CS573 Data Privacy and Security

Statistical Databases

Li Xiong

Department of Mathematics and Computer Science
Emory University
• Statistical databases
  – Definitions
  – Early query restriction methods
  – Output perturbation and differential privacy (next lecture)
Statistical Database

• A statistical database is a database which provides statistics on subsets of records

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

• two types:
  – pure statistical database:
    • only stores statistical data. e.,g., a census database.
  – ordinary database with statistical access
    • contains individual entries
    • some users have normal access, others statistical
Statistical Database

• Objective: provide statistical users with the aggregate information without compromising the confidentiality of any individual entity represented in the database

• Database administrator must prevent, or at least detect, the statistical user who attempts to gain individual information through one or a series of statistical queries

• Inference control to prevent inference from statistics to individual records

Slide credit: Dr Lawrie Brown (UNSW@ADFA) for “Computer Security: Principles and Practice”, 1/e, by William Stallings and Lawrie Brown, Chapter 5 “Database Security”.
Statistical Database Security

• Statistics are derived from a database by means of a characteristic formula $C$
  – logical formula over the values of attributes
  – E.g., $C = (\text{Age} = 42) \& (\text{Sex} = \text{Male}) \& (\text{Employer} = \text{ABC})$

• Query set $X(C)$ of characteristic formula $C$ is the set of records matching $C$

• Statistical query is a query that produces a value calculated over a query set
  • E.g., $\text{COUNT}(\text{Age}=42)$
Inference from a Statistical Database

• Statistical user is restricted to obtaining only aggregate, or statistical, data from the database and is prohibited access to individual records

• Inference problem:
  – user may infer confidential information about individual entities represented in the SDB
  – Such an inference is called a compromise

• Positive compromise – determine an attribute has a particular value

• Negative compromise – determine an attribute does not have a particular value

• In some cases, a sequence of queries may reveal information
Inference from a Statistical Database

• The inference problem for an SDB can be stated as follows:
  – A characteristic function $C$ defines a subset of records (rows) within the database
  – A query using $C$ provides statistics on the selected subset
  – If the subset is small enough, perhaps even a single record, the questioner may be able to infer characteristics of a single individual or a small group
Methods

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

- Data perturbation introduces noise in the data
- Provides answers to all queries, but the answers are approximate
Query Restriction

- Rejects a query that can lead to a compromise
- The answers provided are accurate.
Output Perturbation

- perturbs the answer to user queries while leaving the data in the SDB unchanged
- generate statistics that are modified from those that the original database would provide
Methods

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

- Simplest form of query restriction
- A query-set size control limit the number of records that must be in the result set
- Query \( q(C) \) is permitted (allows the query results to be displayed) only if the number of records that match \( C \) satisfies the condition

\[
K \leq |X(C)| \leq L - K
\]

where \( L \) is the size of the database and \( K \) is a parameter that satisfies \( 0 \leq K \leq \frac{L}{2} \)
Query Set Size Control

Original Database

Query 1
Query 1 Results

Query 2
Query 2 Results

Query 1 Results

Query 2 Results

K

NO

YES
Tracker

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

What if B = A+1?
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)

- if response to Q3 is B
- if response to Q3 is A
Positively or negatively compromised!
Query set size control

• If the threshold value $K$ is large, then it will restrict too many queries
  – And still does not guarantee protection from compromise
• The database can be easily compromised within a frame of 4-5 queries
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 the number of records that match $C$ satisfies:
  \[ |X(C) \cap X(D)| \leq r, \quad r > 0 \]
  for all $q(D)$ that have been answered for this user, and where $r$ is a fixed integer greater than 0.
Query-set-overlap control

• Statistics for a set and its subset cannot be released – limiting usefulness
• High processing overhead – every new query compared with all previous ones
• Multiple users - need to keep user profile, need to consider collusion between users
• Still no formal privacy guarantee
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
• Only “efficient” methods for special types of queries
Audit Expert (Chin 1982)

• Query auditing method for SUM queries
• Given sensitive values $x_1, ..., x_L$, any SUM query on those values can be modeled as an equation $q = a_1 x_1 + a_2 x_2 \ldots + x_L x_L$
• where $a_i = 1$ if $x_i$ (record $i$) belongs to the query set and $a_i = 0$ otherwise, and $q$ is the query result
• A set of $m$ SUM queries can be thought of as a system of linear equations $AX = D$ where $A$ is an $m \times L$ binary matrix, $X$ is the vector of sensitive values, and $D$ is the vector of query result
• 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

- $O(L^2)$ time complexity
- Further work reduced to $O(L)$ time and space when number of queries $< L$
- Only for SUM queries
Auditing – recent developments

• Online auditing
  – “Detect and deny” queries that violate privacy requirement
  – Denial themselves may implicitly disclose sensitive information
  – Prevents privacy breaches on-the-fly

• Offline auditing
  – Check if a privacy requirement has been violated after the queries have been executed
  – Not to prevent - objective to check for compliance of privacy requirement
Methods

- Data perturbation/anonymization
- Query restriction
- Output perturbation
  - Differential privacy