Protecting confidential data with math

December 16 2011

Statistical databases (SDBs) are collections of data that are used to gather and analyze information from a variety of sources. The data may be derived from sales transactions, customer files, voter registrations, medical records, employee rosters, product inventories, or other compilations of facts and figures.

Because database security requires multiple processes and controls, it presents huge security challenges to organizations. With the computerization of databases in healthcare, forensics, telecommunications, and other fields, ensuring this kind of security has become increasingly important.

In a paper published Thursday in the *SIAM Journal on Discrete Mathematics*, authors Rudolf Ahlswede and Harout Aydinian analyze a security-control model for statistical databases.

"Providing privacy and confidentiality in SDBs is not a new issue," Aydinian points out. "Privacy interests have evolved from the very first census in the United States. Recorded protests until the mid-20th century reflect constitutional issues resulting from the requirement for U.S. residents to provide sensitive personal information. Questions on census forms about diseases, mortgage values, and other items have raised many concerns."

While such databases are very helpful in aggregating data, there is a risk that confidential information about an individual's record may be deliberately compromised. "Since such data sets also contain sensitive information, such as the disease of an individual, or the salary of an employee, it is necessary to provide security against the disclosure of confidential information," says Aydinian. "Even in cases where a user has no direct access to sensitive information, sometimes confidential data about an individual can be inferred by correlating enough statistics."

Typically, statistical databases are designed to only accept queries that involve specific statistical functions (such as sum, average, count, min, max, etc.). However, the use of these queries may render databases susceptible to compromise. For instance, it may be possible to infer information about specific individuals by putting together data from a sequence of statistical queries, using prior knowledge of an individual, or through collusion among users.

An SDB is considered secure if no protected data can be inferred from available queries. "In the literature, many scenarios of compromise and inference control methods have been proposed to protect SDBs," Aydinian says. "However, to date no one security control method is capable of completely preventing compromise."

Query restriction is one of several general approaches used for security control. A "query request" retrieves a subset of data from a database that meets a set of conditions. In query restriction, the kind and amount of data that can be retrieved by such queries is limited, for example, the size of the data, or the amount of overlap between data that is returned.

In one type of query restriction method, only certain sums of individual records (called "SUM queries") that meet a minimum specified size or number, and satisfy a specified set of conditions, are available to users.

Aydinian explains with an example. "Consider a company with a large number of employees. Suppose that for each member of the company, the sex, age, rank, length of employment, salary etc. is recorded. The salaries of individual employees are confidential. Suppose that only SUM queries are allowed, i.e. the sum of the salaries of the specified people is returned. Then one might pose the query: What is the sum of salaries for males, above 50, and during the last 10 years?"

The task addressed in the paper is to provide an optimal collection of SUM queries that prevents compromise of confidential information—such as individual salaries, for instance. A natural solution is to maximize the number of available SUM queries. The authors obtain tight bounds for the maximum number of such queries that return subsets of data without compromising groups of entries.

"Future work in the query-restriction approach includes evaluation of new security-control mechanisms, which are easy to implement and guarantee absolute security," says Aydinian. "At the same time, it is desirable that these methods satisfy other criteria like richness of available queries, consistency, cost etc. It also seems promising to develop methods combining different security control mechanisms."

More information: On Security of Statistical Databases. Rudolf Ahlswede and Harout Aydinian, *SIAM Journal on Discrete Mathematics*, 25, pp 1778-1791 (Online publish date: December 15, 2011)

Provided by Society for Industrial and Applied Mathematics

Citation: Protecting confidential data with math (2011, December 16) retrieved 9 April 2024 from https://phys.org/news/2011-12-confidential-math.html