

### III. PET for Protecting Usee Identities

#### 10. Statistical Database Security

##### 10.1 Statistical Database:

collection of  $N$  records (corresponding to  $N$  individuals), where each record contains  $M$  fields

| <u>Record No.</u> | <u><math>A_1</math>.....</u> | <u><math>A_j</math>.....</u> | <u><math>A_M</math></u> |
|-------------------|------------------------------|------------------------------|-------------------------|
| 1                 | $x_{11}$ .....               | $x_{1j}$ .....               | $x_{1M}$                |
| .                 | .                            | .                            | .                       |
| .                 | .                            | .                            | .                       |
| $i$               | $x_{i1}$ .....               | $x_{ij}$ .....               | $x_{iM}$                |
| .                 | .                            | .                            | .                       |
| .                 | .                            | .                            | .                       |
| $N$               | $x_{N1}$ .....               | $x_{nj}$ .....               | $x_{NM}$                |

Attribute  $A_j$  has  $n_j = |A_j|$  possible values  $x_{j1}, \dots, x_{jn_j}$ .

$x_{ij}$ : value of attribute  $A_j$  for record (individual)  $i$ .

$1 \leq i \leq N, 1 \leq j \leq M$

Values can be numerical or non-numerical.

## Example:

| <b>Record No.</b> | <b>Name</b> | <b>Sex</b> | <b>Age</b> | <b>Major</b> | <b>GP</b> |
|-------------------|-------------|------------|------------|--------------|-----------|
| 1                 | Müller      | m          | 20         | CS           | 2         |
| 2                 | Maier       | f          | 18         | CS           | 4         |
| 3                 | Schulz      | m          | 21         | Math         | 3         |
| 4                 | Hack        | m          | 21         | Math         | 2         |
| 5                 | Baier       | f          | 20         | Math         | 1         |
| 6                 | Fischer     | m          | 21         | Math         | 2         |
| 7                 | Kunz        | f          | 20         | Math         | 1         |
| 8                 | Schmidt     | f          | 21         | CS           | 2         |
| 9                 | Kohn        | m          | 19         | CS           | 2         |
| 10                | Sveniek     | m          | 18         | CS           | 2         |
| 11                | Otto        | f          | 19         | CS           | 4         |
| 12                | Mocker      | f          | 19         | Math         | 4         |
| 13                | Andre       | m          | 23         | CS           | 4         |
| 14                | Frank       | m          | 22         | Math         | 4         |

### Statistical queries:

$q(C,U)$  (or simply:  $q(C)$ )

$q$ : statistical function

$C$ : characteristic formula, logical formula over the values of attributes using the operators OR, AND, NOT

$U$ : subset of attributes

### Example:

$q$  = correlation coefficient

$U$  = (weight, blood pressure)

$C$  = (Sex = f) AND ((Age = 32) OR (Age = 33))

"C" also denotes the query set =  
set of records whose values match a characteristic formula C

$|C|$  : query set seize

$C = ALL$  : formula whose query set is the entire database  
(  $|ALL| = N$  )

## Example of statistics:

### Counts (frequencies):

$$\text{COUNT}(C) = |(C)|$$

### Relative frequencies:

$$\begin{aligned} \text{RFR}(C) &= \text{COUNT}(C) / N \\ &= |(C)| / N \end{aligned}$$

### Sums:

$$\text{SUM}(C, X_j) = \sum_{i \in C} x_{ij}$$

### Average:

$$\text{AVG}(C, X_j) = \text{SUM}(C, X_j) / |(C)|$$

### Variance:

$$\text{VAR}(C, X_j) = (1 / |Z(C)|) * \sum_{i \in C} x_{ij}^2 - \text{AVG}(C, X_j)^2$$

Counts, sums are additive statistics:

$q(C)$  is additive, if:

$C = C_1 \cup \dots \cup C_k$ , and  $C_i$  all disjoint  $\Rightarrow$

$$q(C) = q(C_1) + q(C_2) + \dots + q(C_k).$$

## 10.2 Small and Large Query Set Attacks

Suppose: attacker knows individual  $i$  represented in the database who satisfies characteristic formula  $C$ .

If  $\text{COUNT}(C) = 1 \Rightarrow i$  is uniquely identified with  $C$

If  $A_i$  is numeric  $\Rightarrow \text{SUM}(C, A_i)$  reveals the value of  $A_i$  for individual  $i$

Otherwise: Attacker can learn whether  $i$  has an additional characteristic  $D$ :

$$\text{COUNT}(C \text{ AND } D) = \begin{cases} 1: i \text{ has } D \\ 0: i \text{ does not have } D \end{cases}$$

Attack may work even if  $i$  cannot be uniquely identified:

Suppose it is known that  $i$  satisfies  $C$  and  $\text{COUNT}(C) > 1$ :

If  $\text{COUNT}(C \text{ AND } D) = \text{COUNT}(C)$

$\Rightarrow i$  must also have characteristic  $D$ .

It is not sufficient to restrict only (sensitive) statistics with small query sets:

If  $\text{COUNT}(C) = 1 \Rightarrow \text{COUNT}(\text{NOT } C) = N-1$

IF  $X_i$  numeric:

$\text{SUM}(C, A_i) = \text{SUM}(\text{ALL}, A_i) - \text{SUM}(\text{NOT } C, A_i)$

Otherwise:

$\text{COUNT}(\text{NOT}(C \text{ AND } D)) = \begin{cases} N: i \text{ does not have } D \\ N-1: i \text{ has } D \end{cases}$

**Protection: Query Set Seize Control:**

A statistic  $q(C)$  is permitted only if  
 $n \leq |(C)| \leq N-n$  for parameter  $n \geq 2$

$q(\text{ALL})$  can be computed from:

$q(\text{All}) = q(C) + q(\text{NOT } C)$   
for  $C$  such that  $n \leq |(C)| \leq N-n$

However: Tracker attacks can still compromise security !

## 10.3 Individual Tracker

Suppose:  $q(C)$  is rejected, because  $|C| = 1$

$C = C1 \text{ AND } C2$

$$n \leq |C1| \leq N - n$$

$$n \leq |C1 \text{ AND NOT } C2| \leq N - n$$

Individual Tracker:  $\{ C1, C1 \text{ AND NOT } C2 \}$

**Individual Tracker Compromise:** ( $q$  is additive)

$$q(C) = q(C1 \text{ AND } C2) = q(C1) - q(C1 \text{ AND NOT } C2)$$

$$q(C \text{ AND } D) = q((C1 \text{ AND NOT } C2) \text{ OR } (C1 \text{ AND } D)) - q(C1 \text{ AND NOT } C2)$$

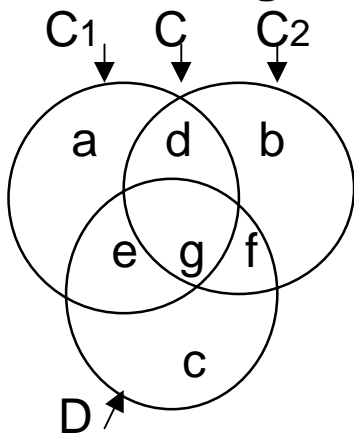
**Example:**

$\text{SUM}((\text{Job} = \text{Programmer}) \text{ AND } (\text{Sex} = \text{f}), \text{Salary})$

=  $\text{SUM}(\text{Job} = \text{Programmer}, \text{Salary})$

-  $\text{SUM}((\text{Job} = \text{Programmer}) \text{ AND NOT } (\text{Sex} = \text{f}), \text{Salary})$

**Venn-Diagram:**



$$q(C1) = a + e + d + g = (a + e) + (d + g) \\ = q(C1 \text{ AND NOT } C2) + q(C)$$

$q((C1 \text{ AND NOT } C2) \text{ OR } (C1 \text{ AND } D))$

$$= a + e + g = (a + e) + g$$

$$= q(C1 \text{ AND NOT } C2) + q(C \text{ AND } D)$$

## 10.4 General Tracker:

Characteristic Formula T such that  
 $2n \leq |(T)| \leq N - 2n$  ,  $n \leq N/4$

### General Tracker Compromise:

$$q(\text{ALL}) = q(T) + q(\text{not } T)$$

If  $|(C)| < n$ :

$$q(C) = q(C \text{ or } T) + q(C \text{ or not } T) - q(\text{ALL})$$

If  $|(C)| > N - n$ :

$$q(C) = q(\text{ALL}) - q(\text{not } C) =$$

$$2 * q(\text{ALL}) - q(\text{not } C \text{ or } T) - q(\text{not } C \text{ or not } T)$$

**Example:**  $T = (\text{Sex} = m)$

SUM ((Job = Programmer ) AND (Sex = f), Salary)

= SUM (((Job = Programmer) AND (Sex = f))

OR (Sex = m), Salary)

+ SUM(((Job = Programmer) AND (Sex = f))

OR NOT (Sex = m), Salary)

- [SUM (Sex = m, Salary) + SUM (Sex = f, Salary)]

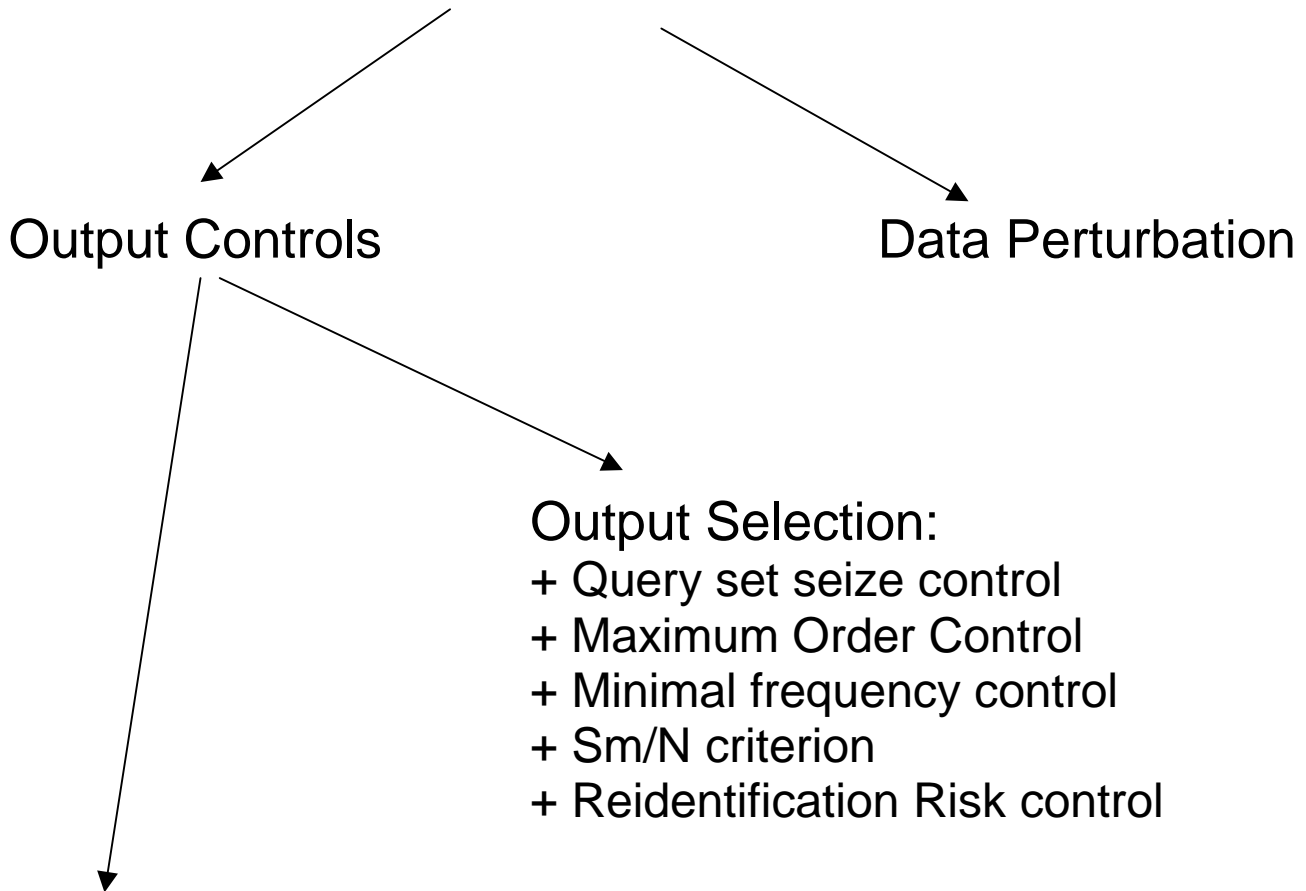
### Venn-Diagram:

|       |   |       |
|-------|---|-------|
|       | T | not T |
| C     | a | b     |
| not C | c | d     |

$$q(\text{All}) = a + c + b + d = q(T) + q(\text{not } T)$$

$$\begin{aligned} q(C \text{ or } T) + q(C \text{ or not } T) &= (a+b+c) + (a+b+d) \\ &= (a+b) + (a+b+c+d) \\ &= q(C) + q(\text{All}) \end{aligned}$$

## 10.5 Security Mechanisms for Statistical Databases:



### Output Modification:

- + rounding
- + adding random numbers
- + random sample queries

### Classification according to:

- information loss
- security
- costs|

## 10.5.1 Output Selection Controls:

### **QuerySet Seize Control:**

$q(C)$  is permitted if and only if  $n \leq |(C)| \leq N - n$ ,  $n \geq 2$

### **Maximum Order Control:**

$q(C, U)$  is permitted if the number of attributes in  $C$  and  $U \leq K$  for parameter  $K$ .

### **Minimal Frequency Control:**

$q(C)$  with the characteristic attributes  $A_1 \dots A_m$  is rejected, if

$$\prod_{i=1}^m \min(r(A_i)) \leq 1/K$$

$\min(r(A_i))$ : smallest relative frequency occurring among the values in the domains of  $A_i$ .

$K$ : Parameter

### **Sm/N -Criterion:**

$q(C)$  with the characteristic attributes  $A_1 \dots A_m$  is rejected, if

$$Sm/N = \prod_{i=1}^m |A_i| > K \quad \text{for parameter } K.$$

### **Reidentification Risk Control:**

$q(C)$  is rejected, if  $RR(q(C)) > K$  for parameter  $K$ .

where  $RR(q(C)) = \begin{cases} 2 H(A_1, \dots, A_m)/N & \text{if } 2 H(A_1, \dots, A_m) \leq N \\ 1 & \text{else.} \end{cases}$

## **10.5.2 Output Modification Controls:**

### Advantages:

- more sensitive statistics can be released
- benefits from difference between absolute and relative errors

### Desirable properties:

- systematic error (bias) should be low  
systematic error =  $q - E(r(q))$   
q: true value of statistic  $q(C)$   
 $E(r(q))$  : expected value of modified value
- consistencies of values (lack of contradictions, paradoxes)
- no error removal by "averaging"

## Systematic Rounding:

value  $q$  of  $q(C)$  is rounded up or down to the nearest multiple of some base  $b$

### **Example:**

Base = 10

|                |   |   |   |   |   |    |    |    |    |    |    |
|----------------|---|---|---|---|---|----|----|----|----|----|----|
| true value:    | 0 | 1 | 2 | 3 | 4 | 5  | 6  | 7  | 8  | 9  | 10 |
| rounded value: | 0 | 0 | 0 | 0 | 0 | 10 | 10 | 10 | 10 | 10 | 10 |

### **Or output of Intervals:**

|             |       |     |       |        |     |   |   |   |   |   |    |
|-------------|-------|-----|-------|--------|-----|---|---|---|---|---|----|
| true value: | 0     | 1   | 2     | 3      | 4   | 5 | 6 | 7 | 8 | 9 | 10 |
|             | :     |     |       |        |     |   |   | : |   | : |    |
| intervals:  | [0,4] | ... | [0,4] | [5,10] | ... |   |   |   |   |   |    |

### Advantages:

- no error removal by averaging
- average systematic error is almost zero
- intervals in frequency tables are consistent

### Disadvantages:

- attacks based on comparison of interval estimates
- tracker attacks still possible

## Example of disclosure under systematic rounding:

Let  $C_1 \dots C_m$  be disjoint query sets, and let

$$C_{m+1} = C_1 \cup \dots \cup C_m.$$

$$\Rightarrow q_{m+1} = q_1 + \dots + q_m \text{ für } q_i = \text{SUM}(C_i, A_j)$$

Let  $[L_i, U_i]$ : be the interval estimates for each rounded  $r(q_i)$

$$L = \sum_{i=1}^m L_i, \quad U = \sum_{i=1}^m U_i$$

$$1.) \quad \boxed{\text{If } U = L_{m+1} \Rightarrow q_i = U_i \text{ (} 1 \leq i \leq m \text{), } q_{m+1} = L_{m+1}}$$

Example:

|           | $r(q_i)$ | $L_i$     | $U_i$     |                          |
|-----------|----------|-----------|-----------|--------------------------|
| $q_1$     | 15       | 13        | 17        | $\Rightarrow$            |
| $q_2$     | 10       | 8         | 12        |                          |
| $q_3$     | 15       | 13        | 17        |                          |
| $q_4$     | 20       | <u>18</u> | <u>22</u> |                          |
|           |          | L: 52     | U: 68     | $q_{m+1} = L_{m+1} = 68$ |
| $q_{m+1}$ | 70       | 68        | 72        |                          |

$$2.) \quad \boxed{\text{If } L = U_{m+1} \Rightarrow q_i = L_i, \quad q_{m+1} = U_{m+1}}$$

## Random rounding

Rounds a statistic  $q$  up or down to the following rule:

$$r(q) = \begin{cases} q & \text{if } d = 0 \\ q-d & \text{with the probability } 1-p \text{ (round down)} \\ q+(b-d) & \text{with the probability } p \text{ (round up)} \end{cases}$$

$b = \text{basis}$

$d = q \bmod b$

$p = d / b$

$\Rightarrow$

$$\begin{aligned} E(r(q)) &= (1-p) \cdot (q-d) + (q+b-d) \cdot p \\ &= q-d + b \cdot p = q-d + b(d/b) = q \end{aligned}$$

$\Rightarrow$  Systematic error (bias) = 0

### Disadvantages:

- attacks based on comparison of interval estimates / tracker attacks
- error removal by averaging
- consistency problems

## Addition of Random Numbers:

$r(q)$  is returned instead of the true value  $q$ , with:

$$r(q) = q + Z$$

$Z$  : random variable with  $E(Z) = 0$ , variance  $\sigma_Z^2 > 0$

Systematic error (bias) =

$$\begin{aligned} q - E(r(q)) &= q - E(q + Z) \\ &= q - E(q) - E(Z) \\ &= -E(Z) \\ &= 0 \end{aligned}$$

## Disadvantage:

error removal by averaging (number of necessary queries increases as variance increases)

## Random Sample Queries:

The result of  $q(C)$  is computed on a random sample  $C^*$  of the query set  $C$

### Advantages:

- low sampling errors due to large samples ( 80 - 90%)
- attackers cannot control composition of query sets  
→ tracker attacks not possible

Selection function  $g_p(C,i)$  is applied to each record  $i \in C$ .

Record  $i$  is chosen for  $C^*$  with sampling probability  $p$ ,

$$C^* = \{ i \in C \mid g_p(C, i) = 1 \}.$$

### Example of implementation of selection function:

$$p = 1 - 1/2^k \quad (p \geq 1/2).$$

$r(i)$ : function that maps  $i$ th record into random sequence of  $m \geq k$  bits.

$s(C)$ : function that maps  $C$  into random sequence of length  $m$  over the alphabet  $\{0, 1, *\}$ , with  $k$  bits and  $m-k$  asterisks (don't care)

$$1 \text{ if } r(i) \text{ does not match } s(C)$$

$$g_p(C,i) = \{$$

$$0 \text{ if } r(i) \text{ matches } s(C)$$

### Example:

Let  $p = 7/8$ ,  $m = 8$

$s(C) = *10*1***$

If  $r(i) = 11011000 \rightarrow$  record  $i$  is excluded from  $C^*$ .

Expected size of  $C^*$ :  $7/8 |C|$

### Advantages:

- Expected value of sampled relative frequency:  
 $RFR^*(C) = |C^*| / pN = p |C| / pN = |C| / N$   
 $\rightarrow$  sampled relative frequency is an unbiased estimator of the true relative frequency
- sampled average has negligible bias
- no direct error removal by averaging

### However:

Error removal by posing different but "equivalent" queries (e.g., use of different formulas to specify the same query set).

### Example:

$q((\text{Sex} = m) \text{ and } (\text{Age} = 32))$  could be estimated by:

$q^*((\text{Sex} = m) \text{ and } (\text{Age} = 32))$

$q^*(\text{NOT}(\text{Sex} = f) \text{ AND } (\text{Age} = 32))$

$q^*(( (\text{Sex} = m) \text{ AND } (\text{Major} = \text{Bio}) \text{ AND } (\text{Age} = 32))$   
 $(\text{OR } (\text{Sex} = m) \text{ AND NOT}(\text{Major} = \text{Bio}) \text{ AND } (\text{Age} = 32)))$

->

Selection function should be a function of the query set C rather than characteristic formula C

Still:

Averaging that uses disjoint subsets of query sets is possible !

**Example:**

$q((\text{Sex} = m) \text{ and } (\text{Age} = 32))$  could be estimated by:

$$q^*( (\text{Sex} = m) \text{ AND } (\text{Age} = 32) \text{ AND } (\text{Major} = \text{Bio})) + \\ q^*( (\text{Sex} = m) \text{ AND } (\text{Age} = 32) \text{ AND NOT } (\text{Major} = \text{Bio}))$$

$$q^*( (\text{Sex} = m) \text{ AND } (\text{Age} = 32) \text{ AND } (\text{Major} = \text{CS})) + \\ q^*( (\text{Sex} = m) \text{ AND } (\text{Age} = 32) \text{ AND NOT } (\text{Major} = \text{CS}))$$

$$q^*( (\text{Sex} = m) \text{ AND } (\text{Age} = 32) \text{ AND } (\text{Major} = \text{Math})) + \\ q^*((\text{Sex} = m) \text{ AND } (\text{Age} = 32) \text{ AND NOT}(\text{Major} = \text{Math}))$$

**However:**

High number of queries required for accurate estimates ->  
detection of systematic attacks by intrusion detection systems