Statistical DataBase Security

SDB- Basic Concepts

- The reference SDB is the relational form.
- Let N is the number of entities and M is the # of attributes of the SDB schema.
- The reference is in the next page where $X_{ij}$ denotes the value of the j-th attribute $A_j$ for the i-th record in the SDB...
- Each attribute $A_j$ has $|A_j|$ possible values.
Special-purpose SDBs, such as survey DBs, release statistics in the form of a table (called macrostatistics), e.g. next slide, Which is a three-dimensional for count.

- Statistics queries can be issued through either keys or characteristics formulas.
- A key-based statistics is the form of Sum (C, Salary), with C= (Ali, Taghi, Naghi, Javad).
- Characteristic formula (indicated by A, B, …) is a logical formula of attributed combined by Boolean operators OR, AND, NOT.
- Example:
  \[
  A = (\text{Sex}=F) \land ((\text{Dept-Code}=\text{Dept1}) \lor (\text{Dept code}=\text{Dept2})) \land (\text{Birth-Year}<1965)
  \]
A characteristic formula (CF) specifies a set of records called a query set. The query set of A indicated as $X(A)$ and the number of records in $X(A)$ is $|X(A)|$.

All: for each CF, $C$, $X(C) \subseteq X(\text{All})$, meaning that the query set of any $C$ is a subset of the whole SDB.

Elementary sets are a special query sets which are not decomposable.

- $C=(A_1=a_1) \land (A_2=a_2) \land \ldots \land (A_M=a_M)$, where $A_j$ is an attribute of the SDB and $a_j$ is one of the $|A_j|$ values.
- The number $E$ of the elementary sets in an SDB with $A_1 \ldots A_M$ attributes is $E=|A_1| \times |A_2| \times \ldots \times |A_M|$, and some of them may be empty.
- If $g$ indicates the size of the largest elementary set, and $N$ is the # of records, if $N \leq E$, then $g = 1$. 

---

**Example, BSD Table**

<table>
<thead>
<tr>
<th>Birth-Year</th>
<th>Sex</th>
<th>Dept-Code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dept1</td>
</tr>
<tr>
<td>1941-1951</td>
<td>M</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>1</td>
</tr>
<tr>
<td>1952-1962</td>
<td>M</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>20</td>
</tr>
<tr>
<td>&gt;1962</td>
<td>M</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>10</td>
</tr>
</tbody>
</table>

Rasool Jalili; 2nd semester 1384-1385; Database Security, Sharif Uni. of Tech.
The main statistical queries in an SDB are Count, sum, Rfreq, Avg, Medium, Max, Min:

- Count(c) = |X(C)|
- Sum(C, Aj) = \sum_{i \in X(C)} X_{ij}
- Rfreq(C) = \frac{\text{Count}(C)}{N}, relative frequency of the query set.
- Avg(C,Aj)
- Max(C,Aj)
- Min(C,Aj)
- Median(C,Aj) = \left\lfloor \frac{|X(C)|}{2} \right\rfloor, used to compute the median value in an ordered set of a numerical value of an attribute Aj.

A sensitive statistic is a statistic that can lead to identifying conf. inf on a single entity in the SDB.

**Inference Protection Techniques**

- Aim at preventing users from inferring conf. info about entities represented in SDBs, causing SDB to be compromised.

- Let Ai be a non-numerical conf. attr. or a numerical attr. and Xj describing an entity in SDB:

- Exact compromise: occurs if a user through some statistical query, can state for Ai the value 1 (non-numerical) or the exact value for numerical attr. for a record j in SDB.

- Partial compromise: occurs if a user through some statistical query, can state for Ai the value 0 (non-numerical) or obtain an estimator of the actual value such that its variance < k^2, where k is defined by DBA.
Techniques

- Conceptual, address the inference problem at a conceptual level, dealing with a conceptual data model of the SDB.

- Restriction-based, provide inference protection by restricting some statistical queries; queries containing a very small/very large number of records. Should take care of usability of SDB due to the large number of false denials on queries.

- Perturbation-based, provide inference protection through introducing modification of info used to answer statistical queries. The modification can be on either the stored data in SDB, or the computed results, before releasing to the user.

  Crucial problems:
  - The bias introduced in the responses which should be as small as possible to assure inference protection and also accuracy of responses.
  - Modifications must guarantee the consistency of the released results.

Conceptual Techniques

- The main conceptual techniques called the lattice model and the partitioning of the SDB entities into populations.

- The lattice model provides a conceptual representation of the records stored in an SDB, based on the lattice structure of m-dimensional tables.

- Next slide shows a m-table for the count statistic computed on Birth-year, Sex, and Dept-Code, also called macrostatistics form
Example, BSD Table

<table>
<thead>
<tr>
<th>Birth-Year</th>
<th>Sex</th>
<th>Dept-Code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dept1</td>
</tr>
<tr>
<td>1941-1951</td>
<td>M</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>1</td>
</tr>
<tr>
<td>1952-1962</td>
<td>M</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>20</td>
</tr>
<tr>
<td>&gt;1962</td>
<td>M</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>10</td>
</tr>
</tbody>
</table>

- An m-dimensional table represents a collection of correlated statistics of m order on m SDB attributes A₁, A₂,...,Aₘ.
- If each Aᵢ has |Aᵢ| value, the total number of table elements: Sₘ = |A₁|x|A₂|x…x|Aₘ|
- The possible statistics can be derived from the table is 2^{Sₘ} -1.
- The m-table related to a given statistic corresponds to a lattice structure, as in the next slide.
Table Lattice on Birth-Date, Sex, Dept-Code attributes

- $T_{al}$ (Sum in respect to an attribute)
- $T_B$
- $T_S$
- $T_D$
- $T_{BD}$
- $T_{BS}$
- $T_{SD}$
- $T_{BSD}$ (Count statistic on the whole SDB)

Rasool Jalili; 2nd semester 1384-1385; Database Security, Sharif Uni. of Tech.
• Sensitive statistics: may lead to disclosure of confidential info on single entities, even in negative.
• The criterion: \textit{n-respondent, %k dominance} is used. It means that \textit{n or fewer records represent more than %k of the total}.
• A statistic holding a query set of size 1 is sensitive. Sum of such unitary cells is also sensitive.

Conceptual Partitioning
• Used to protect inference during the conceptual design phase of the SDB.
• It is based on the definition of populations (set of entities) on which statistics can be released. And the conditions that must be verified to avoid inference.
• The Data Abstraction (D-A) model, based on the \textit{aggregation} and \textit{generalization} abstractions, is used to model the set of SDB entities on which statistical queries can be computed.
• \textit{Aggregation} allows object relationship to be represented as aggregated objects.
• \textit{Generalization} allows object classes to be represented as generic objects. \(\Rightarrow\) the real world is represented as generalization and aggregation hierarchies.
• For statistical purposes, generalization hierarchies are of interest.
• Moving from the root to leaves is the hierarchy, population is decomposed into sub-populations, down to the \textit{atomic} populations.
• Disjoint sub-populations holding a common parent form a \textit{cluster}. 
The conceptual model applied to the Employee SDB

- The model is used to define which statistics (count, sum, average, etc.) can be released and yet avoid inference.
- To support the definition of statistical security requirements, the SSMF (Statistical Security Management Facility) proposed.
Restriction-Based Techniques

- The techniques protect against inference by restricting (not releasing) the statistical sensitive queries. Includes:
- Query-set size control: controls the size of the query set associated with a statistical query. A statistic \( q(C) \) is permitted only if its query set \( X(C) \) satisfies:
  \[
  k \leq |X(C)| \leq N-k
  \]
  where \( N \) is the number of SDB records and \( k \geq 0 \) is a fixed parameter. \( k \) must satisfy \( 0 \leq k \leq N/2 \) in order not to release \( q(\text{All}) \), where \( q(\text{All}) = A(C) + Q(\neg C) \)
  This prevents simple attacks based on very small and very large.
- If a user knows a certain individual satisfies characteristic \( C \), if the \( q1=\text{Count}(C) \) returns 1, then …
• If $q_2 = \text{Count} (C \land C')$ also returns 1, then the intruder understand that the entity has $C'$. If $q_2$ returns 0, then the entity does not have $C'$ characteristic.

• Large query sets [e.g. $q(C)$] are also restricted, as $q(\neg C)$ is small and may be abused.

• More complex attacks, tracker.

• A tracker is a set of characteristic formulas can be used to pad out small-size query sets with additional records in the allowed range $[k, N-k]$.

---

**Example:**  
$C = (\text{Dept-Code}=\text{Dept} \land \text{Sex}=\text{F} \land \text{Birth-Year}=1951)$  
specifies an employee Hanieh.

• The query-size control has parameter $k=3 \Rightarrow$ the query is not permitted!

• A tracker is obtained by defining $A = (\text{Sex}=\text{F})$ and $B = (\text{Dept-Code}=\text{Dept} \land \text{Birth-Year}=1951)$ and $T = (A \land \neg B)$  
Now, $\text{Count}(C) = \text{Count}(A) - \text{Count}(T)$

• Moreover it is possible to infer additional info:  
$\text{Count}(C \land \text{Salary} \geq 20) = \text{Count} (T \lor A \land \text{Salary} \geq 20) - \text{Count}(T)$
• Expand query-size control
• A solution is to expand the # of query sets to be controlled to decide if info is released.
• i.e. given a characteristic formula C, we can define implied query set, query sets identified by characteristics directly derived from C. These query sets must also be controlled.
• Given an m-order statistics of the form:
  \[ q(A_1=a_1 \land A_2=a_2 \land \ldots \land A_m=a_m), \]
• There exist \(2^m\) implied query sets corresponding to the following statistics

\[ q(A_1=a_1 \land A_2=a_2 \land \ldots \land A_m=a_m) \]
\[ q(A_1=a_1 \land A_2=a_2 \land \ldots \land \neg A_m=a_m) \]
\[ \ldots \]
\[ q(A_1=a_1 \land \neg A_2=a_2 \land \ldots \land \neg A_m=a_m) \]
\[ q(\neg A_1=a_1 \land A_2=a_2 \land \ldots \land A_m=a_m) \]
\[ \ldots \]
\[ q(\neg A_1=a_1 \land \neg A_2=a_2 \land \ldots \land \neg A_m=a_m) \]

⇒ An m-order statistics is allowed iff all the \(2^m\) implied query sets obey the query-set size control.
Query-Set Overlap Control

- Another way to improve query-size control is to act on the number of common records.
- Precisely, the Query-Set Overlap Control technique permits a requested statistic \( q(C) \) only if
  \[ |X(C) \cap X(D)| \leq \alpha, \, \alpha > 0; \]
  \[ \alpha \text{ is called the overlap threshold.} \]

Perturbation-based Techniques

-