Principles of Computer Security

Alvin Lin

January 2018 - May 2018

Statistical Inferencing

Statistical databases provide aggregate statistics for a subset of entities. It is difficult to provide useful data while protecting the confidentiality of each individual entity due to inferencing. A user can combine the information in the database with publicly available information to retrieve personally identifiable information. The data in the database is secure if the set of publicly known and unknown data are disjoint, but they often are not.

Statistical databases containing medical information are often used for research, but must be implemented to comply with laws and respect data sensitivity. Researchers need accurate meaningful data while respective patient privacy and confidentiality laws.

Static and Dynamic Databases

Static databases are created once and never changed, such as the US Census, which occurs every 10 years. Dynamic databases reflect real-time data changes. Most online research databases are of this form.

Methods for Inference Handling

- Access restriction: databases normally have different access levels for different types of users. Doctors and healthcare representatives may need full access to the information, while researchers should only have access to partial information.
- Query set restriction: only allow aggregate queries and not specific queries on database elements. This is problematic because individual values can be

deduced from a few queries of the value averages. Query set restriction cannot prevent disclosure, but can make it arbitrarily hard.

- Microaggregation: the raw data is grouped into small aggregates before publication such that the average value of the group replaces each individual value. Data with the most similarity is grouped together to maintain data accuracy.
- Data perturbation: noise is added to the raw data to prevent true values from being disclosed if unauthorized data is accessed. However, data perturbation runs the risk of presenting biased data.
- Output perturbation: noise is added to the output of queries. This still runs the risk of bias as before, but it is less severe than with data perturbation.
- Random sampling: only a sample of records satisfying the query requirements are dhown. However, logically equivalent queries can result in different results from different query sets.
- Auditing: each query made by each user is tracked to check if it is malicious. This is usually done after-the-fact.

Inference prevention is a hard problem. Many techniques exist, but none are completely satisfactory. Some are too costly to apply, while others adversely impact data integrity. When comparing these methods, we have to take into account the following:

- Security: the possibility of exact or partial disclosure
- Information Richness: the amount of non-confidential information eliminated, bias, precision, and consistency
- Cost: the effort needed for the initial implementation and any additional query processing overhead

Method	Security	Information Richness	Costs
Query Set Restriction	Low	Low	Low
Micro-aggregation	Moderate	Moderate	Moderate
Data Perturbation	High	High-Moderate	Low
Output Perturbation	Moderate	Moderate-Low	Low
Sampling	Moderate	Moderate-Low	Moderate
Auditing	Moderate-Low	Moderate	High

You can find all my notes at `http://omgimanerd.tech/notes`. If you have any questions, comments, or concerns, please contact me at alvin@omgimanerd.tech