Inference Control in Statistical Databases

Ernst L. Leiss
Department of Computer Science
University of Houston
coscel@cs.uh.edu

Partially funded under NSF Grant #1241772

Any opinions, findings, conclusions, or recommendations expressed herein are those of the author and do not reflect the views of the National Science Foundation

1. **Confidential Data: Legal Requirements and Societal Expectations (20 min)**
2. **Data Aggregation (20 min)**
3. **Statistical Databases: Inference Control (70 min)**
4. **Conclusion (10 min)**

1. Confidential Data: Legal Requirements and Societal Expectations

With larger databases and greater data collection capabilities, more and more confidential data are aggregated in data repositories

Technology: Unexpected and unknown capabilities

May clash with a society's expectations and ethics

The expectation of privacy: may be contradicted by new technological capabilities

Legal landscape: US vs. EU

The European Union has far more stringent privacy and confidentiality requirements than the US. EU requirements are very general (referring to all data that can be used to identify an individual) while US privacy provisions are sectoral (typically financial, medical, student data).

How can the confidentiality of personal data (an important aspect of privacy) be safeguarded? It may be legally required, or it may be good customer relations to do so.

2. Data Aggregation

Two entirely innocuous databases may be combined to identify uniquely an individual

Example:

Consider Date of Birth (DoB) and consider the US 5-digit zip code

The US has approximately 320M inhabitants (2013).

There are no more than 100 000 (5-digit) zip codes: on average, there are at least 3 200 inhabitants with the same zip

Assuming a cut-off at 80, there are about 365 x 80 birth dates: on average, there are about 12 000 people with the same DoB

Therefore, each item in each of the two databases is entirely safe since there are on average thousands of individuals that correspond to any such item (DoB or zip)

However, combining DoB and zip identifies uniquely many, if not most individuals, since

100 000 x 365 x 80 = 2 920 000 000 >> 320M

Note: This is a statistical argument; there is no guarantee that a specific combination will not be satisfied by more than one individual (e.g., twins living in the same household).

3. Statistical Databases: Inference Control

The notion of a statistical database

Two different access mechanisms:

Ordinary database access

Access to statistics

Ordinary access

Of primary concern are confidential data

Requires proper authorization, typically need to know

Involves access to individual entries

Example: A hospital's medical database

Treating physicians and nurses must have access to patient data, <u>but only</u> their patients', not of those of other healthcare providers

Access to statistics

No access to individual entries, but rather to statistics over sets of individuals

E. g., averages, medians

Example continued:

Medical database containing diagnoses and related information

A public health professional wants to design a campaign to educate people about HIV infections. It would be useful to be able to target the most susceptible population group within the area served by the hospital. Thus, it would be useful to know more about the "average person infected with HIV" so the campaign may be more specifically tailored to that population. This necessitates statistical queries.

Clearly, this public health official must not have access to any information about individuals in the database (e. g., is John Doe HIV-positive?).

Confidential data, legal requirements to maintain confidentiality:

The use of statistical queries must guarantee confidentiality

Inference control:

It must be impossible to use legitimate statistical queries in order to obtain access to individual confidential data

Of particular concern is combining various statistical queries so that access to individual entries is achieved

In practice, inference control is extremely difficult to attain

Extensive literature, for many approaches, restrictions, and models of statistical databases

Example

Assume ordinary SQL queries

- **Each query involves a set of elements**
- **Typical: Median, average, sum**
- **Confidential information in statistical queries is numerical**

Assume query type SUM:

A query defines a set of elements and returns the sum of the confidential information associated with all the elements in the set

In SQL queries, we can do or (forming the union of the underlying sets) and not (forming the complement of the underlying set)

Note that at the technical/implementation level, the queries will always be based on (sets of) individual entries; the difference between ordinary queries (on a need-to-know basis) and statistical queries is what is being divulged to the user who poses the query

In order to obtain a secure statistical database, we impose the following restriction:

A statistical query q is legal if and only if

$$h <= NU(q) <= N-h$$

N is the total number of elements in the database, h is an arbitrary value, and NU(q) is the number of elements in the set underlying the query q

In other words, a legal statistical query must match more than only a few entries (at least h entries) and may not match too many entries (at least h entries must not be involved in the query). The value h is a parameter that can be chosen depending on the preference of the database administrator.

Larger h == more paranoia, smaller h == less paranoia.

Thus there are two bad cases, when q is too small (matches not enough entries) and when q is too large (matches too many entries)

This requirement appears to be reasonable and eminently intuitive.

Unfortunately, if fails spectacularly to prevent inference control.

First, define a general tracker GT:

This is any query GT such that

2h <= NU(GT) <= N-2h

In other words, it must be a bit larger (2h instead of h) than an ordinary legal statistical query and it must be a bit bigger (N-2h instead of N-h) than such a query

Consider now an illegal query q_{bad}:

q_{bad} may be either too small (NU(q_{bad})<h) or too large (NU(q_{bad})>N-h).

The expectation is that the value of q_{bad} cannot be determined.

This is wrong, as its value is obtained as follows:

First compute x := SUM(GT) + SUM($\underline{\text{not}}$GT)

SUM(q_{bad}) = SUM(q_{bad} $\underline{\text{or}}$ GT) + SUM(q_{bad} $\underline{\text{or}}$ $\underline{\text{not}}$ GT) – x q_{bad} **$\underline{\text{too small}}$**

SUM(q_{bad}) = 2x – [SUM($\underline{\text{not}}$ q_{bad} $\underline{\text{or}}$ GT) + SUM($\underline{\text{not}}$ q_{bad} $\underline{\text{or}}$ $\underline{\text{not}}$ GT) q_{bad} **$\underline{\text{too large}}$**

In either case, one can verify that the necessary queries are legal statistical queries and that their combination yields the supposedly confidential information, the value of the illegal query q_{bad}. While one does not know beforehand why the query is illegal (too small or too large), only one of the two cases applies and only that case will yield legal queries.

Upshot

Impossible to maintain the confidentiality of the information of individual entries

Similar results hold for virtually all type of restrictions that are imposed on the queries

There is a large body of literature to that effect.

Randomizing is a possible solution, except the responses are falsified (slightly)

In other words, instead of providing the true response, the actual response is slightly modified, either by adding a randomly selected element from the database or deleting one of the elements involved in the query, and then returning the statistic based on the modified set rather than the original set of entries. This change in the value of the response can be exploited to prevent reasonably well the disclosure of confidential information thereby providing inference control. However, some users find the randomization unacceptable.

4. Conclusion

With the advent of larger and more centralized databases, statistical information based on these databases will become more important.

Inference control is paramount when the information collected in these databases is considered confidential.

Confidentiality may be legally required (e. .g, medical, financial, student data) or the consumer may insist on it.

Most mechanisms designed to achieve inference control do not work.