Security Control Methods for Statistical Database

Li Xiong

CS573 Data Privacy and Security
A statistical database is a database which provides statistics on subsets of records

OLAP vs. OLTP

Statistics may be performed to compute SUM, MEAN, MEDIAN, COUNT, MAX AND MIN of records
Types of Statistical Databases

- **Static** – a static database is made once and never changes
  - Example: U.S. Census

- **Dynamic** – changes continuously to reflect real-time data
  - Example: most online research databases
Types of Statistical Databases

- **Centralized** – one database
- **Decentralized** – multiple decentralized databases
- **General purpose** – like census
- **Special purpose** – like bank, hospital, academia, etc
Access Restriction

- Databases normally have different access levels for different types of users
- User ID and passwords are the most common methods for restricting access
  - In a medical database:
    - Doctors/Healthcare Representative – full access to information
    - Researchers – only access to partial information (e.g. aggregate information)
- Statistical database: allow query access only to aggregate data, not individual records
## Accuracy vs. Confidentiality

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Confidentiality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researchers want to extract accurate and meaningful data</td>
<td>Patients, laws and database administrators want to maintain the privacy of patients and the confidentiality of their information</td>
</tr>
</tbody>
</table>
Data Compromise

- Exact compromise – a user is able to determine the exact value of a sensitive attribute of an individual
- Partial compromise – a user is able to obtain an estimator for a sensitive attribute with a bounded variance
- Positive compromise – determine an attribute has a particular value
- Negative compromise – determine an attribute does not have a particular value
- Relative compromise – determine the ranking of some confidential values
Security Methods

- Query restriction
- Data perturbation/anonymization
- Output perturbation
Comparison

- Query restriction cannot avoid inference, but they accurate responses to valid queries.
- Data perturbation techniques can prevent inference, but they cannot consistently provide useful query results.
- Output perturbation has low storage and computational overhead, however, is subject to the inference (averaging effect) and inaccurate results.
Statistical database vs. data anonymization

- Data anonymization is one technique that can be used to build statistical database
- Data anonymization can be used to release data for other purposes such as mining
- Other techniques such as query restriction and output perturbation can be used to build statistical database
Evaluation Criteria

- Security – level of protection
- Statistical quality of information – data utility
- Cost
- Suitability to numerical and/or categorical attributes
- Suitability to multiple confidential attributes
- Suitability to dynamic statistical DBs
Security

- Exact compromise – a user is able to determine the exact value of a sensitive attribute of an individual
- Partial compromise – a user is able to obtain an estimator for a sensitive attribute with a bounded variance
- Statistical disclosure control – require a large number of queries to obtain a small variance of the estimator
Statistical Quality of Information

- Bias – difference between the unperturbed statistic and the expected value of its perturbed estimate
- Precision – variance of the estimators obtained by users
- Consistency – lack of contradictions and paradoxes
  - Contradictions: different responses to same query; average differs from sum/count
  - Paradox: negative count
Cost

- Implementation cost
- Processing overhead
- Amount of education required to enable users to understand the method and make effective use of the SDB
Security Methods

- Query set restriction
  - Query size control
  - Query set overlap control
  - Query auditing
- Data perturbation/anonymization
- Output perturbation
Query Set Size Control

- A query-set size control limit the number of records that must be in the result set.
- Allows the query results to be displayed only if the size of the query set $|C|$ satisfies the condition
  
  $$K \leq |C| \leq L - K$$

  where $L$ is the size of the database and $K$ is a parameter that satisfies $0 \leq K \leq L/2$. 
Query Set Size Control
Tracker

- Q1: Count ( Sex = Female ) = A
- Q2: Count ( Sex = Female OR (Age = 42 & Sex = Male & Employer = ABC) ) = B

If B = A+1

- Q3: Count ( Sex = Female OR (Age = 42 & Sex = Male & Employer = ABC) & Diagnosis = Schizophrenia)

Positively or negatively compromised!
Query set size control

- With query set size control the database can be easily compromised within a frame of 4-5 queries.
- For query set control, if the threshold value $k$ is large, then it will restrict too many queries.
- And still does not guarantee protection from compromise.
Query Set Overlap Control

- Basic idea: successive queries must be checked against the number of common records.
- If the number of common records in any query exceeds a given threshold, the requested statistic is not released.
- A query $q(C)$ is only allowed if:
  $$|X(C) \cap X(D)| \leq r, \ r > 0$$
  Where $\alpha$ is set by the administrator.
- Number of queries needed for a compromise has a lower bound $1 + (K-1)/r$.
Query-set-overlap control

- Ineffective for cooperation of several users
- Statistics for a set and its subset cannot be released – limiting usefulness
- Need to keep user profile
- High processing overhead – every new query compared with all previous ones
Auditing

- Keeping up-to-date logs of all queries made by each user and check for possible compromise when a new query is issued
- Excessive computation and storage requirements
- “Efficient” methods for special types of queries
Audit Expert (Chin 1982)

- Query auditing method for SUM queries
- A SUM query can be considered as a linear equation

\[ \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_n x_n = q \]

where \( \lambda_i \) is whether record \( i \) belongs to the query set, \( x_i \) is the sensitive value, and \( q \) is the query result

- A set of SUM queries can be thought of as a system of linear equations

\[ BX = Q \]

- Maintains the binary matrix representing linearly independent queries and update it when a new query is issued
- A row with all 0s except for \( i \)th column indicates disclosure
Audit Expert

- Only stores linearly independent queries
- Not all queries are linearly independent
  - Q1: $\text{Sum(Sex=M)}$
  - Q2: $\text{Sum(Sex=M AND Age>20)}$
  - Q3: $\text{Sum(Sex=M AND Age<=20)}$
Audit Expert

- \( O(L_2) \) time complexity
- Further work reduced to \( O(L) \) time and space when number of queries < \( L \)
- Only for SUM queries
- No restrictions on query set size
- Maximizing non-confidential information is NP-complete
Auditing – recent developments

- Online auditing
  - “Detect and deny” queries that violate privacy requirement
  - Denial themselves may implicitly disclose sensitive information

- Offline auditing
  - Check if a privacy requirement has been violated after the queries have been executed
  - Not to prevent
Security Methods

- Query set restriction
- Data perturbation/anonymization
  - Partitioning
  - Cell suppression
  - Microaggregation
  - Data perturbation
- Output perturbation
Partitioning

- Cluster individual entities into mutually exclusive subsets, called atomic populations
- The statistics of these atomic populations constitute the materials
Microaggregation

Original Data → Averaged → Microaggregated Data

User

Query

Results
Data Perturbation

Original Database → Noise Added → Perturbed Database

User 1 queries Perturbed Database

User 2 queries Perturbed Database
Security Methods

- Query set restriction
- Data perturbation/anonymization
- Output perturbation
  - Sampling
  - Varying output perturbation
  - Rounding
Output Perturbation

- Instead of the raw data being transformed as in Data Perturbation, only the output or query results are perturbed.
- The bias problem is less severe than with data perturbation.
Output Perturbation

Original Database

Query

User 1

Query

Noise Added to Results

User 2

Query

Results

Results
Random Sampling

- Only a sample of the query set (records meeting the requirements of the query) are used to compute and estimate the statistics.
- Must maintain consistency by giving exact same results to the same query.
- **Weakness** - Logical equivalent queries can result in a different query set – consistency issue.
Varying output perturbation

- Apply perturbation on the query set
- Less bias than data perturbation
## Some Comparisons

<table>
<thead>
<tr>
<th>Method</th>
<th>Security</th>
<th>Richness of Information</th>
<th>Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query-set-size control</td>
<td>Low</td>
<td>Low&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Low</td>
</tr>
<tr>
<td>Query-set-overlap control</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Auditing</td>
<td>Moderate-Low</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td>Partitioning</td>
<td>Moderate-low</td>
<td>Moderate-low</td>
<td>Low</td>
</tr>
<tr>
<td>Microaggregation</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Data Perturbation</td>
<td>High</td>
<td>High-Moderate</td>
<td>Low</td>
</tr>
<tr>
<td>Varying Output Perturbation</td>
<td>Moderate</td>
<td>Moderate-low</td>
<td>Low</td>
</tr>
<tr>
<td>Sampling</td>
<td>Moderate</td>
<td>Moderate-Low</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

<sup>1</sup> Quality is low because a lot of information can be eliminated if the query does not meet the requirements
Sources

- Partial slides:  
  http://www.cs.jmu.edu/users/aboutams
