic systems are an elegant and simple solution but do not allow for dynamic situations that could arise with web services and business process software. In such settings, enterprises may provide services in many different ways and may delegate the execution of parts of the service to third parties. This is indeed the case of a virtual organization based on business process for web service where different partners explicitly integrate their efforts into one process.

3.2 Challenges

On the basis of the solution for the exchange between enterprises and customers, Hippocratic databases enforced fine-grained disclosure policies to an architecture at the data level [2]. In the proposed architecture, enterprises declared the purpose for which the data are collected, who can receive them, the length of time the data can be retained, and the authorized users who can access them. Hippocratic databases also created a privacy authorization table shared by all customers, but it does not allow to distinguish which particular method is used for fulfilling a service. Moreover, enterprises are able to provide their services in different ways, and each different method may require different data. For example, notification can be done by email or by mobile phone or by fax. Depending on the different kinds of methods, customers should provide different personal information. Asking for all personal information for different service methods as compulsory would clearly violate the principle of minimal disclosure.

On the server side, a single enterprise usually could not complete all procedures of a service by itself,

rather a set of collaborating organizations participating in the service. Enterprises might need to decompose a generic purpose into more specific sub-purposes since they are not completely able to fulfill it by themselves, and so they may delegate the fulfillment of sub-purposes to third parties. It is up to customers to decide on a strategy of how to get a service fulfilled on the basis of their personal feeling of trust for different service components. A question that many customers have when interacting with a web server, with an application, or with an information source is “Can I trust this entity?”. Different customizations may require different data for which considerations may vary; there might be different trust levels on different partners (sub-contractors). The choice of service customization has significant impact on the privacy of individual customers.

3.3 Privacy protecting access control

Starting from the landmark proposals for Hippocratic databases [2], most privacy-aware technologies use purpose as a central concept around which privacy protection is built. Byun and Bertino [6] 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.

Following [6], privacy level, types of data and possible data usages (i.e., purposes) are defined in Table 3. 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 provider¹ 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*.

¹Data 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 3: 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 4: Privacy information and Metadata

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

Table 5: Private information for *Delivery* purpose

Table 4 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*.

3.4 Challenges

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 5. 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.

The use of data generalization² can significantly increase the comfort level of data providers; i.e., the personal information can be collected in a generalized form.

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

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.

²Data generalization refers to techniques that “replace a value with a less specific but semantically consistent value.”

4 Methodology and Significance

In this section, we briefly discuss the approaches that we adopt in dealing with the potential research problems and the significance of our research.

4.1 Privacy protection in hippocratic database with delegation

In the last few years data and privacy protection have become critical issues in the development of information systems. This reflects the growing attention of customers to their personal information and the increasing number of laws, policies, and regulations that are intended to safeguard it. Hippocratic databases offer mechanisms for enforcing privacy rules in database systems for inter-organizational business processes, but do not allow to distinguish which particular method is used for fulfilling a particular purpose. To meet this requirement, We will organize purposes into purpose directed graphs through AND/OR decomposition, which supports task delegations and distributed authorizations. Specially, customers have controls of deciding how to get a service fulfilled on the basis of their personal feeling of trust for any service customization. Quantitative analysis will be performed to characterize privacy penalties dealing with privacy cost and customer’s trust. Finally, efficient algorithms will be given to guarantee the minimal privacy cost and maximal customer’s trust involved in a business process.

4.2 Privacy-aware access control with generalization boundaries

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?

To answer this question, we will 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 will discuss how to manage a valid access process and analysis the access control policy.

4.3 Significance

My PhD research focuses exclusively on how to specify and enforce policies for protecting private information in relational database systems. This research will extend the current mechanisms in order to support inter-organizational business processes in Hippocratic databases. A comprehensive approach for negotiation of personal information between customers and enterprises based on user preferences is developed when enterprises offer their clients a number of ways to fulfill a service. We will continue our work to propose new mechanisms in privacy protecting systems for enhancing current privacy methods with respect to balancing data privacy and usability and by proposing new access control models in relational database systems.

5 Research Outcomes

Related to the above research work, three research papers have been published by Australia & International Conferences.

- M. Li and H. Wang, Protecting Information Sharing in Distributed Collaborative Environment, *In: The 10th Asia-Pacific Web Conference Workshop (APWeb'2008)*, pages:192-200, April 26-28, 2008, Shenyang, China.

- M. Li, H. Wang, A. Plank and J. Yong, Advanced Permission-Role Relationship in Role-Based Access Control, *In: 13th Australasian Conference on Information Security and Privacy (ACISP'2008)*, pages: 391-403, July 7-9, 2008, Wollongong, Australia.

- M. Li and H. Wang, ABDM: An Extended Flexible Delegation Model in RBAC, *In: 8th IEEE International Conference on Computer and Information Technology (CIT'2008)*, pages: 390-395, July 8-11, 2008, Sydney, Australia.

6 Conclusion

Above is a brief introduction of my intended work during the PhD study and we will target at top level conference and high impact journal paper publication. We appreciate any expert suggestion concerning privacy protecting in relational database systems.

References

- [1] N.R. Adam and J.C. Worthmann, Security-control methods for statistical databases: a comparative study. *CSUR*, 21(4):515-556, 1989.
- [2] R. Agrawal, J. Kiernan, R. Srikant and Y. Xu, Hippocratic databases. *In: Proceedings of the 28th International Conference on Very Large Databases (VLDB) (2002)*
- [3] R. Agrawal, A. Evmievski and R. Srikant. Information sharing across private databases. *In Proc. of the 2003 ACM SIGMOD Int. Conf. on Management of Data*. ACM Press, 2003.
- [4] P. Ashley, C.S. Powers and M. Schunter, Privacy promises, access control, and privacy management. *In: Third International Symposium on Electronic Commerce (2002)*
- [5] M. Backes, B. Pfitzmann and M. Schunter, A Toolkit For Managing Enterprise Privacy Policies. In: *Proceedings Of Esorics'03, Lncs 2808*, pp. 162-180. Springer, Berlin Heidelberg New York (2003)
- [6] J.W. Byun and E. Bertino: Micro-views, or on how to protect privacy while enhancing data usability: concepts and challenges. *SIGMOD Record* 35(1): 9-13 (2006).
- [7] J.W. Byun, E. Bertino and N. Li, Purpose based access control for privacy protection in relational database systems. *Technical Report 2004-52*, Purdue University, 2004.
- [8] J.W. Byun, E. Bertino and N. Li, Purpose based access control of complex data for privacy protection. *In Symposium on Access Control Model And Technologies (SACMAT)*, 2005.
- [9] L. Cranor, M. Langheinrich, M. Marchiori, and J. Reagle: The platform for privacy preferences 1.0 (P3P1.0) specification. W3C recommendation (2002). <http://www.w3.org/TR/P3P/>
- [10] X. Dong, A. Halevy, J. Madhavan and E. Nemes, Reference reconciliation in complex information spaces. *In ACM International Conference on Management of Data (SIGMOD)*, 2005.
- [11] E. Ferrari and B.M. Thuraisingham, Security and privacy for web databases and services. In: *Proceedings of the 9th International Conference on Extending Database Technology, LNCS 2992*, pp. 17-28. Springer, New York (2004).
- [12] G. Karjoth, M. Schunter and M. Waidner: Platform for enterprise privacy practices: privacy-enabled management of customer data. In: *Proceedings of PET'02, LNCS 2482*, pp. 69-84. Springer, Berlin Heidelberg New York (2002)
- [13] K. LeFevre, R. Agrawal, V. Ercegovac, R. Ramakrishnan, Y. Xu and D. DeWitt, Disclosure in hippocratic databases. *In: The 30th International Conference on Very Large Databases (VLDB) (2004)*
- [14] F. Massacci and N. Zannone: Privacy is linking permission to purpose. In: *Proceedings of the 12th International Workshop on Sec. protocols (2004)*
- [15] K. Seamons, M. Winslett, and T. Yu. Limiting the Disclosure of Access Control Policies during Automated Trust Negotiation. *In Proc. of NDSS01*, pp. 109-125. IEEE Press, 2001.
- [16] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. *In ACM International conference on Knowledge discovery and data mining (SIGKDD)*, 2002.
- [17] A. Tumer, A. Dogac, and H. Toroslu. A Semantic based Privacy Framework for Web Services. *In Proc. of ESSW03*, 2003.
- [18] A. Westin. E-commerce and privacy: What net users want. *Technical report*, Louis Harris & Associates, June 1998.
- [19] A. Westin. Freebies and privacy: What net users think. *Technical report*, Opinion Research Corporation, July 1999.
- [20] M. Yasuda, T. Tachikawa and M. Takizawa: Information flow in a purpose-oriented access control model. In: *Proceedings of ICPADS'97*, pp. 244-249. IEEE Press, Lausanne (1997)