World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:6, 2008

Hippocratic Database : A Privacy-Aware Database

Norjihan Abdul Ghani, Zailani Mohd Sidek

***Abstract*—**Nowadays, organizations and business has several motivating factors to protect an individual's privacy. *Confidentiality* refers to type of sharing information to third parties. This is always referring to private information, especially for personal information that usually needs to keep as a private. Because of the important of privacy concerns today, we need to design a database system that suits with privacy. Agrawal et. al. has introduced Hippocratic Database also we refer here as a privacy-aware database. This paper will explain how HD can be a future trend for web-based application to enhance their privacy level of trustworthiness among internet users.

***Keywords*—** Hippocratic Database, privacy, privacy-aware.

I. Introduction

WEB-BASED applications owes its dominance over traditional paper-based systems to the ability of information manipulation – to store endlessly, to sort efficiently, to locate effortlessly, and to make decision effectively [1]. More and more data have been exchanged in order to complete a task in a web-based application or online information systems. By using web-based application, people can shop online; no traffic jam, more convenient and of course by only fingertips. Web-based application has increased the quality of services provided by organizations.

However, behind all this advantages of web-based application, the risks of privacy violation are increasing. Not only to keep the databases secure, but at the same time we need to consider the database privacy. It's because the more data disclosure, the more protection should be applied.

Easy access to private information will cause the misuse of data, no control over the information and others. Because of this, it's important to protect the information not only from external threats but also from insider threats. Data disclosure when performing a task in web-based application should be ensured by data security mechanisms. A complete solution to data security must meet three requirements [2] : 1) *secrecy* or *confidentiality* refers to protection of data against unauthorized disclosure, 2) *integrity* refers to the prevention of unauthorized and improper modifications, and 3) *availability* refers to the prevention and recovery from hardware and software errors from malicious denials that make the database system unavailable. In [3], confidentiality involves sharing of information while secrecy is a type of blocking that makes the information unavailable Ensuring the secrecy means protection of data involved in highly protected environment such as military environment. Confidentiality are always refers to type of sharing of private information to third parties.

Because of this reason, privacy-protection is a growing challenge for database security and privacy experts. Privacy Protection is a process of finding appropriate balances between privacy and multiple competing interests [8]. We suggest that in today's era, database system should be secure and private in order to get consumer's trust towards web-based application. The new concept of database system with privacy has been introduced in [7]. Hippocratic database (HD) has been introduced in responding to significant privacy threats that always caused by inference and multilevel security problems. Hippocratic database should be architected to regulate use and disclosure of private information in strict accordance with privacy & security laws, enterprise policies & individual choices. This means that it'll find an agreement between consumer and the organization itself towards privacy. Hippocratic database will protect individual privacy without impeding legitimate and beneficial uses of information. This paper will explain what privacy is and the privacy problems in the current database system and the introduction of HD as a privacy-aware database system.

The structure of the rest of this paper is as follows. Section 2 will give an overview of privacy as a base for confidentiality Section 3 will discuss the main privacy problems in current database system. We define what is Hippocratic Database in Section 4 and how Hippocratic database can be viewed as privacy-aware database system in section 5. Section 6 will conclude the paper discussion.

II. Privacy as a Base for Confidentiality

"The right to be let alone"

Louis Brandeis, 1980
(Harvard Law Review)

Privacy, generally is a central to our dignity and our basic human rights. Privacy is the right of individuals to determine

for themselves when, how, and to what extent information about themselves is communicated to others [7]. It's the ability to control collection, retention and distribution of themselves [9].

It's the right of for them to determine for what purposes their private information is stored and used. Although it seems that security and privacy are one and same, but it's differ. Privacy is concerned with confidentiality of the information. When user thinks that there's a need to keep any information from anybody, means that they want to keep the information as their privacy. Now, privacy become a major concerns when individual interacting with corporation. Organizations and business needs to protect an individual's privacy. The main reason is, to seek the user's trust towards services offer by them. They want users know that they protect their private information in a trustworthy ways. And from the user's perspective, they are realizing that it's is become a compulsory for them to make sure their private information keep as private.

A number of common privacy dimensions have been defined that have gained wide acceptance [8]. They are as follows :

1. *privacy of the person*, sometimes referred to as 'bodily privacy'. This is concerned with the integrity of the individual's body. Issues include compulsory immunisation, blood transfusion without consent, compulsory provision of samples of body fluids and body tissue, and compulsory sterilisation;

2. *privacy of personal behaviour*. This relates to all aspects of behaviour, but especially to sensitive matters, such as sexual preferences and habits, political activities and religious practices, both in private and in public places. It includes what is sometimes referred to as 'media privacy';

3. *privacy of personal communications*. Individuals claim an interest in being able to communicate among them, using various media, without routine monitoring of their communications by other persons or organizations. This includes what is sometimes referred to as 'interception privacy'; and

4. *privacy of personal data*. Individuals claim that data about themselves should not be automatically available to other individuals and organizations, and that, even where data is possessed by another party, the individual must be able to exercise a substantial degree of control over that data and its use. This is sometimes referred to as 'data privacy' and 'information privacy'

In this paper, we apply the term of private information as a notion of confidentiality. As explain above, confidentiality involves sharing of information with the expectation that it will not be revealed to third parties, or it will be revealed under restricted circumstances [10]. Confidentiality is about controlling the access to information and its release according to certain agreement, normally from owner, organizations that own the information and third parties that get the accesses.

Privacy ensures and protects our information from unneeded disclosure. Most people only use web-based application if they feel that their information is secured and private enough. Web-based applications require techniques in privacy-preserving data management. There are three techniques in privacy-preserving data management [2] [8]. First technique in privacy-preserving deals with privacy when released to third party, (also known as data anonymization), second, privacy-preserving in data mining context and third is database tailored to support privacy policy, such as the policies that can be expressed by using the well-known P3P. P3P is an international privacy policy was developed by W3C.

The main challenge in privacy is to share the information while complying the data owner privacy preferences.

III. PRIVACY PROBLEMS IN CURRENT DATABASE SYSTEM

Privacy violations via inferences first considered in statistical databases [4]. The security requirement in statistical databases is to provide access to statistics about data inside the database while protecting the confidentiality of the data. Statistical databases are used to produce statistics on various populations. Individual information such as identity card numbers, credit card information, medical reports are being considered as confidential in any types of database. Users may allow to access statistical information on the population. For example, in human resource database, specifically in employee table, employee's income should be considered as secure and private, which is only can be accessed by an authorized users. But, because of inference problems, anybody can access use the statistical query operation to know the exact income for anyone. There are two types of inference problems; *1) direct attack* refers to finding sensitive information directly with queries that yield only a few records and 2) indirect attacks refers to seeking to infer the final result based on a number of intermediate statistical results (SUM, AVG, etc). Fig. 1 shows the indirect information access via inference channels.

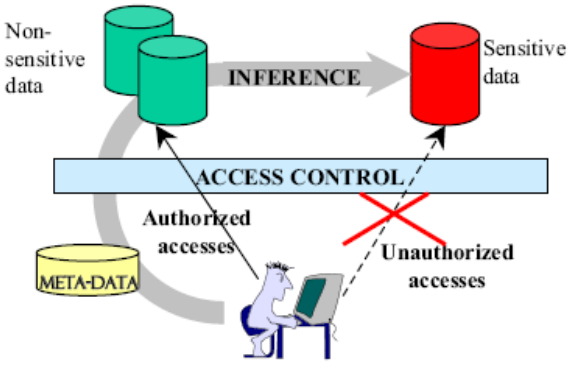

Fig. 1 Indirect information access via inference channel

Let use the example below to explain the inference problems in statistical database. Assume that we have a relation as shown below :

employee (name, identity_num, income, degree, sex, city)

From the relation above, we can see that identity_num (referring to identity number) and income is the main data that is considered private. Suppose that users are allowed to retrieve only the statistical information over this relation by using **SUM**, **AVG**, **MIN**, **MAX**, **COUNT**, and others. Means that users are only allowed to obtain the average income rather than specific employee's income. Then, we submitted a query to find the average income of female staff who have PhD and live in Petaling Jaya.

```
Q1 : select AVG(income)
     from employee
     where degree= "Ph.D"
     and sex = "F"
     and city = "Petaling Jaya"

Q2 : select COUNT(*)
     from employee
     where degree= "Ph.D"
     and sex = "F"
     and city = "Petaling Jaya"
```

If we know that A has PhD and lives in Petaling Jaya and then we want to know his income, we may use the above two queries. When query **Q1** returns one, then the result of query **Q2** is the income of A.
Let's us submit another two queries;

```
Q3 : select SUM(income)
     from employee
     where degree= "Ph.D"
     and sex = "F"
     and city = "Petaling Jaya"
     and name <>"A"

Q4 : select AVG(income)
     from employee
     where degree= "Ph.D"
     and sex = "F"
     and city = "Petaling Jaya"
```

Query **Q3** will return the sum of all incomes except Ahmad, while the query **Q4** will returns the sum of all incomes. So, the differences between these two incomes should be Ahmad's income.

The above query example show how inferences problems occurred in statistical databases. From previous example, we can see that although we don't have the authorization to access certain information, actually we still can access. So, data is not private anymore. Several research have been done in statistical database, able to provide statistical information (**SUM**, **AVG**, **MIN**, **MAX**, **COUNT**) without compromising sensitive information about individuals [5] [6].

A second concern with database security is multilevel security problems. It allows information with different classifications to be available in an information system. At the same time, it's also allows multiple level of security (top secret, secret, confidential, unclassified) to be defined and associated with individual attribute values. Security level of a query should be higher or lower than that individual data item. On the other hand, higher security query can't write a lower security data items.

Because of these two major problems in database security, the concept of HD has been introduced as in [7]. Hippocratic database looks like to overcome this two major concerns in database; inference problems and multilevel security. Hippocratic Database share with statistical database the goal of preventing disclosure of private information and there are a few techniques have been developed for statistical database will find application in HD. For multilevel security, the basic architecture idea of HD was actually inspired from secure database.

IV. HIPPOCRATIC DATABASE

A. Ten Hippocratic Database Principles

Ten guiding principles of Hippocratic databases and initial designs to provide limited disclosure and compliance audition were introduced in [7] :

1. **Purpose Specification** For personal information stored in the database, the purposes for which the information has been collected shall be associated with that information.
2. **Consent** The purposes associated with personal information shall have consent of the donor of the personal information.
3. **Limited Collection** The personal information collected shall be limited to the minimum necessary for accomplishing the specified purposes.
4. **Limited Use** The database shall run only those queries that are consistent with the purposes for which the information has been collected.
5. **Limited Disclosure** The personal information stored in the database shall not be communicated outside the database for purposes other than those for which there is consent from the donor of the information.
6. **Limited Retention** Personal information shall be retained only as long as necessary for the fulfillment of the purposes for which it has been collected.
7. **Accuracy** Personal information stored in the database shall be accurate and up-to-date.

8. **Safety** Personal information shall be protected by security safeguards against theft and other misappropriations.
9. **Openness** A donor shall be able to access all information about the donor stored in the database.
10. **Compliance** A donor shall be able to verify compliance with the above principles. Similarly, the database shall be able to address a challenge concerning compliance.

B. Architecture

Main consideration of HD is to enforce the privacy policy in database system. It has been introduced by Agrawal et al. to incorporate privacy protection in relational database system. An important feature of HD is the use of metadata, consisting of *privacy policies* and *privacy authorizations* stored in privacy-policies tables and privacy-authorizations table respectively.

In this architecture, *purpose* is used as the central concepts. Purpose is a major role in access control [11]. The HD performs privacy checking during the query processing. Every query submitted to the database with the intended purpose. Then, the system will checks either the user who issued the query is authorized to access the information or not by checking either he/she present in the list of authorized user for that purposes in the privacy-authorizations table. Then, if yes, the system will ensures that the query accessed only the attributes that are explicitly listed for query purposes in the privacy-authorizations tables. If yes, the system will ensure that only records whose purposes attribute includes the query purpose are visible to the query. The conclusion is, only attributes for certain purposes are allowed to be accessed by authorized users. Fig. 2 shows the Strawman architecture of a Hippocratic database system.

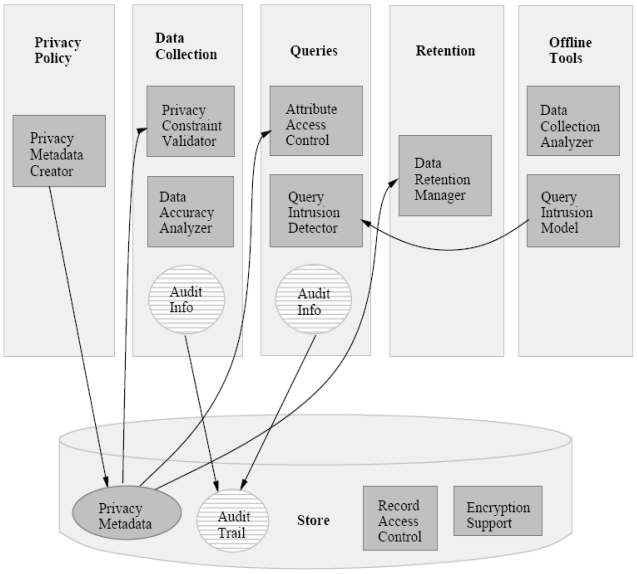

Fig. 2 An Architecture of Hippocratic Database [7]

V. Privacy Protection in Hippocratic Database

As explain in the previous section, the main concept of HD is purpose. Users are allowed to access the information that is available for them as authorized users. During the HD design, we'll collect for what purposes the information was collected and to whom the information can be disclosure.

Hippocratic Database as discussed in previous section is also known as privacy-aware database. This purpose concept has been introduced to implement the privacy-aware access in database system. It has been designed to maximize the privacy protection factor in database system. Whenever the system will limit the access only for intended purposes means that users who not have the authorization access to that purposes can't access the information in a database. The access to database in only permitted based on purposes. This will consider that purposes can limit the access and at the same time, privacy can be obtained.

Privacy in database can't easily achieve using traditional access control models [11]. There a two types of access control that have been considered as traditional access control; Mandatory Access Control (MAC), and Discretionary Access Control (AC) [12]. We're not going to discuss this two access control models since there are lots of research have been done in this. But now, as privacy becomes a major concern for consumers and enterprises, many privacy protection access control models have been introduced. In a world today, traditional access control in not enough to implement itself in web-base application. Hippocratic database is not an access control technique but it's more on database system which the main concept is to protect the privacy. In HD, in order to preserve the privacy of information providers, every information access must comply with the privacy policies on which information providers have agreed. A typical privacy policy for a HD includes purpose(s), retention and authorized users. It states that the particular information can be accessed only for the specific purpose(s) on the specific condition. The retention indicates how long the information can be retained. In [11], purpose is defined as to describe the reason(s) for data collection and data access. Hippocratic Database controls the access to a database by using the purpose concept.

VI. Conclusion

As the world are moving forward to a new era of web-based application, it's important to highlight the privacy. Hippocratic Database exists when the world really needs something for privacy. Hippocratic Database, based on purpose concept, identify who can access our information, which information and for what purposes. On the way, it's also depends on owner preferences. This paper discuss on how Hippocratic Database solve the two major concerns in database privacy, and also discussing on how accessing the information through Hippocratic Database can be considered as private information.

REFERENCES

[1] J. H. Moor. "Towards a Theory of Privacy for the Information Age". *Computers and Society*, 27(3):27-32, 1997.
[2] Elisa Bertino, and Ravi Sandhu, "Database Security—Concepts, Approaches, and Challenges", *IEEE Transactions on Dependable and Secure Computing*, Vol. 2, No. 1, January-March 2005, pp 2 – 19.
[3] Sabah S. Al-Fedaghi, "Privacy as a base for Confidentiality". Presented in the Fourth Workshop on the Economics of Information Security, Harvard University, Cambridge, MA, 2005.
[4] Csilla Farkas, Sushil Jajodia. "The Inference Problem : A Survey". *SIGKDD Explorations*, Volume 4, issues 2, pp 6-11.
[5] N. R. Adam & J. C. Wortman. "Security-control methods for Statistical Databases". *ACM Computing Surveys*, 21(4):515 – 556, Dec 1989.
[6] A. Shoshani. "Statistical Databases : Characterictics, Problems and Some Solutions". In Proc. of the Eighth International Conference on Very large Databases, pages 208 -213, Mexico, September 1982.
[7] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. Hippocratic databases. In *The 28th International Conference on Very Large Databases (VLDB)*, 2002.
[8] Clarke, R. 1999. Introduction to Dataveillance and Information Privacy, and Definitions and Terms.[Online] Available : http://www.anu.edu.au/people/Roger.Clarke/DV/Intro.html#Priv
[9] Goldberg, I., Wagner, D., Brewer, E. "Privacy-Enhancing Technologies for the Internet". Proceedings of IEEE COMPCON '97, 1997, 103 – 109.
[10] Marx, G. T., 2001. "Identity and Anonimity: Some Conceptual Distinctions and Issues for Research", *In J. Caplan and J. Torpey, Documenting Individual Identity* (Princeton University Press, 2001)
[11] Ji-Won Byun, Ninghui, "Purpose Based Access Control for Privacy protection in relational Database Systems". *The VLDB Journal,* 2006.
[12] (Book Chapter) Sabrina De Capitani di Vimercati, Sarah Foresti, Pierangela Samarati, "Authorization and Access Control". *Privacy & Trust in Modern Data Management*.
[13] Silcana Castano, Mariagrazia Fugini, Giancarlo Martella, Peirangela Samaranti, "Database Security", Addison Wesley1994.

Norjihan Abdul Ghani is with Department of Information Science, Faculty of Computer Science and Information Technology, University of Malaya 50603 Kuala Lumpur, Malaysia norjihan@um.edu.my

Zailani Mohd Sidek is with the Centre for Advanced Software Engineering (CASE), Universiti Teknologi Malaysia, City Campus,Jalan Semarak, 54100 Kuala Lumpur, zailani@citycampus.utm.my