Differential Privacy

Based on slides from: Ken Birman

The Pitfalls of Anonymization

- "Anonymous" datasets (with PII removed) have been connected to specific individuals.
  - AOL search histories.
  - Netflix prize.
  - Human genetic datasets.
- All of these cases involved *auxiliary information*.
- A technology is required to give incentives (or at least to remove disincentives) to contribute to these datasets.
  - "First, do no harm."

Statistical databases

- The purpose of a statistical database is to inform.
- These databases will be exposed to users.

The goal is to reveal information, so the right definition of privacy is not obvious.

A Definition Inspired by Semantic Security (Encryption)

An attempt at a workable definition of privacy in this setting:

"Anything that can be learned about an individual from the statistical database should be learnable without access to the database."

Naturally relates to **semantic security** in cryptosystems.

Paradox

- Violet Mcguire's salary is $15K higher than the average Canadian woman's salary `DB allows computing average salary of Canadian women `This DB breaks Violet Mcquire's privacy according to this definition… **even if her record is not in the database**!
- This has been extended to a general proof of the inadequacy of this definition.
- This definition fails as it punishes the database for revealing any information at all, which is the purpose of a statistical database.

How to Define Privacy?

* Hard to determine if something satisfies privacy
* Easy to determine if something is **NOT** private

Blatant non-privacy

A system is blatantly non-private if an adversary can construct a replica database that matches the real database in 99% of its entries.

The adversary gets at most 1% wrong.

Simple Example

A "Database" is simply a 0/1 column vector $d = d_1, \ldots, d_n$

A query corresponds to a subset $S \subseteq [n]$ of indeces

A query response is the number of 1's in the locations contained in S:

$$\sum_{i \in S} d_i$$

How much noise do we need to add to each query response in order to avoid "blatant non-privacy"?

Theorem 1: Let M be a mechanism that adds noise bounded by E to each query. Then there exists an adversary that can reconstruct the database to within $4E$ positions.

PROOF: Let d be the true database. The adversary can attack in two phases:

1. **Estimate the number of 1's in all possible sets:** Query M on all subsets $S \subseteq [n]$.

2. **Rule out "distant" database:** For every candidate database $c \in \{0,1\}^n$, if, for any $S \subseteq [n]$, $|\sum_{i \in S} c_i - M(S)| > E$, then rule out c. If c is not ruled out, then output c and halt.

$M(S)$ never errs by more than E, so the real database will not be ruled out and thus this algorithm must return some database. Call its output c.

Let I_0 be the indices in which $d_i = 0$. Given the second step of the algorithm, it must be that $|M(I_0) - \sum_{i \in I_0} c_i| \leq E$. By assumption $|M(I_0) - \sum_{i \in I_0} d_i| \leq E$. It follows from the triangle inequality that c and d differ in at most $2E$ positions in I_0.

Let I_1 be the indices in which $d_i = 1$. The same argument holds, so that c and d differ in at most $2E$ positions in I_1.

Thus, c and d agree on all but at most $4E$ positions.

To avoid blatant non-privacy, we must add noise bounded by $n/400$. A bound of $n/401$ or lower is provably non-private.

Lessons

* Noise cannot be bounded (i.e. add a value $E \in [-B, B]$ to the true query response.)
* Unlimited number of queries cannot be allowed
  * Noise will grow with number of queries

Differential privacy

- It should not harm you or help you as an individual to enter or to leave the dataset.
- To ensure this property, we need a mechanism whose output is nearly unchanged by the presence or absence of a single respondent in the database.
- In constructing a formal approach, we concentrate on pairs of databases (*D, D'*) differing on only one row, with one a subset of the other and the larger database containing a single additional row.

Differential privacy

Definition 2. A randomized function K gives ϵ-differential privacy if for all data sets D and D' differing on at most one row, and all $S \subseteq Range(K)$,

$$\Pr[K(D) \in S] \leq \exp(\epsilon) \times \Pr[K(D') \in S],$$

where the probability space in each case is over the coin flips of K.

Differential privacy

- An equivalent expression of this idea is given as a ratio bounded by R:

$$\frac{Pr[K(D) \in S]}{Pr[K(D') \in S]} \leq exp(\epsilon) = R$$

- The closer R is to 1, or ε to 0, the more difficult it will be for an attacker to determine an individual's data.

- ε is a publicly known characteristic of our database. It defines the level of privacy maintained and it informs users of the amount of error to expect in the responses it yields.

Differential privacy

- An important property of this definition is that any output with zero probability is invalid for all databases.
  - An output with a probability of zero in a given database must have a probability of zero in both neighboring databases and by induction, in any other database as well.
- It immediately follows that sub-sampling fails to implement differential privacy.
  - A row cannot be present in a sub-sample if that person has previously left the dataset.

Noise properties

- We know we can add noise to query responses to disguise the true contents of the database.
- We know the level of disguise required for differential privacy.
- What distribution should we employ to generate this noise?

Simple Example

A "Database" is simply a 0/1 column vector $d = d_1, \ldots, d_n$

A query corresponds to a subset $S \subseteq [n]$ of indeces

A response is the number of 1's in the locations contained in S:

$$\sum_{i \in S} d_i$$

How much and what type of noise can we add to a query response to achieve differential privacy *for a single query*?

Go Back to Simple Example

Adding or removing an element of $\boldsymbol{d}$ can only change the answer by 1. Let x be the outcome of a query in d, (x+1) is the outcome in d'. Let z be the added noise. We need to show that:

$$e^{-\epsilon} \leq \frac{\Pr[x + z \in S]}{\Pr[(x+1) + z \in S]} \leq e^{\epsilon}$$

Assume that for every $x \backslash in [n], v \backslash in R$, the noise z comes from a distribution with PDF ψ that satisfies

$$e^{-\epsilon} \leq \frac{\psi(z = v - x)}{\psi(z = v - x - 1)} \leq e^{\epsilon}$$

Then

$$e^{-\epsilon} \leq \frac{\Pr[x + z \in S]}{\Pr[(x+1) + z \in S]} \leq e^{\epsilon}$$

Which noise distribution has this property?

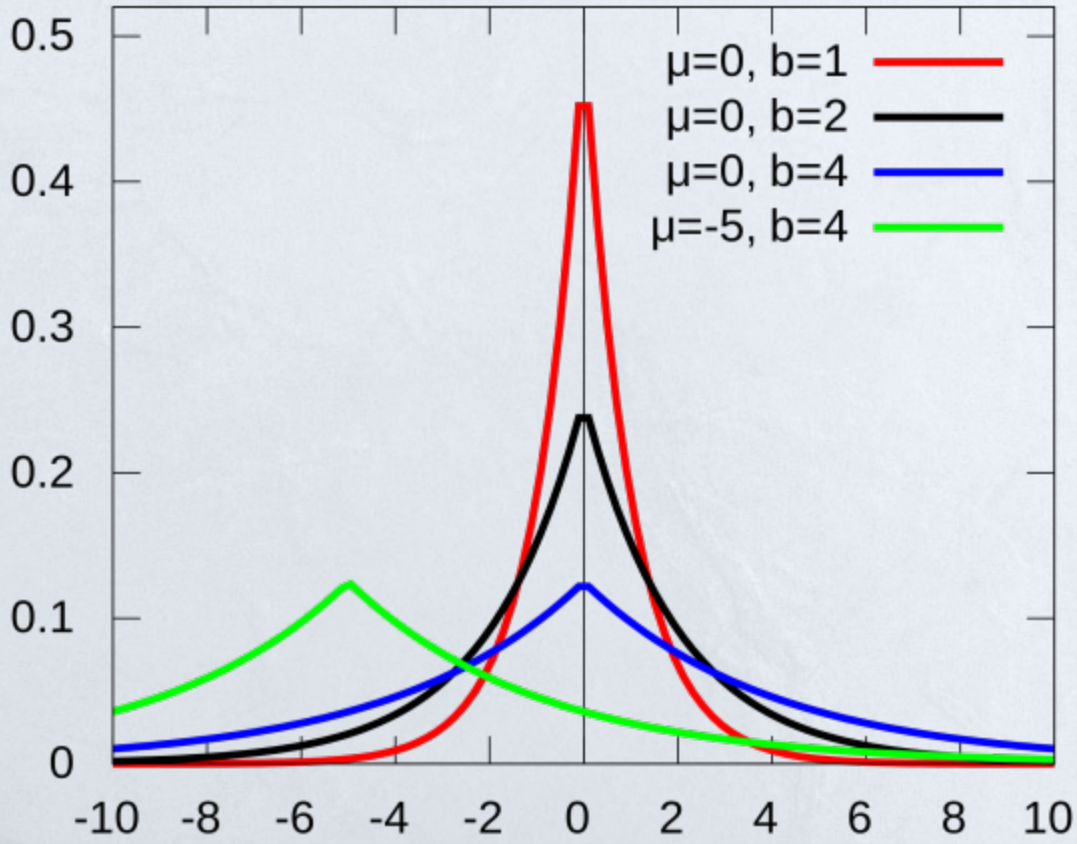
Laplace distribution

- We generate noise using the Laplace distribution.
- The Laplace distribution, denoted Lap(b), is defined with parameter b and has density function:

$$P(z|b) = \frac{1}{2b} e^{\frac{-|z|}{b}}$$

- Taking $b = 1/\varepsilon$ we have immediately that the density is proportional to $e^{-\varepsilon|z|}$.
- This distribution has its highest density at 0.
- **For any z, z' such that $|z - z'| \leq 1$, the density at z is at most e^{ε} times the density at z', satisfying the condition we outlined in the simple case.**
- The distribution is symmetric about 0.
- The distribution flattens as ε decreases. More likely to deviate from the true value.

Laplace distribution



Wikipedia, Creative Commons license

General case

- What about multiple queries? Or queries whose output value can change by more than 1 when a row is added or removed?
- To do this, we must consider the *sensitivity* of the function that will generate the response.
  - In the simple case, the sensitivity was 1.

DEFINITION 3. For $f : D \rightarrow \mathbf{R}^d$, the L_1 sensitivity of f is

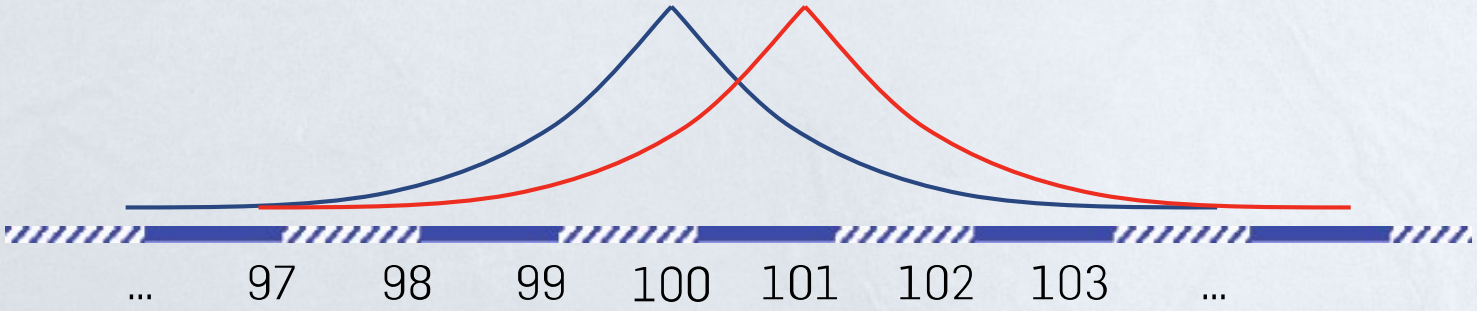
$$\Delta f = \max_{D,D'} ||f(D) - f(D')||_1$$
$$= \max_{D,D'} \Sigma_{i=1}^{d} |f(D)_i - f(D')_i|$$

for all D, D' differing in at most one row.

- The sensitivity defines the difference that the noise must hide.

Final theorem

Theorem 4. For $f : D \rightarrow \mathbf{R}^d$, the mechanism K that adds independently generated noise with distribution $Lap(\Delta f/\epsilon)$ to each of the d output terms enjoys ϵ-differential privacy.



In this figure, the distribution on the outputs, show in gray, is centered at the true answer of 100, where $\Delta f = 1$ and $\varepsilon = \ln 2$. In orange is the same distribution where the true answer is 101.

PROOF. Consider any subset $S \subseteq Range(K)$, and let D, D' be any pair of databases differing in at most one row. When the database is D, the probability density at any $r \in S$ is proportional to $e^{-||f(D)-r||_1(\epsilon/\Delta f)}$. Similarly, when the database is D', the probability density at any $r \in Range(K)$ is proportional to $e^{-||f(D')-r||_1(\epsilon/\Delta f)}$.

$$\begin{aligned}
\frac{e^{-||f(D)-r||(\epsilon/\Delta f)}}{e^{-||f(D')-r||(\epsilon/\Delta f)}} &= \frac{e^{||f(D')-r||(\epsilon/\Delta f)}}{e^{||f(D)-r||(\epsilon/\Delta f)}} \\
&= \frac{e^{||f(D')-r||(\epsilon/\Delta f)}}{e^{||f(D)-r||(\epsilon/\Delta f)}} \\
&= e^{(||f(D')-r||-||f(D)-r||)(\epsilon/\Delta f)} \\
&\leq e^{(||f(D')-f(D)||)(\epsilon/\Delta f)}
\end{aligned}$$

where the inequality follows from the triangle inequality. By the definition of sensitivity, $||f(D') - f(D)||_1 \leq \Delta f$, and so the ratio is bounded by e^{ϵ}. Integrating over S yields ϵ-differential privacy.