Graduate School ETD Form 9
(01/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis Acceptance

This is to certify that the thesis prepared

By Ji-Won Byun

Entitled Toward Privacy-Preserving Database Management Systems
-- Access Control and Data Anonymization

Complies with University regulations and meets the standards of the Graduate School for originality and quality

For the degree of Doctor of Philosohpy

Final examining committee members

Elisa Bertino , Chair

Ninghui Li

Mikail Atallah

Sunil Prabhaka

Approved by Major Professor(s): Elisa Bertino

Ninghui Li

Approved by Head of Graduate Program: Susanne Hambrusch

Date of Graduate Program Head's Approval: March 1, 2007

TOWARD PRIVACY-PRESERVING DATABASE MANAGEMENT SYSTEMS – ACCESS CONTROL AND DATA ANONYMIZATION

A Thesis

Submitted to the Faculty

of

Purdue University

by

Ji-Won Byun

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2007

Purdue University

West Lafayette, Indiana

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

To my family for their absolute love and support.

ACKNOWLEDGMENTS

I thank my advisors, Prof. Elisa Bertino and Prof. Ninghui Li. Their dedication to research has been truly inspirational, and their tremendous insight and knowledge in the field of information security have guided me throughout my graduate study. It has been my greatest honor to work with and learn from them, and I humbly hope that this privileged relationship will be continued even after I leave Purdue.

I also owe a great debt to the many wonderful people in the Center for Education and Research in Information Assurance and Security (CERIAS). The educational environment provided by CERIAS has enabled me to learn and explore many topics in information security.

Lastly, but certainly not least, I also thank all the faculty, staffs, and graduate students in the department of Computer Science from whom I have benefited in various ways.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Byun, Ji-Won Ph.D., Purdue University, May, 2007. Toward Privacy-Preserving Database Management Systems – Access Control and Data Anonymization. Major Professors: Elisa Bertino and Ninghui Li.

In this thesis, we identify basic requirements for privacy-preserving DBMS and focus on two core techniques, namely *purpose-based access control* and *data anonymization*, that are essential to address some of the requirements. Specifically, purpose-based access control enables DBMS to tightly control data access with respect to privacy requirements and preferences, and data anonymization provides a way to guarantee privacy protection in data itself even if the control of access is not feasible. We present formal models and develop mechanisms for realizing such models. In addition, we introduce two conceptual models, *micro-view* and *integrity-control*, which are designed to enhance data utility and integrity, respectively.

1 INTRODUCTION

The phenomenal advance in information technology over the past few decades has literally transformed our life. Particularly, the explosive growth of the Internet and e-commerce has enabled people to carry out daily activities online, for example, online shopping, e-banking, and even consulting a doctor over the Internet. Although most people are not aware of it, such prevalent online activities imply that a vast amount of personal data is electronically produced and collected continuously. In fact, a survey conducted by Federal Trade Commission (FTC) in March 2000 showed that as many as 90 million Americans were using the Internet on a regular basis and that 97 percent of web sites were collecting at least one type of identifying information such as name, e-mail address, or postal address of consumers [1]. The proliferation of ubiquitous computing also significantly increases the collection of personal data. The fast growing number of various hand-held devices, GPS, and RFID tags elevates the amount and the resolution of personal data that can be massively collected in real-time.

Such collected data represent an important asset today as they can be used for various purposes, ranging from scientific research to demographic trend analysis or marketing purposes. For instance, medical researchers may observe breakthrough patterns from a collection of patient records, or government agencies may make critical decisions based on various data collected by them. In addition to the useful knowledge that can be extracted from data, the monetary reward of using collected data itself is also considerably high, and many enterprises gain a huge profit by either using or selling personal data. For example, Equifax, a multinational credit reporting company, earns nearly two billion dollars per year from collecting and distributing personal information [2].

While the collection and use of personal data is accepted as a common business practice today, this trend raises a significant concern for information privacy. In fact, having observed many privacy related incidents [3–6], individuals are afraid that their personal information might fall into a wrongful hand and be abused against their will. A survey of FTC [1] well illustrates this fear: 92 percent of consumers were concerned about the misuse of their personal information online. In what follows, we discuss the implication of personal information for privacy and develop a clear understanding of information privacy which is the focal issue addressed in this thesis.

1.1 Information Privacy

Although privacy has long been recognized as a fundamental human right, privacy is a difficult concept to comprehend in a concise manner, mainly due to its heavy dependency on context and social/cultural setting. Colin Potts, a professor of College of Computing at Georgia Institute of Technology, illustrated in his presentation [7] that privacy could be interpreted to mean many different concepts. According to Potts, computer scientists often consider privacy as a confidentiality issue and place privacy within the scope of information security. However, lawyers and ethicists view privacy as the right "to be let alone", which places privacy within the bounds of personal freedom. There are some people who believe that personal information belongs to users, which turns privacy into a matter of intellectual property.

Although there exist various interpretations for privacy, in this thesis we define information privacy as the ability of a person to control the availability of information and exposure of oneself[1]. We believe that this definition precisely echoes the concern that individuals have over their personal information today. As discussed previously, individuals are concerned that their personal information might be misused against their intention. In addition to potential harm or loss imposed by privacy violations,

[1] Long before the information age began, Alan F. Westin, a professor emeritus of public law and government at Columbia University, defined privacy as "the right to select what personal information about me is known to what people" [8].

Table 1.1
Privacy regulations: Fair information practice and basic privacy principles

Fair Information Practices	**1. Notice.** Provide consumers clear notice of their information practices, including what information they collect, how they collect it, how they use it, etc. **2. Choice.** Offer consumers choices as to how their personal identifying information is used beyond the use for which the information was provided. **3. Access.** Offer consumers reasonable access to the information a Web site has collected about them, including a reasonable opportunity to review information and to correct inaccuracies or delete information. **4. Security.** Take reasonable steps to protect the security of the information they collect from consumers.
Basic Privacy Principles	**1. Collection Limitation Principle.** There should be limits to the collection of personal data and any such data should be obtained by fair means and, where appropriate, with the knowledge or consent of the data subject. **2. Data Quality Principle.** Personal data should be relevant to the purposes for which they are to be used and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date. **3. Purpose Specification Principle.** The purposes for which personal data are collected should be specified and the subsequent use limited to the fulfillment of those purposes. **4. Use Limitation Principle.** Personal data should not be disclosed, made available or otherwise used for purposes except: a) with the consent of the data subject; or b) by the authority of law. **5. Security Safeguards Principle.** Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data. **6. Openness Principle.** There should be a general policy of openness about developments, practices and policies with respect to personal data. **7. Individual Participation Principle.** An individual should have the right: a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him; b) to have communicated to him, data relating to him; c) to be given reasons if a request made under subparagraphs (a) and (b) is denied, and to be able to challenge such denial; and d) to challenge data relating to him and, if the challenge is successful to have the data erased, rectified, completed or amended. **8. Accountability Principle.** A data controller should be accountable for complying with measures which give effect to the principles stated above.

individuals also feel uncomfortable with the fact that the personal information they release for a particular purpose might be used for various other purposes without their consent or awareness. In other words, the primary focus of information privacy lies on the awareness of data collection and the control of personal data. In fact, the awareness and control are the key issues addressed in many privacy-related legislations. For example, Table 1.1 summarizes two sets of privacy regulations, Fair Information Practice from FTC [1] and Basic Privacy Principles from Organization for Economic Co-operation and Development (OECD) [9], both of which explicitly address the issue of awareness and control. Note that more recent regulations, such as the Health Insurance Portability and Accountability Act (HIPPA) [10], the Graham-Leach-Bliley Act (GLBA) [11], and EU Data Directive [12], establish similar ground rules.

As an additional note, we emphasize that information privacy implies much more than the confidentiality of personal information. Stewart Baker, assistant secretary for policy for the Department of Homeland Security and former general counsel of the National Security Agency, acknowledges this challenging issue in [13].

> The biggest threats to our privacy in a digital world come not from what we keep secret but from what we reveal willingly. We lose privacy in a digital world because it becomes cheap and easy to collate and transmit data, so that information you willingly gave a bank to get a mortgage suddenly ends up in the hands of a business rival or your ex-spouse's lawyer.

In other words, the problem of information privacy is not about how to conceal personal information, but how to ensure that such information is disclosed only under appropriate circumstances.

1.2 Requirements for Privacy-Preserving DBMS

The recent privacy-related regulations have made many organizations aware of the importance of privacy protection. However, such regulations are not the only incentive for organizations to protect individuals' privacy. For many businesses, especially e-commerce, consumers' concern for privacy is directly translated to a huge financial loss. A survey report from Forrester Research [14] states that individual privacy concerns reduced 15 billion dollars in e-commerce in 2001 alone. Privacy is a key concern for other types of organizations as well. Many companies and agencies nowadays try to demonstrate good privacy practices in order to build solid trust and provide more confidence to their customers [15]. Therefore, the demand for privacy protection technology today is stronger than ever. The W3C's Platform for Privacy Preference (P3P) [16] is a solution widely adopted to address this need. P3P is an industry standard that provides a simple, automated method for users to gain control over the use of their personal information on Web sites they visit. P3P allows Web sites to encode their privacy practice, such as what data is collected, who can access those data for what purposes, and how long the data will be stored by the sites, in a machine-readable XML format.

Languages for specification of privacy promises, however, represent only one of the components in a comprehensive solution to privacy [17]. It is crucial that once data are collected, privacy promises be enforced by the information systems managing them. Because in today's information systems data are in most cases managed by DBMS, the development of DBMS properly equipped for the enforcement of privacy promises and of other privacy policies is crucial. Here we discuss a set of requirements towards the development of such DBMS. Some of those requirements, such as the support for purpose meta-data and privacy obligations, are derived directly from P3P. Other requirements are not directly related to P3P; however, they are crucial for the development of DBMS able to support a wide range of privacy policies, going beyond the ones strictly related to P3P. In what follows, we use the terms *data users* to denote the active entities, trying to gain accesses to the data, and *data subjects* the passive entities, whose privacy is to be protected.

R1. Support for rich privacy related metadata An important characteristic of P3P is that very often privacy policy, that is, statements specifying the use of the data by the party collecting them, include the specification of the intended use of the data by the collecting party as well as other information. Examples of this additional information are how long the data will be kept and possible actions that are to be executed whenever a subject accesses the data. Supporting this additional information calls for the need of privacy-specific metadata that should be associated with the data, stored in the database together with the data, and send with the data whenever the data flow to other parties in the system. Metadata should be associated with the data according to a range of possible granularities. For example in a relational database, one should be able to associate specific metadata with an entire table, with a single tuple, or even with a column within a single tuple. Such flexibility should not, however, affect the performance; thus we need to develop highly efficient techniques for managing these metadata in particular when dealing with query executions. Query executions may need to take into account the contents of such metadata in order to filter out the data that cannot be accessed because of privacy constraints from the data to be returned.

R2. Support for expressive attribute-based descriptions of data users We see an increasing trend towards the development of access control models that relies on information concerning data users. Examples of such models are represented by trust negotiation systems [18, 19], that use credentials certifying relevant properties of data users. Such access control models are crucial in the context of privacy because they provide a high-level mechanism able to support a very detailed specification of the conditions that data users must verify in order to access data. As such, fine-grained privacy-preserving access control policies can be supported. They also make it easy to formulate and maintain privacy policies and verify their correctness. Moreover, such high-level models can provide better support for interoperability because they can, for example, easily integrate with Security Assertion Markup Language (SAML) assertions. However, current database technology is very poor in the representation of data users. At the best current DBMS provide support for roles in the context of the well-known role-based access control (RBAC) model [17, 20]. However, apart from this, DBMS do not provide the possibility of specifying application-dependent user profiles for use in access control and privacy enforcement. It can be argued that such profiles should perhaps be built on top of the DBMS or even be supported externally. However, in such a case, it is not clear how efficient access control and privacy enforcement could be supported. It also important to notice that RBAC does not support data user attributes. Extensions of RBAC models supporting such a feature should be devised.

R3. Support for obligations Obligations specify privacy-related actions that are to be executed upon data accesses for certain purposes. There is a large variety of actions that can be undertaken, including modifications to the data, deletion of the data, notifications of data access to the individual to whom the data are related or to other individuals, insertion of records into privacy logs. These obligations should be possibly executed, or at least initiated by, the DBMS because their execution is tightly coupled with data accesses. An important issue here is the development of expressive languages supporting the specification obligations, and analysis tools to verify the correctness and consistency of obligations. A viable technology to support obligations is represented by trigger mechanisms, currently available in all commercial DBMS. The main question is however whether current trigger languages are adequate to support the specification of obligations.

R4. Fine-grained access control to data The availability of a fine-grained access control mechanism is an important requirement of a comprehensive solution to privacy. Conventional view mechanisms, the only available mechanism able to support in some ways a very fine granularity in access control, have several shortcomings. A naive solution to enforce fine-grained authorizations would require specifying a view for each tuple or subset of a tuple that are to be protected. Moreover, because access control policies are often different for different users, the number of views would further increase. Furthermore, applications programs would have to code different interfaces for each user, or group of users, because queries and other data management commands would need to use for each user, or group of users, the correct view. Modifications to access control policies would also require creation of new views with consequent modifications to application programs. Alternative approaches that address some of those issues have been proposed that are based on the idea that queries are written against base tables and then automatically re-written by the system against the view available to the user. These approaches do not require to code different interfaces for different users, and thus address on of the main problems in the use of conventional view mechanisms. However, they introduce other problems, such as inconsistencies between what the user expects to see and what the systems returns; in some cases, they return incorrect results to queries rather than rejecting them as unauthorized. Different solutions thus need to be investigated. These solutions must not only address the specification of fine-grained access control policies but also their efficient implementation in current DBMS.

R5. Privacy-preserving information flow In many organizations, data flow across different domains. It is thus important that privacy policies related to data "stick" with the data when these data move within an organization or across organizations. It is crucial to assure that if data have been collected under a given privacy promise from an individual, this promise is enforced also when the data are passed to parties different from the party that have initially collected them. Information flow has been extensively investigated in the past in the area of multi-level secure databases. An important issue is to revisit such theory and possibly extend it for application in the context of privacy.

R6. Development of advanced view mechanisms An important type of access control for database systems is represented by the view mechanism. Such a mechanism, that has been generalized by techniques such as Oracle VPD [21] and materialized views, not only supports content-based access control, but also allows one to give users (or sets of users) a particular vision of the data through transformation. Examples of such transformations include aggregation of data, resulting in aggregate views, and elimination of columns. However, such views are not sufficient to assure privacy. Today, the enormous amount of personal information, combined with powerful data mining techniques [22–24], has become a very serious threat to individuals' privacy. Therefore, more powerful view mechanisms are needed to be able to support more sophisticated data transformation. In particular, an important class of transformation is represented by anonymization where a certain amount of information is hidden (or modified) so that the data does not reveal the identity of data subjects. Some important requirements for a view mechanism integrated with anonymization techniques include the maximization of data quality and the support for dynamic updates to the source database. In particular, it is critical to ensure that knowledge extracted from an anonymized view of the data be as much as possible of high quality; this requirement is crucial in assuring that useful knowledge can still be extracted from such a view. Reflecting changes to the source database into the anonymized view is also important with respect to the timeliness requirement. Such requirement must, however, be addressed while preserving privacy.

1.3 Contributions and Organization of This Work

This thesis aims to help developing a comprehensive privacy-preserving DBMS by addressing some of the essential requirements mentioned in the previous section. Some efforts have already been reported dealing with DBMS specifically tailored to support privacy policies. In particular, Agrawal et al. [25] have recently introduced the concept of Hippocratic databases, incorporating privacy protection in relational database systems. Their work introduces the fundamental principles underlying Hippocratic databases and proposes a reference architecture. An important feature of their architecture is that it uses some privacy metadata, consisting of privacy policies and privacy authorizations stored in privacy-policies tables and privacy-authorizations table respectively. Although some follow-up effort has been made, the development of privacy-preserving DBMS is yet at a very preliminary stage. It is important to notice that privacy-preserving DBMS may have to be combined with collateral tools, such as data anonymizer and metadata manager, in order to provide comprehensive platforms for supporting flexible and articulated privacy-preserving information management. The main objective of this thesis is to develop models and techniques for building a privacy-preserving DBMS with this regard.

1.3.1 Privacy Respecting Access Control

In Chapter 2, we introduce Purpose-Based Access Control (PBAC) model, which directly addresses the issue of individuals' control over their personal data. A key challenge of privacy-centric access control lies in the fact that traditional access control models are not designed to support privacy protection. That is, while traditional access control models mainly focus on which user is performing which action on which data object, privacy policies are concerned with which data object is used for which purpose(s). As such, the notion of purpose plays a major role in PBAC. Specifically, PBAC utilizes privacy-related metadata, namely *intended purposes*, which specify the intended usage of data, and *access purposes*, which specify the purposes for which a given data element is accessed. To facilitate the purpose management, both intended purposes and access purposes are specified with respect to a hierarchical structure that organizes a set of purposes for a given enterprise. In Section 2.1, we formally introduce the notions of purposes and also develop the notion of *purpose compliance*, which is the basis for verifying that the purpose of a data access complies with the intended purposes of the data.

As another key contribution, we address the problem of how to determine the purpose for which certain data are accessed by a given user in Section 2.4. In our approach, users are required to state their access purposes along with their queries, and the system validates the stated access purposes by ensuring that the users are indeed allowed to access data for the particular purposes. To ease the management of access purpose authorizations, we develop an approach which relies on the well-known Role Based Access Control (RBAC) model [20, 26–28]. This method has a great deployment advantage as many systems are already using RBAC mechanisms for the management of access permissions. This approach is also reasonable as access purposes can be granted to the tasks or functionalities over which roles are defined within an organization. However, using an RBAC mechanism for the management of both access permissions and access purposes may increase the complexity of the role engineering tasks. To address this problem, we introduce a simple extension to RBAC, which simplifies the role administration and also provides increased flexibility.

Another important issue that we address in the chapter is the granularity of data labeling; that is, the units of data with which purposes can be associated. In Section 2.5.1, we address this issue in the context of relational databases and propose four different labeling schemes, each providing a different granularity. Using our approach it is thus possible to associate a purpose (or a set of purposes) with an entire table, with each column within a table, with each tuple within a table, or with each attribute within a tuple. In Section 2.5.2, we present an approach to representing purpose information, which results in very low storage overhead. Furthermore, we exploit query modification techniques to support data filtering based on purpose information. Such techniques ensure efficient query processing even in the case of fine-grained purpose labeling. We demonstrate the efficiency of our method with detailed experimental result in Section 2.5.3.

Although the relational model is perhaps today's most common data model, the use of many advanced data management systems, such as the ones based on XML and the ones based on the object-relational data model, has been increasing. In such models, objects are complex, have hierarchical structures, and are characterized by several semantic relationships. We thus develop a more sophisticated purpose management model for such complex data model in Section 2.6.

1.3.2 Data Privacy through Anonymization

In Chapter 3, we address the issue of data privacy by developing efficient data anonymization techniques. Anonymity is an important concept for privacy, and data anonymity is particularly crucial in public databases such as census data or health records collected by government agencies. Data anonymity can also be useful in the private sector, for example, when an organization wishes to allow third parties to access its customer data. Note that in such a case, it cannot be guaranteed that the privacy policy of the data will be always respected by the third parties. Thus, the organization must assure customers' privacy by removing all information that can link data items with individuals.

A key difficulty of data anonymization comes from the fact that data utility (i.e., data quality) and data privacy are conflicting goals. Intuitively, data privacy can be enhanced by hiding more data values, but it inevitably decreases data utility; on the other hand, revealing more data values increases data utility, but it may decrease data privacy. A recent approach addressing this difficulty relies on the notion of *k-anonymity* [29, 30]. In this approach, the data privacy is guaranteed by ensuring that any record in the released data is indistinguishable from at least $(k-1)$ other records with respect to a set of attributes called *quasi-identifier*. The k-anonymity problem has recently drawn considerable interest from research community, and a number of algorithms have been proposed [31–36].

Current solutions, however, suffer from high information loss mainly due to reliance on pre-defined generalization hierarchies [32–34, 36] or total order [31, 35] imposed on each attribute domain. In Section 3.2, we address such limitations by developing a data anonymization algorithm that utilizes clustering techniques. Intuitively, the k-anonymity requirement can be naturally transformed into a clustering problem where we want to find a set of clusters (i.e., equivalence classes), each of which contains at least k records. In order to maximize data quality, we also want the records in a cluster to be as similar to each other as possible. This ensures that less distortion