Proceedings of the 4th WSEAS Int. Conf. on Information Security, Communications and Computers, Tenerife, Spain, December 16-18, 2005 (pp180-185)

Context-dependent Security Enforcement of Statistical Databases

VIVEK MISHRA, ANDREW STRANIERI, MIRKA MILLER, JOE RYAN
School of Information Technology and Mathematical Sciences
University of Ballarat
PO Box 663,Ballarat, Victoria 3353
AUSTRALIA

Abstract: - A statistical database system is a database system that contains information about individuals (companies, organisations) which enable its authorized users to retrieve aggregate statistics such as sample total, mean and count. The security problem for a statistical database is to limit the use of the database so that only statistical information is available to an authorized statistical user and no sequence of queries is sufficient to infer protected information about any individual. If, however, as a result of a statistical query, individual confidential information is obtained by a statistical user, then the database is said to be compromised. In order to prevent compromise, we use a knowledge based system approach. A statistical user can pose only statistical queries to the database. The knowledge based system will infer a query result by answering questions from the statistical user in such a way that the individual information is kept confidential. Earlier models modelled SDB compromise using two kinds of knowledge [7]. In this paper we enhance this model by including another important knowledge, namely, legal knowledge, and we describe an implementation of a knowledge base that supports the decision making for the protection of privacy in statistical databases, while taking into account not only the working and (known) supplementary knowledge but also legal knowledge.

Key-Words: - Statistical database, Knowledge based system, Privacy, Supplementary knowledge, Working knowledge, Legal knowledge.

1 Introduction

A statistical database (SDB) is a database that is used for statistical queries (aggregates, averages, counts) on subsets of the database entities. The security problem for statistical databases is to limit the use of a statistical database so that, while statistical information is available, no sequence of queries is sufficient to infer protected information about any individual. When such information is inferred, the SDB is said to be compromised.

For example, using the sample data in Table 1, an authorized user may issue a query such as: the number of females in Smalltown and receive the result "1". This is illustrated in Figure 1. Smalltown has a very small population of just one individual who may be identified so that answering a statistical query about Smalltown will divulge information about just one individual, contrary to the purpose of a statistical database which is to provide statistical information about groups or subpopulations of the database, not individuals.

To date, some measure of security for statistical databases has been achieved with various noise addition and query restriction techniques. Noise addition techniques involve either input perturbation or output perturbation or data swapping. Such techniques have been applied by [1], [2] and [3]. Query restriction techniques can be based on query set size, query set overlap, partitioning, cell suppression, auditing and have been applied by [4] and [5].

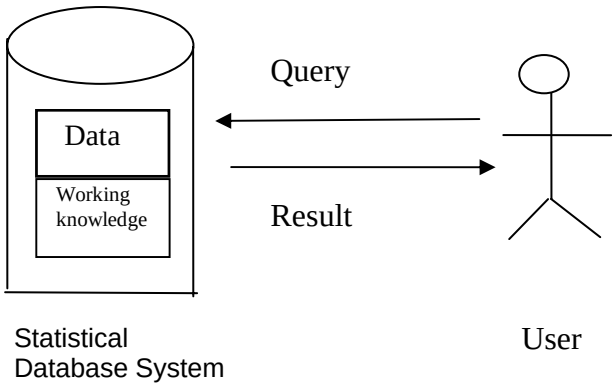

Figure 1 User database interaction

One drawback of the restriction techniques is that they are fixed techniques and will not release any confidential information, whenever a query is deemed unanswerable due to potential compromise, regardless of any other considerations.

In some jurisdictions, the security of statistical database is important but at the same time revealing the answer may be more vital for the "public good" than would be the continued provision of confidentiality of individuals. Figure 2 illustrates an authorized user issuing the same query as in Figure 1. Instead of simply answering the query if it is safe and not answering it if it could result in a compromise, this time a knowledge based system (KBS) intervenes and prompts the user to see if a public health issue is at stake. If it is then the system may

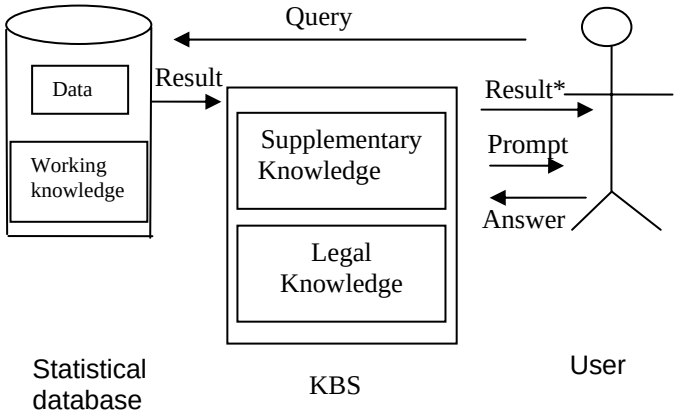

Figure 2 KBS enhances privacy

release the confidential information to an authorized user. Since the Result from the SDB is not necessarily passed on to the user, we denote the response obtained by the user from the SDB via the KBS as Result*. It may be the same as the Result but for legal reasons Result* may be different from Result. Such a situation is not uncommon and occurs for example in Australia, where the privacy laws

Name	City	Sex	Status	Age	Child	Salary
White	Smalltown	F	S	41	1	48700
White	Brisbane	M	M	19	4	16400
Mlynar	Adelaide	F	W	61	7	30000
Brown	Sydney	F	M	36	2	30000
Baker	Sydney	F	M	25	0	40000
Baker	Brisbane	F	D	25	1	17000
Peter	Melbourne	M	S	29	2	40900
Brown	Sydney	M	D	37	3	25000
Black	Melbourne	M	D	50	0	17000
Ling	Hobart	M	M	60	0	15000
Ling	Perth	M	S	42	2	16500
White	Brisbane	F	S	50	3	25500
Brown	Canberra	F	S	25	1	25000
Brown	Perth	F	W	44	2	22400

Table 1 Sample database

allow an individual's privacy to be breached in the event of a public health benefit according to section 95A of the Privacy Act 1988.

Securing a database using a knowledge based system enables some context-dependency in dealing with a database for those authorized users who need private information for public safety.

In the next section we discuss sources of knowledge for securing a SDB, namely, working knowledge, supplementary knowledge and legal knowledge. Section 3 briefly discusses knowledge based systems. Section 4 completes the paper with conclusion and directions for further research.

2 Sources of Knowledge

According to [7], two sources of knowledge useful for perpetrating a compromise of a SDB have been identified as working knowledge and supplementary knowledge. Working knowledge includes knowledge about the attributes and tables in a database.

For example, an authorized user needs to know that the attribute labeled "sal" represents an individual's annual salary. This is working knowledge. Typically, some attributes in a SDB represent publicly known, non-confidential data and others represent confidential data. Knowledge of the confidential status of each attribute is also working knowledge.

Supplementary knowledge is not working knowledge but is background or context knowledge about the external world that could be used to infer data that represents a compromise.

For example, an authorized user may personally know most of the inhabitants of Smalltown. That supplementary knowledge could be applied to the result of a query to enable the user to infer the identity of the resident over forty that had an income of $750,000 last year.

Knowledge about the legislative context of the SDB is also a kind of supplementary knowledge. In Australia, the Privacy Act 1988 protects an individual's privacy. However, an individual's right to privacy is curtailed if the data may be beneficial to a police investigation. The supplementary knowledge brought to bear on the Smalltown individual will be overridden by legal knowledge that the confidentiality is to be breached for a criminal investigation.

In this paper, to our best knowledge for the first time in the context of security of statistical databases, we consider including this new type of knowledge, the

legal knowledge. Such knowledge is not part of the working knowledge nor is it in the same category as supplementary knowledge, but rather, is orthogonal to the other two types of knowledge that can pertain to the problem of security of statistical databases. Interpolating from our knowledge of real-life scenarios, we can assume that the three types of knowledge are independent of each other.

We will next consider in turn each of the three types of knowledge in more detail. We shall assume throughout the rest of this paper that Salary is the only confidential attribute in the database.

2.1 Working Knowledge

Working knowledge is the user's knowledge of the attributes and their values in the database. The user should have knowledge of the non-confidential attributes in order to avoid meaningless queries. He or she needs to know the form in which the attributes are represented in the database.

For example, in Table 2, there is an attribute which specifies an individual belongs to a particular location and this attribute is called "City" in the database.

The user needs to know the legal values of attributes and also whether they are recorded using lowercase or uppercase letters or in numeric or abbreviated form.

All this knowledge is necessary for the efficient use of a statistical database. Otherwise, precious time would be spent in trying to guess the correct form in which the name of attribute is placed in the database, for example, is the value of attribute used to hold values for Sex M or Male or male or 1?

Attribute A_j	Values	$\|A_j\|$
City	Smalltown, Brisbane, Adelaide, Sydney, Melbourne, Hobart, Perth, Canberra	8
Sex	M, F	2
Status	S, M, D, W	4
Child	0,1,2,3,4,5,6,7	8

Table 2 Attribute values for database in Table 1

2.2 Supplementary Knowledge

Supplementary knowledge is the pre-knowledge of the user about the database which is not directly derived from the database but from external sources. The three basic types of supplementary knowledge used in dealing with database compromise are classified as supplementary knowledge of type 1, type II and type III [7].

A user has supplementary knowledge of type I, if he has knowledge to build a characteristic formula which uniquely identifies the individual or a group of individuals in a database.

For example, a malicious Smalltown resident who wishes to confirm that none of the single women in Smalltown have children may initially issue a query to discover the average number of children to single mothers:

AVG(number of children)
WHERE (City = Smalltown AND Sex = F AND status=S)
and receive the result => 0.25.

Supplementary knowledge leads the resident to infer that at least one single mother exists in Smalltown. A count query confirms that there are four single women in Smalltown. Supplementary knowledge leads the resident to infer that the only way the average number of children could be 0.25 is if there were 3 women with no children and one with a child. The resident recalls the name of four single women likely to be in the SDB but does not know which one has the child. A query that lists the average age of the Smalltown children with single mothers yields a result of => 41. The resident again draws on supplementary knowledge of the age of the four single mothers and infers that the one with a child must be Ms White.

A user has supplementary knowledge of type II, if a user has knowledge about the confidential X-value of particular individual in a database.

For example, if the user knows the Salary of Ms. White living in Smalltown is $48700 then the user has supplementary knowledge of type II. Such knowledge can be obtained directly from other sources external to the SDB or as a data inference, for example, by issuing two queries: a query which counts the Salary of all the persons in the database, and a query that gives the sum of Salaries of all persons aged over 20,

SUM(Salary of all)
and
SUM(Salary of all individuals aged over 20)

Since Ms. White of Brisbane is the only individual included in the first query and not in the second one, by taking the difference of the answers to the two queries, we can infer that Ms White's Salary is $48700. That is, if the (exact) results of both the queries are released then the statistical database will be compromised.

The supplementary knowledge of type III is the knowledge of the user other than type I and type II. An example of supplementary knowledge of type III is a functional dependency. For example, consider a database containing attributes called Position and Level. It is possible that there is a relationship between the values for attributes Position, Level and Salary, for example, the combination of Position and Level may functionally determine the Salary. If the user knows somebody's Position is "Programmer" and the Level is "level6", then it leads to an exact compromise even if we keep the attribute Salary itself confidential. The problem occurs when Salary is a confidential attribute while Position and Level are not and when there is the functional dependency FD{Position, Level} -> Salary (such situation is not uncommon for example, in government organizations).

2.3 Legal Knowledge

The decision tree in Figure 4 represents privacy guidelines [10]. According to Australian privacy law, if some confidential information about individual is needed to be used for social welfare than it can be shown to the user.

As shown in the decision tree, the record is shown to the user if the information about an individual is used for preventing serious threats to an individual or the public or for investigating a serious crime.

If the information is used for research relevant to public health or safety and if it is feasible to obtain the individual's consent then the knowledge based system will advise that consent from compromised individuals should be obtained from an Ethics committee. If the committee agrees to show the record and it is used to prevent crime or is necessary for enforcement of laws relating to the confiscation of the proceeds of crime or necessary for protecting public revenue or necessary for preventing improper conduct than the confidential record can be shown to user.

For any other reason the information would remain confidential.

3 Knowledge Base for Secure SDBs

A knowledge-based system is a computer system that models real world knowledge. Typically, knowledge is heuristic in nature based on rules of thumb rather than absolute certainties.

Generally, an expert system that requires a knowledge base contains a database of extracted human knowledge. Building this expert system knowledge base is a difficult task as the representation of knowledge must be precise and accurate. In our approach presented in this paper, there are three sources of knowledge, namely, working knowledge, supplementary knowledge and legal knowledge.

The significance of knowledge based system in securing statistical database is that it considers real world knowledge about the database. The user will query the database and the knowledge based system will find out the user's purpose to use the information.

The knowledge based system does so by asking several questions to determine if the compromise should occur, if a release of the confidential information about an individual is used for public or any individual safety or for public good then it will release the information. In the case where KBS finds that the information is confidential and it may be used in order to harm any individual or to threaten any individual's life then the query result will not be shown to the user.

4 Conclusion and Further Research

The security of statistical databases using query restriction and perturbation techniques is rigid and aims to prevent any compromise of protected information about individuals.

In some cases, where security is an important issue but at the same time confidential information needs to be used for social good then, according to Australian Privacy Law, it must be released to the authorized user. This can be done using a knowledge based system which after receiving the response(s) from the user for corresponding questions, decides whether to release the confidential information according to privacy law requirements.

Currently, we are implementing the knowledge base that encodes privacy law and supplementary knowledge using a decision and argument tree representation described in [12].

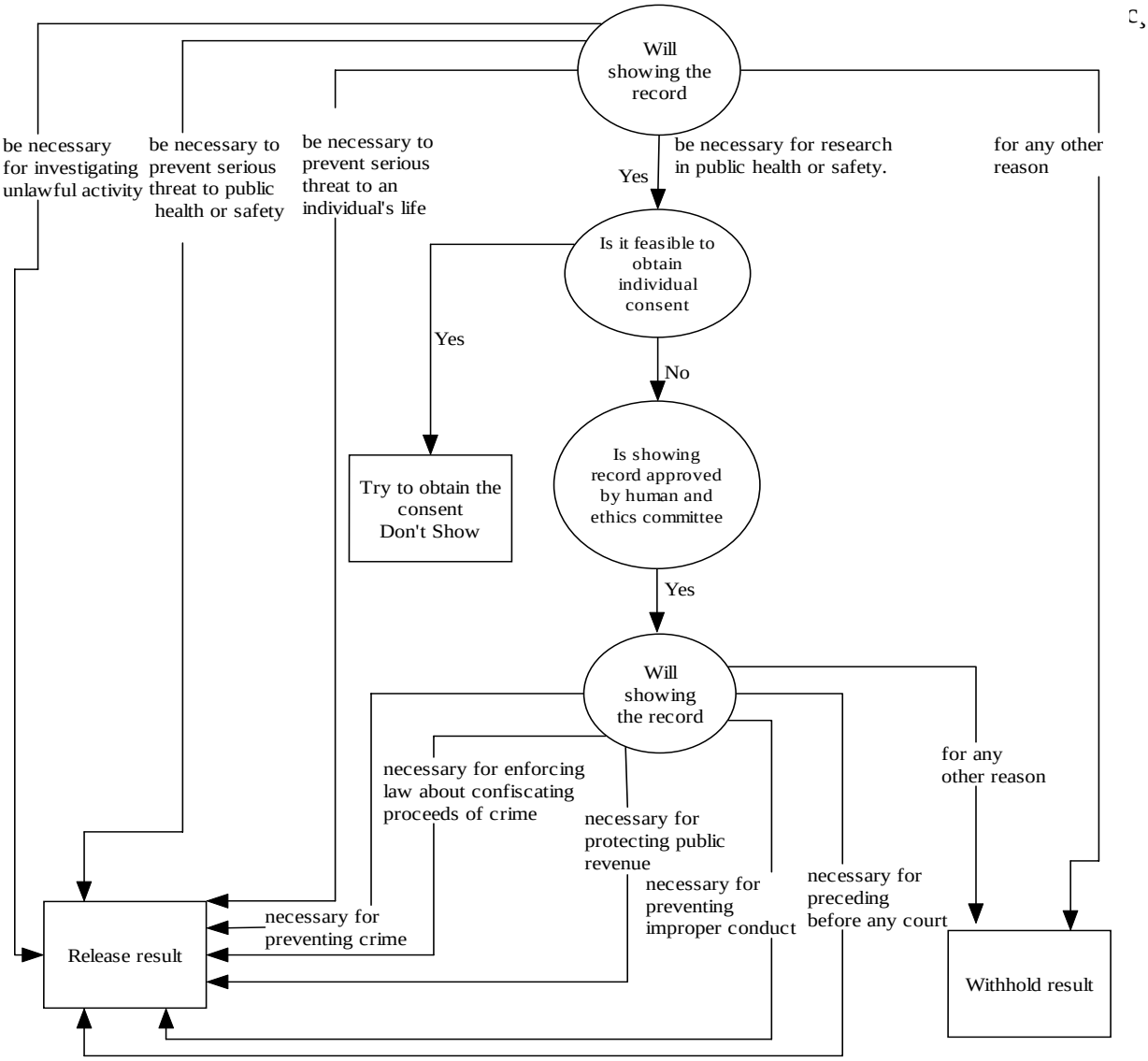

Figure 4 Decision tree privacy legislation

References:

[1] Norman S. Matloff, Another Look at the Use of Noise Addition for Database Security, *IEEE*, 1986, pp.173-180

[2] Krishnamurty Muralidha and Rahul Parsa and Rathindra Sarathy, A General Additive Data Perturbation Method for Database Security, *Management Science*, October 1999, Vol. 45, No. 10, pp. 1399–1415

[3] Krishnamurty Muralidha and Rathindra Sarathy, Security of Random Data Perturbation Methods, *ACM Transactions on Database Systems*, December 1999, Vol. 24, No. 4, pp. 487–493.

Measures in Statistical Disclosure Control of Tabular Data, *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*, 2002, ISBN:0-7695-1632-7, pp.227-231.

[5] Dorothy E. Denning and Peter J. Denning, The Tracker: A Threat to Statistical Database Security, *ACM Transactions on Database Systems*, March 1979, Vol. 4, No. 1, pp.76-96.

[6] Ljilijana Brankovic and Mirka Miller, Introduction to Statistical Database Security, Selected Topics of Cryptography and

Information Security, *Communications of the CCISA*, 2003,Vol.9, No.4, pp.1-30.

[7] Mirka Miller, A model of statistical database compromise incorporating supplementary knowledge. In: *Databases in the 1990's (B Srinivasan and J. Zeleznikow (editors), world Scientific*, 1991, pp 97-113.

[8] N.R.Adam and J.C. Wortmann, Security-control methods for statistical databases: A comparative study, *ACM Computing Surveys*, Vol.5, No.3, 1980, pp. 316-338.

[9] D.E.R Denning, *Cryptography and Data Security*, *Addison*-Wesley, 1982.

[10] *Guidelines approved under Section 95A of the Privacy Act 198,* Dec.2001, National Health Medical Research council.

[11]Zbigniew Miochalewicz, Functional Dependencies and their connection with security of statistical databases, *Inform. Systems*, Vol.12, No.1, 1987, pp.17-27.

[12] A. Stranieri, J. Zeleznikow and J. Yearwood, Argument structures that integrate dialectical and non-dialectical reasoning, *The Knowledge Engineering Review*, 16, issue 4, pp. 331-348.