Journal of Official Statistics
Vol. 9. No. 2, 1993, pp. 507–510
© Statistics Sweden

Discussion
Computer Security

Teresa F. Lunt[1]

1. Introduction

The paper by Sallie Keller-McNulty and Elizabeth Unger attempts to bring together the fields of statistical data disclosure control and secure database system technology. Although the communities of interest have the common goal of protecting information, they have traditionally had very little interaction. As a result, each community is largely ignorant of the significant amount of work that has been accomplished in the other community.

The work that has been accomplished has taken different paths in the two communities. The statistical data disclosure control community has focused on properties of the released information that make it difficult for an attacker to infer sensitive information. In contrast, the database security community has focused on the development of technology to protect sensitive information from those not authorized. Thus, although their approaches have been different, both communities can benefit from further interaction.

2. Computer Security

Computer security is concerned with the ability of a computer system to enforce a security policy governing the disclosure, modification, or destruction of information. The security policy may be specific to the organization, or may be generic. For example, the U.S. Department of Defense (DoD) *mandatory security* (or multilevel security) policies restrict access to classified information to cleared personnel. *Discretionary security* policies, on the other hand, define access restrictions based on the identity of users (or groups), the type of access (e.g., select, update, insert, delete), the specific object being accessed, and perhaps other factors (time of day, which application program is being used, etc.). Different types of users (system managers, database administrators, and ordinary users) may have different access rights to the data in the system. The access controls commonly found in most database systems today are examples of discretionary access controls.

A multilevel database systems supports data having different classifications or *access classes* and users having different clearances. In the most general case, the ability to individually classify atomic facts in a database is required. In relational database systems, this means that data are classified at the level of individual data elements. Special cases of multilevel relations may be classfied at the attribute level (i.e., all the data associated with a particular attribute have the same classification); at the row level (i.e., every tuple has a single classification); or at the relation level (i.e.,

[1] SRI International, Computer Science Laboratory, 333 Ravenswood Avenue, Menlo Park, CA 94025.

all the data in the table have the same classification). Although these mandatory policies are generally phrased in military terms, many believe that civil and commercial sector security policies can be formed in similar terms.

Computer security is also concerned with the ability to provide convincing arguments or proof that the security mechanisms work as advertised and cannot be disabled or subverted. In building multilevel database systems, providing such assurance is especially challenging because large, complex mechanisms may be involved in the enforcement of the security policy. To provide high assurance in trusted computer systems, the Department of Defense has developed a set of criteria for evaluating such systems. The evaluation criteria center on the notion of a *security kernel.* A security kernel is a mechanism that mediates all access attempts to data objects by users or processes acting on their behalf. The security kernel must be shown to be tamperproof and nonbypassable, and it must be small enough to be verified to be correct and secure with respect to the policy it enforces. At the highest evaluation class, such verification consists of formal mathematical proof of security. The information that must be protected with such high assurance is that whose compromise or damage could result in significant harm to national security. The sensitivity of the information is reflected in the severe penalties that can be imposed for its inappropriate disclosure, which can entail capital punishment for willful disclosure.

3. Multilevel Security vs. Statistical Database Security

3.1 What is considered sensitive

Statistical database security is concerned with protecting information concerning individuals. It is acceptable and even desirable for data users to be able to form a big picture from the data, but it is considered vital to prevent the inference that a particular data record corresponds to a particular individual. The threat is that a data spy may observe the data set and be able to link a particular record with a specific individual even though all identifying characteristics have been removed and the data have been masked or perturbed. The need for statistical database security derives from ethical and privacy concerns. Government agencies are required by law or regulation to protect the data they collect in the course of administering their programs, and in many cases the use of survey data collected by government agencies or individual researchers outside the government is restricted by the terms of informed consent agreements made between the data collector and the survey respondent.

In contrast, multilevel security is concerned with protecting classified information, and most computer security work to date has been driven by the requirements of the Department of Defense, in particular the consequential harm from the inappropriate disclosure of classified information. Whereas statistical database security is concerned with protecting individual data records while making the big picture available, multilevel database security is concerned with preventing the inference of a big picture from individual data records, since knowledge of the big picture tends to be more highly classified than the individual data items that contribute to it. Thus, in multilevel systems, access to individual data records can be granted until enough items have been released that it may become possible for a user to grasp the overall situation. This is generally referred to as the "aggregation problem" in the multilevel security community.

3.2. Protection techniques

In both multilevel database security and statistical database security use may be made of disinformation, although disinformation is used in different ways. For example, multilevel databases may provide for disinformation in the form of cover stories, whereas statistical data sets may use various data perturbation techniques to mask the data, such as swapping data values, combining adjacent records into groups, suppressing values, or adding noise.

In statistical database security, various methods of obscuring the data while preserving various aggregate statistics are employed to prevent meaningful links to be made between the data and the individuals they represent. Such data masking would not be acceptable, in general, in most multilevel database applications envisioned. However, one could view polyinstantiation, in which there may be different versions of the same real-world entity stored in the database at different security levels, as a simple and benign form of data masking; yet even this degree of data perturbation is causing a great deal of controversy in the multilevel database research community.

Polyinstantiation can take the following forms in a multilevel database:

- Different records with the same record identifier(s) but having different security levels. Thus, there could be a top-secret employee with employee number 12345 and a *different* unclassified employee also having employee number 12345, but the fact that the top-secret employee even exists is not known to unclassified users.
- Different records that are identical except for the value and classification of some nonidentifying attribute(s). Thus, there could be an unclassified employee with employee number 12345 whose unclassified salary is \$45,000 and whose secret salary is \$75,000.

Polyinstantiation will typically be used for implementing cover stories designed to provide the unclassified world with plausible explanations for unavoidably observable information that could otherwise lead to partial or complete inference of sensitive information.

3.3. Preventing undesired inferences

The multilevel inference problem arises whenever some data x can be used to derive partial or complete information about some other data y, where y is classified higher than x. In some cases, even learning of the existence of the information may be unacceptable. The aggregation problem also arises from an attempt to protect a sensitive relationship among otherwise nonsensitive data.

The possibility of statistical inference has been largely ignored in the multilevel database security community, even in projects which have proposed solutions to the multilevel "inference problem". While at first glance the problems appear to be quite different, upon deeper inspection it seems apparent that techniques of statistical inference will have to be considered in any comprehensive approach to the multilevel inference problem.

In contrast with the work to date on multilevel inference and aggregation, which has not resulted in any formal or quantitative definition of the problem, there has been a great deal of quantitative work in the statistical database security community, where the problem has been precisely defined. Most multilevel aggregation problems are phrased in vague and general terms, where an entire collection of data is, say, top-secret, and single data items are unclassified, but certain subsets of the data may allow you to form a larger picture (although still just a part of the big picture) and thus are considered, say, secret. However, these subsets are not generally identified; the problem is too fuzzy.

Government agencies have traditionally taken a conservative approach to the statistical inference problem by making only static datasets available that have been adequately masked so that undesirable inferences cannot be made. In the multilevel world, however, most people see a need for online interactive query of dynamic databases.

Work is beginning at SRI in the area of inferential security for multilevel database systems. We are beginning to investigate data design techniques that will minimize the amount of illegal inferences that can occur. Consideration of possible statistical attacks will have to be considered in any comprehensive approach to the problem. The eventual goal is to design and build tools for the data designer to use when constructing a multilevel application on a multilevel database.

3.4. *Disclosure policies*

Multilevel policies for disclosure are more specific and codified than are security policies for statistical data. Multilevel policies may also be more amenable, in general, to enforcement with high assurance in a computer system, through the use of security kernels. Higher level policies dealing with such issues as informed consent may be more problematic, although it is too early to say until the issue has been investigated. Also, the need for high assurance solutions for statistical database security may be questionable. This is because the attackers are not expected to be as well organized or well financed as the adversaries the DoD is concerned about, and also because there is not necessarily consequential harm if an individual's data are disclosed.

There are also costs to *not* disclosing data. In DoD applications the cost may be loss of mission or even loss of life, whereas in the statistical case the cost is the loss of the benefits that could have been gained by the research that could be performed using the data.

4. Summary

Both the statistical database security community and the multilevel database security community have a great deal to share. As computing technology advances and statistical data are stored on networked computer systems by government agencies and universities, the data become more vulnerable to exposure or even alteration. The computer security community has developed technologies that are now available in commercial products to address some of the vulnerabilities. At the same time, the statistical database security community has developed a large body of knowledge that could profitably be applied to the analysis of the multilevel inference problem.