Today: Differential Privacy

Given:

a dataset with sensitive infromation, such as:

health records, census data, social network activity,...

Goal:

Use this data in desirable ways while protecting the privacy of individuals.

Example: Data is used for predicting likelyhood of disease, for predicting recidivism, and more.

Our private data is used everyday by machile learnning algorithms!

This raises many issues: privacy of our data, fairness, and more.

A great book recommendation: "The Ethical Algorithm: The Science of Socially Aware Algorithm Design" by Michael Kearns and Aaron Roth.

Approach 1: Anonymize the data

Remove explicit identifiers such as name, address and telephone number.

Doesn't work!

Example 1: Re-identification by linking (Sweeney 1997)

The National Association of Health Data Organizations (NAHDO) reported that 37 states in the USA have legislative mandates to collect hospital level data.

The data that it recommends to collect is (and more):

Ethnicity.

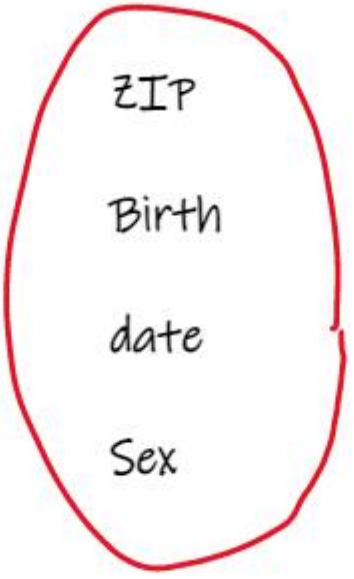
Visit date

Diagnosis

Procedure

Medication

Total charge



In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees.

GIC collected patient specific data along the lines of the those shown above for about 135,000 state employees and their families.

Because the data were believed to be anonymous, GIC gave a copy of the data to researchers and sold a copy to industry.

Sweeney 1997: "For twenty dollars I purchased the voter registration list for Cambridge Massachusetts and received the information of each voter, including the name, address, ZIP code, birth date, and gender.

This information can be linked using ZIP code, birth date and gender to the medical information, thereby linking diagnosis, procedures, and medications to particularly name individuals.

For example, William Weld was governor of Massachusetts at that time and his medical records were in the GIC data. Governor Weld lived in Cambridge Massachusetts. According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code.

The example above provides a demonstration of re-identification by directly linking (or "matching") on shared attributes.

Example 2: Netfix Prize competition

Netflix is a data-driven and statistically-minded company. Their recommendation algorithm is tuned to optimize user engagement. Between 2006 and 2009, they hosted a contest, challenging researchers to improve their recommendation engine. The grand prize was a highly-publicized $1,000,000.

In order to help teams design their strategies, Netflix provided a training dataset of user data. Each datapoint consisted of an (anonymized) user ID, movie ID, rating, and date.

Netflix assured users that the data was appropriately anonymized to protect individual privacy. Indeed, the Video Privacy Protection Act of 1988 requires them to do this.

Unfortunately, Narayanan and Shmatikov (2008) demonstrated that this naive form of anymization was insufficient to preserve user privacy. They took the dataset provided by Netflix, and cross-referenced it with public information from the online movie database IMDb, which contains hundreds of millions of movie reviews. In particular, they tried to match users between the two datasets by finding users who gave similar ratings to a movie at similar times.

Conclusion: Anonymization does not work due to external information!

Question: What if we release only statistics (instead of anonymized dataset)?

Example:

Suppose we have a dataset $(x_1,\ldots,x_n)$ where person i has a sensitive bit x_i in $\{0,1\}$.

Suppose the adversay gets only the subset sums of x_i.

Of course, if the adv gets to choose which subset sums he receives then he can attack by choosing the subset $S=\{i\}$.

Even if we insist that S is at least size n/10 the adversary can attack by asking for the subset sum of a random set S which does not contain i and a subset sum of S including i.

What if these subsets are random? (In particular the adv cannot choose them.)

Claim: If the number of sums released is n, then with high prob. an adversary can reconstruct entire dataset.

Lesson learned: Too many statistics reveal individual information!

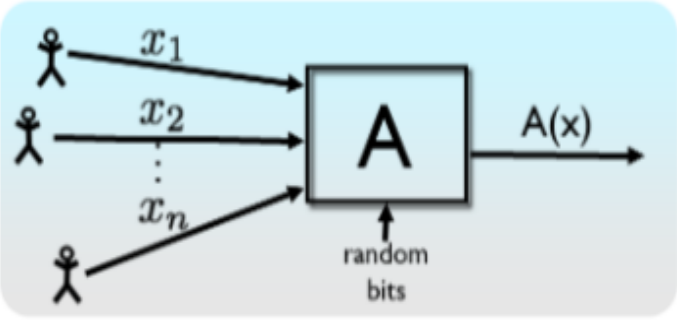
Problem: Even a few statistics can reveal senstitive information!

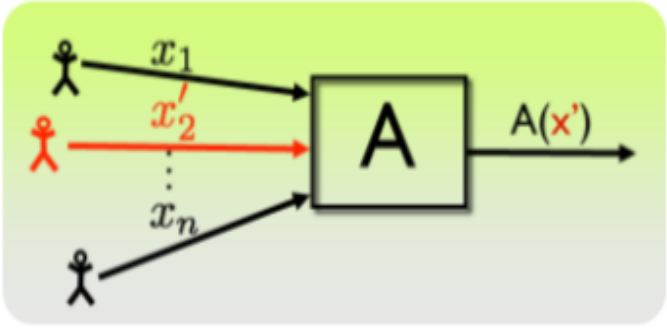
Example: Release the average salary before and after the resignation of an individual

Too many "too accurate" statistics reveal individual information

Differential Privacy: Defined by Dwork, McSherry, Nissim, and Smith in 2006.

When is an algorithm private?





Alg A is private if for every neighboring datasets (that differ in a single row)

A(x) and A(x') are close.

Definition: A is ϵ-differentially private if,

for all neighbors x, x',

for all subsets S of outputs

$1+\epsilon$

$$\Pr[A(x) \in S] \leq e^{\epsilon} \Pr[A(x') \in S]$$

ϵ is a leakage measure

Simple approach: Add noise

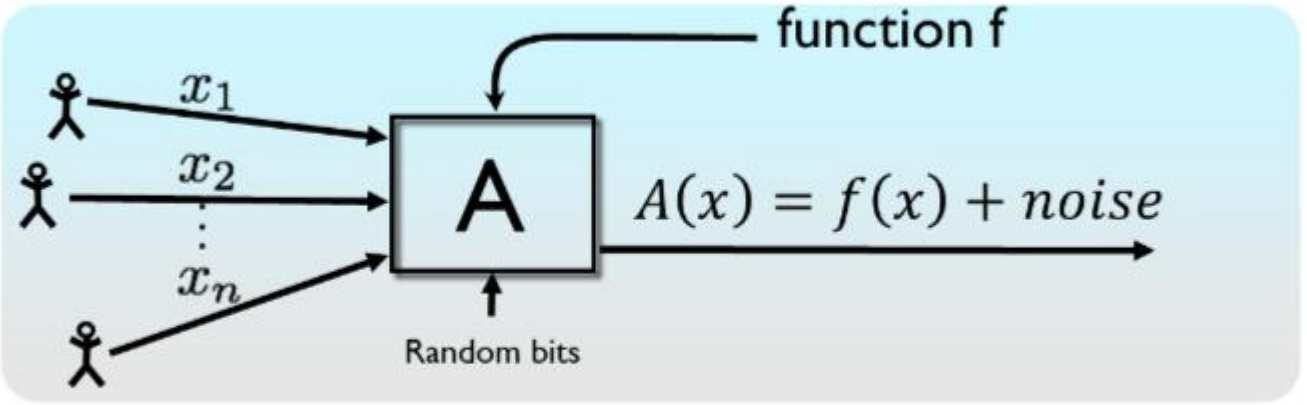
This approach dates back to [Warner 1965] who proposed to use ranodm noise as a way to incentivize individuals to be truthful.

Example:

Suppose we want to release the statistics of the fraction of students who cheated on the test. Instruct the students to answer honestly with probability 2/3 and to send an opposite answer with probability 1/3.

Using the law of large numbers one can use this noisy data to get a good estimate of the fraction of cheaters.

Laplace Mechanism



- Say we want to release a summary $f(x) \in \mathbb{R}^d$
  - e.g., proportion of cheaters: $x_i \in \{0,1\}$ and $f(x) = \frac{1}{n}\sum_i x_i$

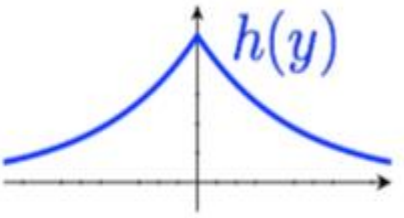
Simple approach: Add noise!

Question: How much noise is needed to ensure ϵ-differnetial privacy?

- Global Sensitivity: $GS_f = \max_{\text{neighbors } x,x'} \|f(x) - f(x')\|_1$

Example: $GS_{\text{proportion}} = \frac{1}{n}$

Laplace distribution: $h(y) = \frac{1}{2b} e^{-|y|/b}$



Theorem: If noise is added from Laplace distribution with $b \geq \frac{GS_f}{\epsilon}$, then A is ϵ-DP.

Proof:

$$\frac{\Pr[A(x) = v]}{\Pr[A(x') = v]} = \frac{e^{-\frac{|v-f(x)|}{b}}}{e^{-\frac{|v-f(x')|}{b}}} = e^{\frac{|v-f(x')|-|v-f(x)|}{b}} \leq e^{\frac{|f(x')-f(x)|}{b}} \leq e^{\epsilon}$$

Interpreting Differential Privacy:

A naive hope:

~~Your beliefs about me are the same after you see the output as they were before~~

Impossible!

Suppose you know that I smoke.

Suppose a clinical study shows that smoking and cancer are correlated.

You learned something about me!

Differential privacy implies:

No matter what you know ahead of time, you learn almost the same about me whether or not I partipated in the dataset.

DP composes well:

Leaking k e-DP functions is ke-DP.

Differential privacy does not imply:

Privacy for a group of individuals

DP is deployed:

Apple, Google, Census bureau, and Uber

Mainly used for counting and average statistics.

Challenges for DP in practice

Managing privacy loss over time.

Dealing with small datasets.

Analysts are used to working with raw data, as oposed to querying