Sanitization of Databases

- Real Database (RDB)
  - Add noise, delete names, etc.
  - Sanitized Database (SDB)
  - Health records
  - Census data
  - Protect privacy
  - Provide useful information (utility)
Recall Dalenius’ Desideratum

A 1977 paper of Dalenius [6] articulated a desideratum that foreshadows for databases the notion of semantic security defined five years later by Goldwasser and Micali for cryptosystems [15]: access to a statistical database should not enable one to learn anything about an individual that could not be learned without access.²
**Impossibility Result**

- **Privacy**: For some definition of “privacy breach,”
  \[ \forall \text{ distribution on databases}, \forall \text{ adversaries } A, \exists A' \]
  such that \[ \Pr(A(\text{San})=\text{breach}) - \Pr(A'(\cdot)=\text{breach}) \leq \varepsilon \]

  - Result: For reasonable “breach”, if San(DB) contains information about DB, then some adversary breaks this definition

- **Example**
  - Terry Gross is two inches shorter than the average Lithuanian woman
  - DB allows computing average height of a Lithuanian woman
  - This DB breaks Terry Gross’s privacy according to this definition… even if her record is not in the database!

[Dwork, Naor 2006]
(Very Informal) Proof Sketch

- Suppose DB is uniformly random
  - Entropy $I( DB ; \text{San}(DB) ) > 0$
- “Breach” is predicting a predicate $g(DB)$
- Adversary knows $r, H(r ; \text{San}(DB)) \oplus g(DB)$
  - $H$ is a suitable hash function, $r=H(DB)$
- By itself, does not leak anything about DB
- Together with $\text{San}(DB)$, reveals $g(DB)$
Relaxed Semantic Security Databases?

Example

- Terry Gross is two inches shorter than the average Lithuanian woman
- DB allows computing average height of a Lithuanian woman
- This DB breaks Terry Gross’s privacy according to this definition… even if her record is not in the database!

- Suggests new notion of privacy
  - Risk incurred by joining database
Differential Privacy (informal)

Output is similar whether any single individual's record is included in the database or not. C is no worse off because her record is included in the computation.

If there is already some risk of revealing a secret of C by combining auxiliary information and something learned from DB, then that risk is still there but not increased by C’s participation in the database.

C is no worse off because her record is included in the computation.
Differential Privacy [Dwork et al 2006]

Randomized sanitization function \( \kappa \) has \( \varepsilon \)-differential privacy if for all data sets \( D_1 \) and \( D_2 \) differing by at most one element and all subsets \( S \) of the range of \( \kappa \),

\[
\Pr[\kappa(D_1) \in S] \leq e^\varepsilon \Pr[\kappa(D_2) \in S]
\]
Achieving Differential Privacy

- How much noise should be added?

- Intuition: \( f(D) \) can be released accurately when \( f \) is insensitive to individual entries \( x_1, \ldots, x_n \) (the more sensitive \( f \) is, higher the noise added)
Sensitivity of a function

We will achieve $\epsilon$-differential privacy by the addition of random noise whose magnitude is chosen as a function of the largest change a single participant could have on the output to the query function; we refer to this quantity as the sensitivity of the function $^{9}$.

**Definition 3.** For $f : \mathcal{D} \rightarrow \mathbb{R}^d$, the $L_1$-sensitivity of $f$ is

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1$$

(2)

for all $D_1, D_2$ differing in at most one element.

- **Examples:** $\Delta \text{count} \leq 1$, $\Delta \text{histogram} \leq 1$
- **Note:** $\Delta f$ does *not* depend on the database
Achieving Differential Privacy

The privacy mechanism, denoted $\mathcal{K}_f$ for a query function $f$, computes $f(X)$ and adds noise with a scaled symmetric exponential distribution with variance $\sigma^2$ (to be determined in Theorem 4) in each component, described by the density function

$$\Pr[\mathcal{K}_f(X) = a] \propto \exp(-\|f(X) - a\|_1/\sigma)$$

This distribution has independent coordinates, each of which is an exponentially distributed random variable. The implementation of this mechanism thus simply adds symmetric exponential noise to each coordinate of $f(X)$.

**Theorem 4.** For $f : D \rightarrow \mathbb{R}^d$, the mechanism $\mathcal{K}_f$ gives $$(\Delta f/\sigma)$$-differential privacy.
Proof of Theorem 4

Proof. Starting from (3), we apply the triangle inequality within the exponent, yielding for all possible responses $r$

$$\Pr[K_f(D_1) = r] \leq \Pr[K_f(D_2) = r] \times \exp(||f(D_1) - f(D_2)||_1/\sigma). \quad (4)$$

The second term in this product is bounded by $\exp(\Delta f/\sigma)$, by the definition of $\Delta f$. Thus (1) holds for singleton sets $S = \{a\}$, and the theorem follows by a union bound.
Sensitivity with Laplace Noise

**Theorem**

If $A(x) = f(x) + \text{Lap}\left(\frac{\text{GS}_f}{\varepsilon}\right)$ then $A$ is $\varepsilon$-indistinguishable.

Laplace distribution $\text{Lap}(\lambda)$ has density $h(y) \propto e^{-\frac{\|y\|_1}{\lambda}}$

**Proof idea:**

- $A(x)$: blue curve
- $A(x')$: red curve
- $\delta = f(x) - f(x') \leq \text{GS}_f$