

SmallDB algorithm

Assumption : 1. We know the query set in advance. It is f_1, \dots, f_m . (So the results should be $f_1(T), \dots, f_m(T)$)
 (we actually do not need this assumption)

2. Function f_i can be written in the form

$$f_i(T) = \sum_{j=1}^{|T|} f_i(T_j)$$

T_j → record j of Table T
 $|T|$ → size of T (#records)

$f_i(T_j)$ must be between 0 and 1

Example

Name	Chosen Party	Age < 40
Alice	Melan-pun	30
Bob	Melan-pun	40
Charles	Gyu-don	20
Doe	Gyu-don	20

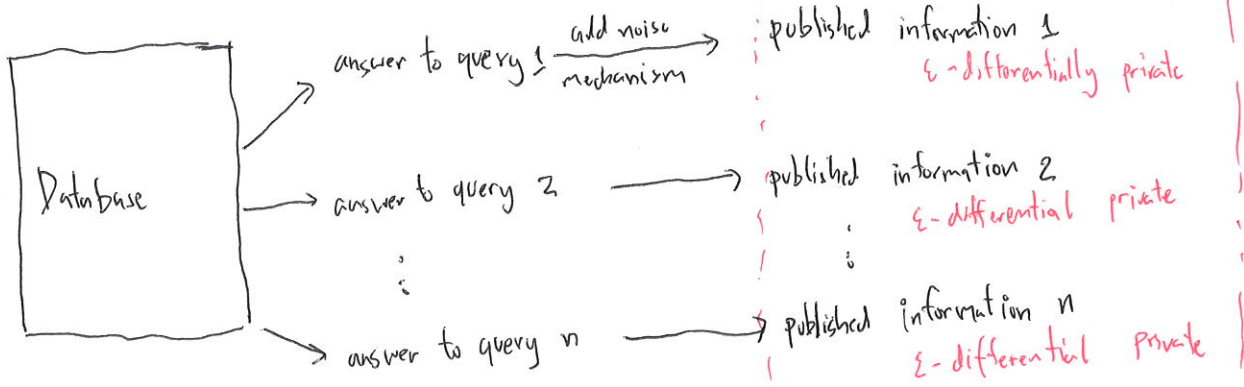
How many persons "choose Gyu-don" and "age smaller than 30"?

$$f_i(N, P, A) = \begin{cases} 1 & \text{when } P = \text{Gyu-don and } A < 30 \\ 0 & \text{otherwise} \end{cases}$$

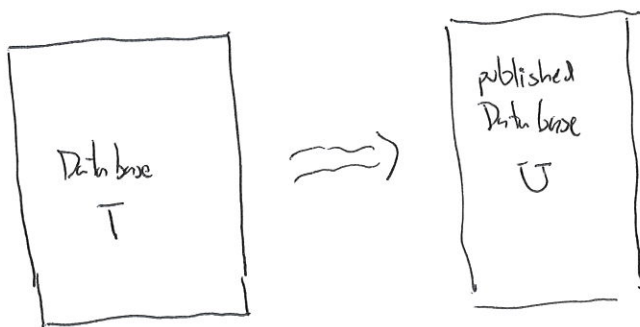
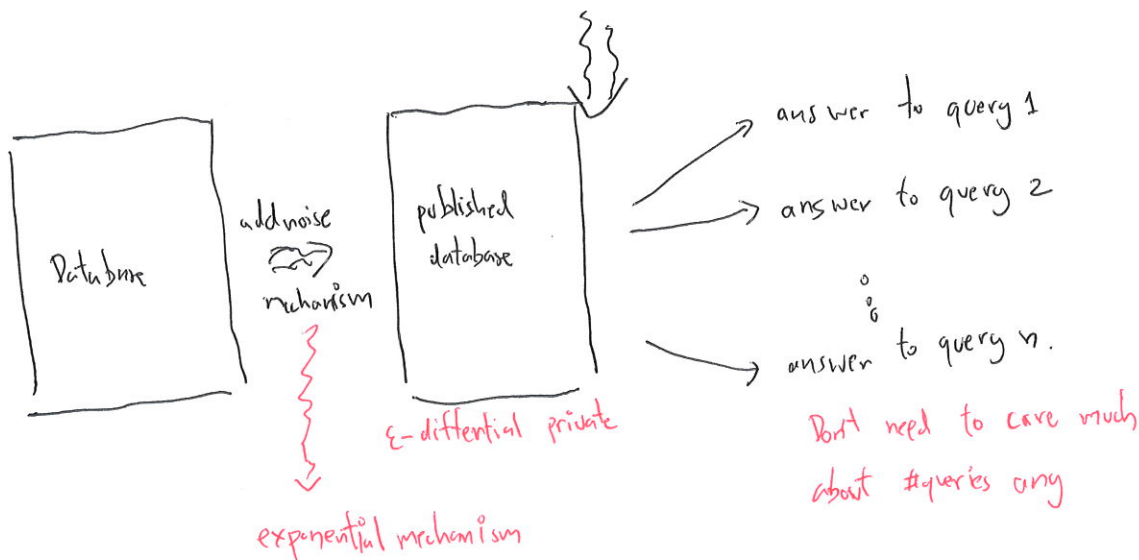
$$f_i(T) = \frac{\# \text{ persons who "choose Gyu-don" and "age smaller than 30"}}{|T|} \rightarrow \# \text{ records}$$

$$\# \text{ persons} = f_i(T) \cdot |T|$$

Until now, ...



can be risky when n is large \leftarrow ϵn -differentially private



Utility $z(T, U)$ \rightarrow larger large if U is similar to T
Small otherwise

$$h_T(U) = \exp\left(\frac{\epsilon \cdot \text{Utility}(T, U)}{2 \Delta \text{Utility}}\right)$$

$$Pr_T[\text{Out}(T) = U] = \frac{h_T(U)}{\sum_{\text{all possible databases } U'} h_T(U')}$$

\rightarrow Takes a lot of time to compute.

too many possible databases

at least 2^n possible databases for n persons!

Average Age?

$$f_p(\text{Age}, A) = \frac{A}{200}$$

$$f_i(T) = \frac{\sum_i A_i / 200}{\|T\|} = \frac{\sum_i A_i}{\|T\|} \cdot \frac{1}{200}$$

$$\text{Average Weight} = f_i(T) \cdot 200$$

Small Database (Small DP) Algorithm

Use Exponential Mechanism

$\mathcal{P} := \{ \text{Database with } \log \frac{m}{\epsilon} \}$
possible report $\epsilon \rightarrow$ parameter to define later

$$\text{Utility}(T, U) := - \max_i |f_i(T) - f_i(U)|$$

\rightarrow Different in query results.
 \rightarrow take the maximum difference
 \rightarrow the larger difference, the smaller utility.

Choose $U \in \mathcal{P}$ using exponential mechanism

Privacy: ϵ -differentially private because of the exponential mechanism.

Chernoff Bound Let X_1, \dots, X_n be independent random variables. Let

$$S = \frac{1}{n} \sum_{i=1}^n X_i$$

average values and μ is expected value of S .
 $\mu_1 + \dots + \mu_n \rightarrow$ expected value of $\sum X_i$.

We have

$$\Pr[S > \mu + \epsilon] \leq e^{-2m\epsilon^2}$$

larger $m =$ smaller probability
larger $\epsilon =$ smaller probability

$$\Pr[S < \mu - \epsilon] \leq e^{-2m\epsilon^2}$$



$$\Pr[|S - \mu| > \epsilon] \leq 2 \cdot e^{-2m\epsilon^2}$$

Theorem: For any table T and function f_i , there exists table U with $\frac{\log(m)}{\epsilon^2}$ records

such that

$$\max_i |f_i(T) - f_i(U)| \leq \epsilon \rightarrow \text{larger } \epsilon = \text{smaller } \# \text{ records}$$

Proof:

We will construct U as follows:

Each record U_j is randomly picked from Table T , i.e.

$$\Pr[U_j = T_i] = \frac{1}{|U|} \text{ for all } i, j$$

$f_i(U_j)$ has expected value $\sum_j f_i(T_j) / n = f_i(T)$.

Let $X_j = f_i(U_j) / |U|$. The expected value of X_j is $f_i(T) / |U|$

$$S = f_i(U) = \sum_j \frac{f_i(U_j)}{|U|} = \sum_j X_j$$

Expected value of $S = f_i(T)$.

By Chernoff Bound

$$\Pr \left\{ |f_i(U) - f_i(T)| > \epsilon \right\} \leq 2e^{-\frac{|U| \epsilon^2}{2m}} \text{ for any } i$$

$$\Pr \left\{ \text{There exist } i \text{ such that } |f_i(U) - f_i(T)| > \epsilon \right\} \leq 2 \cdot m \cdot e^{-\frac{|U| \epsilon^2}{2m}}$$

$$= 2 \cdot m \cdot \exp\left(-2 \cdot \frac{\log m}{\epsilon^2} \cdot \epsilon^2\right)$$

$$= 2 \cdot m \cdot \left[\exp(\log m)\right]^{-2}$$

$$= 2 \cdot m \cdot \frac{1}{m^2} \ll 1 \text{ when } m \gg 2$$

~~$\Pr \left\{ \text{There exist } i \text{ such that}$~~

$$\Pr \left\{ |f_i(U) - f_i(T)| \leq \epsilon \text{ for all } i \right\} > 0$$

It is possible to have such U

□