CERIAS Tech Report 2005-25

VISION PAPER: MICRO-VIEWS, OR ON HOW TO PROTECT PRIVACY WHILE ENHANCING DATA USABILITY

by Ji-Won Byun and Elisa Bertino

Center for Education and Research in
Information Assurance and Security,
Purdue University, West Lafayette, IN 47907-2086

Vison Paper: Micro-views, or on How to Protect Privacy while Enhancing Data Usability

Ji-Won Byun Elisa Bertino
CERIAS and Department of Computer Sciences
Purdue University
656 Oval Drive, West Lafayette, IN47906, U.S.A.
{byunj, bertino}@cs.purdue.edu

ABSTRACT

The large availability of repositories storing various types of information about individuals has raised serious privacy concerns over the last ten years. Yet database technology is far from providing adequate solutions to this problem that requires a delicate balance between individual's privacy and convenience and data usability by enterprises and organizations - a database which is rigid and over-protective may render data of little value. Though those goals may seem odd, we claim that the development of solutions able to reconcile them will be an important challenge to be addressed in the next few years. We believe that the next-wave of database technology will be represented by DBMS providing high-assurance privacy and security. In this paper, we elaborate on such challenges. In particular, we claim that we need to provide different views of data at a very fine level of granularity; conventional view technology is able to select only up to a single attribute value for a single tuple. We need to go beyond such level; we need a mechanism by which even a single value inside a tuple's attribute may have different views - which we refer to as micro-views. We believe that such a mechanism can be an important building block, together with other mechanisms and tools, of next-wave database technology.

1. INTRODUCTION

Current information technology enables many organizations to collect, store and use a vast amount of personal information in their databases. The use of innovative knowledge extraction techniques combined with advanced data integration and correlation techniques [7] makes it possible to automatically extract a large body of information from the available databases and from a large variety of information repositories available on the web. Such a wealth of information and extracted knowledge raises, however, serious concerns about the privacy of individuals. As privacy awareness increases, individuals are becoming more reluctant to carry out their businesses and transactions online, and many enterprises are losing a considerable amount of potential profits [11]. Also, enterprises that collect information about individuals are in effect under the obligations of keeping them private and must strictly control the use of such information in order to avoid potential law suits. Thus, information stored in the databases of an enterprise is no longer a valuable property of the enterprise, but a costly responsibility. Consequently, the development of data management techniques providing high-assurance privacy and at the same time avoiding unnecessary restrictions to data accesses is a crucial need. We need to combine the individuals' right to privacy with the need by many enterprises of carrying on information analysis and knowledge extraction. In many cases, such knowledge is used to provide better and tailored services to individuals.

To date issues related to privacy have been widely investigated and several privacy protecting techniques have been developed. The most well known effort is the W3C's Platform for Privacy Preference (P3P) [18]. P3P allows websites to encode their privacy policy in a machine readable format so that consumers can easily compare the published privacy policies against their privacy preferences. P3P, however, does not provide any functionality to enforce these promises to the internal privacy practice of enterprises. To complement P3P's lack of enforcement mechanisms, many privacy-aware access control models have also been investigated [3, 4, 8, 9]. Although all these models do protect privacy of data providers[1], 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., access is either allowed or denied as in most conventional access control models.

We believe that a new generation of privacy-aware access control models is required to be able to maximize information use by exploiting the nature of information privacy. First of all, information privacy is context-specific. For instance, consider address data of consumers. The comfort level of individuals toward the possibility of their address being used for marketing is significantly different from the case when the address information is to be used for shipping. Furthermore, the level of comfort varies from individual to individual. Some consumers may feel that it is acceptable to disclose their purchase history or browsing habits in return for better service; others may feel that disclosing such information may lead to an invasion to their privacy. These differences in individual privacy measures suggest that access control models should be able to cater to a large variation in privacy preferences and to maximize the use of information by taking these preferences into account.

Second, the use of data generalization[2] can significantly increase privacy comfort level of data providers. For example, suppose that an enterprise collects annual incomes of its consumers. This is indeed sensitive information, and many individuals may not be comfortable in letting this informa-

[1] By data providers, we refer to the subjects to whom the stored data is related.

[2] Data generalization refers to techniques that "replace a value with a less specific but semantically consistent value" [15].

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

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

tion to be used. Suppose now that the enterprise promises its consumers that the income information will be generalized before being used; e.g., $123,345 will be generalized to a categorical value $100-150K. This assurance will be surely more comforting to many consumers even though individual reactions may vary. Clearly, privacy enhancing access control models should be able to better utilize information by employing data generalization techniques.

The development of DBMS able to address the above requirements is a challenging task, requiring revisiting theoretical foundations of data models as well as languages and architectures. A core DBMS component which is crucial in such context is represented by the access control system. Current access control systems are fundamentally inadequate with respect to the above goals. For example, fine-grained access control to data, an important requirement for privacy, poses several difficult problems and to date no satisfactory solution exists. We have yet to understand which the relevant technical requirements are.

In this paper, we pose as a new challenge the development of a new generation of access control systems. As an example, we propose a radically new access control model able to exploit the subtle nature of information privacy to maximize the usability of private information for enterprises with privacy guarantees. Our model has not to be considered as a complete solution; rather it is meant to show some of capabilities that, in our opinion, a suitable model should provide. In particular, our model is based on the notion of *micro-view*. A micro-view applies the well known idea of views at the level of the atomic components of tuples, that is, to attribute's values. By using these different values, one is able to finely calibrate the amount of precision in the information released by queries.

The remainder of this paper is organized as follows. In Section 2, we present a high-level description of our access control model, and some technical challenges imposed by our model are discussed in Section 3. We provide a brief survey of related work in Section 4 and conclude our discussion in Section 5.

2. A SKETCH OF OUR "NAIVE" MODEL

Our model is based on what we could consider a typical "life-cycle" of data concerning individuals. During the data collection phase, data providers specify the level of privacy they require for their data and the type of possible usage of these data. Such user's preferences are then stored in the database along with data; access and use of the data is strictly controlled according to the user's preferences. In particular, in our model, answers to queries may have different precisions, depending on the privacy requirements for the data and the purpose of the query. The different precision levels are obtained by using different micro-views of the data. In this section, we first illustrate how data collection process is carried out in our model and what type of data preprocessing is required to support our model. Then we discuss our access control model more in details.

2.1 Data collection and preprocess

As previously mentioned, data providers specify the level of privacy they require for each type of data and each type of possible data usage when they release their personal information. Thus, enterprises must clearly define, based on the pre-established privacy policy, the levels of privacy, the types of data collected and the types of data usage (i.e., purposes) and make their consumers aware of such options. Table 1 describes these concepts and provides some examples.

Although any arbitrary number of privacy levels is possible, to ease the illustration we will limit the level of privacy to three levels in this paper: *Low*, *Medium* and *High*. We also consider only *name*, *address* and *income* as data types and *admin* and *marketing* as data usage types in our discussion for the same reason.

Data providers specify privacy requirements for their information by specifying a privacy level for each data type and each data usage. For instance, a consumer may select *Low* on *Address* for *Admin*, which means that he/she does not have any privacy concern over the address information when it is used for administrative purpose. Thus, the address information can be used for such purpose as it is. However, the same consumer 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 marketing purpose; thus, the address information should be used only in a sufficiently generalized form[4].

While specified privacy requirements are stored into database, the actual data items are preprocessed before being stored in the following way. Each data 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 into three levels: detailed address for *Low*, city and state for *Medium* and state for *High*.

Table 2 illustrates some fictional records and privacy requirements stored in a "conceptual" database relation. Notice that every data is stored in three different generalization levels, each of which corresponds to a privacy level. *PL_Admin* and *PL_Marketing* are metadata columns[5] storing the set of privacy levels of data for *Admin* and *Marketing*, respectively. For instance, {L, L, M} in *PL_marketing* indicates that the privacy levels of *Name* and *Address* are both *Low* while the privacy level of *Income* is *Medium*.

Note that exactly how such data is organized and stored in

[4] Note that if one wishes to allow data providers to completely opt out from any use of data, another privacy level (e.g., Opt-out) can be added to indicate that the particular data should not be used in any circumstance.

[5] The metadata columns may be viewable to any user, but they should be modifiable to only authorized users.

CustID[3]	Name		Address		Income		PL_Admin	PL_Marketing
1001	*L*	Alice Park	*L*	123 First St., Seattle, WA	*L*	45,000	{L, M, H}	{H, H, H}
	M	Alice P.	*M*	Seattle, WA	*M*	40K-60K		
	H	A.P.	*H*	WA	*H*	Under 100K		
1002	*L*	Aaron Parker	*L*	491 3rd St, Lafayette, IN	*L*	121,000	{L, L, M}	{H, M, H}
	M	Aaron P.	*M*	Lafayette, IN	*M*	120K-140K		
	H	A.P.	*H*	IN	*H*	Over 100K		
1003	*L*	Carol Jones	*L*	35 Oval Dr, Chicago, IL	*L*	64,000	{L, L, L}	{L, M, H}
	M	Carol J.	*M*	Chicago, IL	*M*	60K-80K		
	H	C.J.	*H*	IL	*H*	Under 100K		

Table 2: Private information and metadata

databases is a crucial issue as it determines the performance and storage efficiency of a system. However, this issue is beyond the scope of this paper and not discussed further here.

2.2 Access Control

In our model, users query the database using SQL statements. However, the data accessible to each query varies depending on the privacy levels of the data and the purpose of the query[6]. That is, each query runs as if it is running on a view that is defined by the purpose of the query and the privacy levels of data. We call such views as *privacy views*. Tables 3 and 4 illustrate this effect. For instance, any query against the base table in Table 2 with *Admin* purpose will return a result that is equivalent to the result of the query run on the privacy view in Table 3. As the privacy views directly reflect the information that is allowed by each data provider, querying against these views does not violate privacy.

Note that the major difference of our model from conventional database models is that in our model, different sets of data may be returned for the same query, depending on the privacy levels of data and the purpose of the query. For instance, suppose that the following query is written against the base table in Table 2: "SELECT * FROM Customer WHERE CustID = 1002". If the purpose of this query is *Admin*, then the system will return a tuple ⟨'Aaron Parker', '491 3rd St, Lafayette, IN', '120K-140K'⟩ as Aaron's privacy levels for *Admin* are specified as {L, L, M}. On the other hand, if the purpose of the query is *Marketing*, then a tuple ⟨'A. P.', 'Lafayette, IN', 'Over 100K'⟩ will be retrieved as his privacy levels for *Marketing* is {H, M, H}.

Another important issue to be addressed is how to associate a particular purpose with each query. In fact, it is almost impossible to correctly infer the purpose of a query as it means that the system must correctly figure out the real intention of database users. However, if we assume that users are trusted, then the problem of associating a purpose with each query becomes relatively easy; i.e., users themselves can specify the purpose of their queries with an additional clause[7]. For instance, a simple select statement "SELECT name FROM customer" can be extended to a form of "SELECT name FROM customer FOR marketing". In fact, this is not a flawed assumption at all. Many privacy violations may occur from accidentally accessing unauthorized information, and thus it is important to protect database users from committing such accidental violations.

CustID	Name	Address	Income
1001	Alice Park	Seattle	Under 100K
1002	Aaron Parker	491 3rd St, Lafayette, IN	120K-140K
1003	Carol Jones	35 Oval Dr, Chicago, IL	64,000

Table 3: Privacy-view for *Admin* purpose

CustID	Name	Address	Income
1001	A. P.	WA	Under 100K
1002	A. P.	Lafayette, IN	Over 100K
1003	Carol Jones	Chicago, IL	Under 100K

Table 4: Privacy-view for *Marketing* purpose

3. CHALLENGES

The full development of the approach we have sketched in the previous section and its integration in a DBMS architecture requires addressing several interesting challenges.

Policy specification language. The core of our model is that data providers can specify their privacy requirements using a privacy level for each data category. There is thus a strong need for a language in which privacy specifications can be precisely expressed. A challenge is that the language must be powerful enough to express every possible requirement, yet simple enough to avoid any ambiguity or conflict. Thus usability is a crucial issue. Especially as we cannot assume that every data provider would be an expert in privacy or any type of technology, GUI tools that are intuitive and instructive must be provided for them. We believe that many valuable lessons can be learned from existing technology related to P3P and APPEL [18, 17] and previous work on user interaction design [19]. It is important that data providers have a clear understanding of the guarantees provided by each privacy level.

Data generalization. Needless to say, devising a quality data generalization technique is one of the key challenges. There are two important issues to be considered here. The first issue is that the generalization process must preserve meaningful information from actual data as inadequate information would not be of any use. For example, although numeric or structured data may be relatively easy to be generalized into multi-levels that are meaningful, it is unclear how unstructured data (e.g., textual data) should be generalized into multi-levels. We need also to devise generalization policies and ontologies supporting systematic and consistent data generalization across the database. The other important issue is that generalization process must produce a sufficient level of data privacy by effectively suppressing distinctive information in individual data. For instance, consider name information of individuals. There are certain

[6] For now, assume that each query is associated with a specific purpose.

[7] A more sophisticated approach which validates whether users are indeed authorized to use their claimed purposes is thoroughly investigated in [5].

names that are more infrequent than others, and inadequate generalization techniques would not be able to hide the uniqueness of such names. Moreover, if the content of database dynamically changes, the task of distinct information hiding becomes much more challenging. Clearly, a big challenge in data generalization is to well balance the trade-off between information preservation and information suppression. Generalization must also be efficiently performed; in many cases, the system will have to perform data generalization "on the fly" while processing a query; in other cases, a privacy post-processing of queries will be required, because what has to be returned may depend on for example the cardinality and statistics of the results as well as on past accesses. Many valuable lessons can be learned from various generalization techniques that are available in statistical databases [1]. The main challenge here is that these techniques may have to be used dynamically in a variety of settings, ranging from data storage to query processing.

Metrics for data privacy and data quality and usability. So far, we have claimed that both privacy and usability of data can be achieved when data is sufficiently generalized. However, the key question is: how can we determine whether or not a certain generalization strategy provides a sufficient level of privacy and usability? As one can generalize data in various ways and degrees, we need metrics that methodologically measures privacy and usability of generalized data. It is clear that such metrics are necessary to devise generalization techniques that can satisfy the requirements of both data providers and data users.

Metadata storage. In our "naive model" we have assumed that collected data is generalized and stored into various levels at the preprocessing stage. This approach is simple and effective, yet may require huge storage space. For instance, suppose there are n numbers of privacy levels in a system. This means that the required storage space would be n times larger than the size of the collected data. Another approach is to postpone the generalization process to the time of data access. This method does not require any additional storage and also helps reducing unnecessary data generalization[8]. However, the overall performance may significantly suffer. Another possible solution is to use both pre-generalization and post-generalization selectively. For example, only data that are expected to be frequently accessed are pre-generalized and stored. Then other data that are not pre-processed should be generalized when they are actually accessed. Also, for better performance the post-generalized data may be cached in a special space. Using this approach, one can try to reduce the overall cost of generalization process. However, a challenge here is to balance the trade-off between storage and performance. Yet another approach could be based on the use of views, which would have to be extended with innovative capabilities for value generation and transformation.

Complex query processing. In this paper we have considered only simple queries; i.e., queries without join, subquery or aggregation. The key question here is whether complex queries can be introduced in our model. Even though it seems that they can be correctly processed in the model, it is not clear whether the results of such queries would be still meaningful.

Applicability to general-purpose access control. Although we have limited our discussion to access control for privacy protection, we believe it is possible to extend our model to a general-purpose access control model. For instance, each user can be assigned to a trust level[9], and the access control system can control, based on user's trust level, degrees of precision on accessible information. This approach is indeed very similar to multilevel secure database systems [14, 12, 6], where every piece of information is classified into a security level and every user is assigned a security clearance. However, the main difference is that our approach can provide much finer level of control as 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". This type of finer grained access control can be extremely useful for internal access control within an organization as well as information sharing between organizations. Even though such extension seems very promising at this point, further investigation is required to confirm this possibility.

Other issues. There are many other issues that require careful investigation, such as problems of polyinstantiation [13, 10], inference and integrity. Addressing such issues is also crucial for the development of comprehensive access control models for high-assurance privacy.

4. RELATED WORK

To date, several approaches have been reported dealing with various aspects of the problem of high-assurance privacy systems. Here we briefly discuss the approaches that have provided some initial solutions that can certainly be generalized and integrated into comprehensive solutions to such problem.

The W3C's Platform for Privacy Preference (P3P) [18] 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 consumer's set of privacy preferences which are specified in a privacy preference language such as A P3P Preference Exchange Language (APPEL) [17], also designed by the W3C.

The concept of Hippocratic databases, incorporating 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. A privacy policy defines for each attribute of a table the usage purpose(s), the external-recipients and retention period, while a privacy authorization defines which purposes each user is authorized to use.

Byun et al. presented a comprehensive approach for pri-

[8] Note that data items are not equally accessed. That is, some data items are accessed much more frequently than the others. It is also reasonable to assume that some data items are rarely accessed.

[9] The trust level will not be chosen by users, but assigned to users by authorized personnel.

vacy preserving access control based on the notion of purpose [4, 5]. 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 fully discussed in details in [4], and a systematic approach to implement the notion of access purposes, using roles and role-attributes is presented in [5].

Previous work on multilevel secure relational databases [6, 12, 14] 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. These constraints ensure that there is no information flow from a lower security level to a higher security level and that subjects with different clearances see different versions of multilevel relations.

In order to prevent re-identification of anonymized data, Sweeney introduced the notion of k-anonymity [16]. K-anonymity requires that information about each individual in a data release be indistinguishable from at least k-1 other individuals with respect to a particular set of attributes. Sweeney also proposed a technique using generalization and suppression of data to achieve k-anonymity with minimal distortion [15].

5. CONCLUSIONS

In this paper, we discussed a new approach to access control that maximizes the usability of private information for enterprises while, at the same time, assuring privacy. We believe that one direction for next-generation DBMS technology is represented by DBMS with high-assurance security and privacy. The "naive" model we presented in the paper provides an example of an access control system for such a new DBMS. Based on this model, we discussed many challenges that need to be addressed. We would like to conclude the paper by saying that ultimately a suitable solution to access control systems with high-privacy assurance will be built by integrating techniques such as view mechanisms, statistical databases, anonymization, privacy-preserving computation and data mining. The main challenge is how to integrate such techniques in a full-fledged DBMS ensuring good performance.

6. REFERENCES

[1] Nabil Adam and John Wortmann. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys (CSUR)*, 21, 1989.

[2] Rakesh Agrawal, Jerry Kiernan, Ramakrishman Srikant, and Yirong Xu. Hippocratic databases. In *The 28th International Conference on Very Large Databases (VLDB)*, 2002.

[3] Paul Ashley, Calvin S. Powers, and Matthias Schunter. Privacy promises, access control, and privacy management. In *Third International Symposium on Electronic Commerce*, 2002.

[4] Jiwon Byun, Elisa Bertino, and Ninghui Li. Purpose based access control for privacy protection in relational database systems. Technical Report 2004-52, Purdue University, 2004.

[5] Jiwon Byun, Elisa Bertino, and Ninghui Li. Purpose based access control of complex data for privacy protection. In *Symposium on Access Control Model And Technologies (SACMAT)*, 2005. To appear.

[6] Dorothy Denning, Teresa Lunt, Roger Schell, William Shockley, and Mark Heckman. The seaview security model. In *The IEEE Symposium on Research in Security and Privacy*, 1998.

[7] Xin Dong, Alon Halevy, Jayant Madhavan, and Ema Nemes. Reference reconciliation in complex information spaces. In *ACM International Conference on Management of Data (SIGMOD)*, 2005.

[8] IBM. *The Enterprise Privacy Authorization Language (EPAL)*. Available at www.zurich.ibm.com/security/enterprise-privacy/epal.

[9] Kristen LeFevre, Rakesh Agrawal, Vuk Ercegovac, Raghu Ramakrishnan, Yirong Xu, and David DeWitt. Disclosure in hippocratic databases. In *The 30th International Conference on Very Large Databases (VLDB)*, August 2004.

[10] Fausto Rabitti, Elisa Bertino, Won Kim, and Darrell Woelk. A model of authorization for next-generation database systems. In *ACM Transactions on Database Systems (TODS)*, March 1991.

[11] Forrester Research. Privacy concerns cost e-commerce $15 billions. Technical report, September 2001. Available at www.forrester.com.

[12] Ravi Sandhu and Fang Chen. The multilevel relational data model. In *ACM Transaction on Information and System Security*, 1998.

[13] Ravi Sandhu and Sushil Jajodia. Polyinstantiation integrity in multilevel relations. In *IEEE Symposium on Security and Privacy*, 1990.

[14] Ravi Sandhu and Sushil Jajodia. Toward a multilevel secure relational data model. In *ACM International Conference on Management of Data (SIGMOD)*, 1991.

[15] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. In *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.

[16] Latanya Sweeney. K-anonymity: A model for protecting privacy. In *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.

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

[18] World Wide Web Consortium (W3C). *Platform for Privacy Preferences (P3P)*. Available at www.w3.org/P3P.

[19] Kaping Yee. User interaction design for secure systems. In *The 4th International Conference on Information and Communications Security*, 2002.