Data Privacy

* Simplest privacy method: Anonymization.
  1. Remove "identifying" bits - eg, names, addresses, etc.
  2. Publish data.

* This has serious problems.

* Why? People's data tends to be very unique. For example:

- Gender
- Position
- Dept
- Ethnicity
- F
- Faculty
- CSE
- S Asian

Only one person, Kosaloka, fits description. "Linkage" information on CSE website.

* Statistics from data also problematic.
  eg. Wong et al study
  Histogram with outliers.

* Need robust, rigorous measures to preserve privacy: tradeoffs.

---

Three Privacy Settings:

1. Local sanitizer
2. Centralized sanitizer
3. Consent management
Local Sanitizer

1. Aggregator is untrusted.
2. Sanitization happens locally before passing on to Aggregator.

Example: Randomized Response. [Werner 1965]
Each person is asked if they use an illegal drug (Yes or No). Everyone takes their answer, flips w.p. \(p\) and returns it.

Privacy offered: deniability.

\[
\Pr\left(\text{Output} = Y \mid \text{True value} = Y\right) = 1 - p
\]
\[
\Pr\left(\text{Output} = N \mid \text{True value} = Y\right) = p
\]

Using Bayes Rule,
\[
\Pr\left(\text{True} = Y \mid \text{Output} = Y\right) = \frac{\Pr\left(\text{Output} = Y \mid \text{True} = Y\right) \Pr\left(\text{True} = Y\right)}{\Pr\left(\text{Output} = Y\right)}
\]

\[
\Pr\left(\text{Output} = Y \mid \text{True} = Y\right) \Pr\left(\text{True} = Y\right) + \Pr\left(\text{Output} = Y \mid \text{True} = N\right) \Pr\left(\text{True} = N\right)
\]

\[
= \frac{(1 - p) \frac{1}{2}}{\frac{1}{2}(1 - p) + \frac{1}{2} \cdot p} = 1 - p.
\]
What is the utility offered?

Suppose we are aggregating $N$ responses, and $f$ is the fraction of drug uses.

\[ \mathbb{E}[\# \text{Yes}] = N \Pr[\text{Output} = Y] \]

\[ \Pr(\text{Output} = Y) = \Pr(\text{Output} = Y | \text{True} = Y) \Pr(\text{True} = Y) \]
\[ + \Pr(\text{Output} = Y | \text{True} = N) \Pr(\text{True} = N) \]
\[ = (1-p)f + p(1-f) = pf + f - 2pf \]

\[ \mathbb{E}[\# \text{Yes}] = N(pf + f - 2pf) = Np + N(1-2p)f \]

\[ \text{Var}(\# \text{Yes}) \leq \sqrt{N} \]

Let \[ T = \frac{\# \text{Yes} - Np}{N(1-2p)} \]

Then \[ \mathbb{E}[T] = f \]

\[ \text{Var}(T) \leq \frac{1}{\sqrt{N(1-2p)}} \]

Comment estimate the fraction of drug users if $p$ is not too close to $\frac{1}{2}$.

**Note**: Privacy - Utility - Data size tradeoff

$p < \frac{1}{2}$: High privacy, but high variance of $T$, so low utility.

Tradeoff is better when $N$ is high.

**Applications**: Data collection systems in companies, eg. Google, Apple, etc.
Centralized Sanitizer

1. Sanitizer is trusted.
2. Sanitizer sees/collects raw sensitive data and "privatizes" it and passes the privatized version to the public sphere.

What is a good notion of privacy for such a setting? Differential privacy.

Main idea: Participation of a single person does not make a difference.

Alice + data $\rightarrow$ Algorithm $\rightarrow$ Output $\rightarrow$ Statistics
Bob + data $\rightarrow$ Algorithm $\rightarrow$ Output $\rightarrow$ Data release.

Example: Medical data, held at a hospital, summaries released.
Example 2: Census (2020 Census)

Adversary: whatever she can learn about Alice from algorithm's output, she can learn even if Alice is not in data.

Example 1: Study shows smoking causes cancer, Adversary knows Alice smokes $\rightarrow$ infers Alice may have cancer.

NOT a privacy violation.
Example 2: A Study in Wang et al with Alice's data.
Adversary knows about Alice's genome \(\rightarrow\) solves equations and finds Alice in Cancer group.

Formally, Differential privacy definition:

A mechanism \(A\) is \(\epsilon\)-DP if for all datasets \(D\) and \(D'\) that differ in the private value of a single person and for all \(t\),
\[
Pr(A(D) = t) \leq e^{\epsilon} Pr(A(D') = t).
\]
\(\epsilon\) = privacy budget
\(Pr\) = over randomization of the algorithm \(A\).

Properties: 1. Post-processing Invariance.

\(\text{Data} \Rightarrow \epsilon\)-DP \(\Rightarrow \text{answer} \Rightarrow \epsilon\)-DP (for any transformation)

2. (Graceful) Composition

Sequential

Parallel

\(\epsilon_1\)-DP
\(\epsilon_2\)-DP
\(\max(\epsilon_1, \epsilon_2)\) - DP
How to get DP?

- The Global Sensitivity Mechanism.
- Global sensitivity of a function $f$:
  \[
  GS(f) = \max_{D, D'} |f(D) - f(D')| \quad |D \setminus D'| = 1
  \]

Mechanism:
Output $f(D) + \frac{GS(f)}{\varepsilon} Z$ where $Z \sim \text{Laplace}(\frac{1}{\varepsilon})$.

Laplace distribution:
\[
 f(Z) = \frac{1}{2b} e^{-|Z|/b} \quad \text{[mean 0, std dev: } \sqrt{2b}]\]

Use this mechanism + composition to get more complex ones.

Utility: Back to the drug user problem, if we use DP, then
\[
 GS \text{ of mean of } n \text{ bits } = \frac{1}{n}
\]
So std dev of noise added = $\frac{\sqrt{2}}{n\varepsilon}$ (as opposed to $\frac{1}{\sqrt{n(1-2p)}}$ for RR)

So better privacy - utility - data size tradeoff