# CS208: Applied Privacy for Data Science
# Reidentification & Reconstruction Attacks

School of Engineering & Applied Sciences
Harvard University

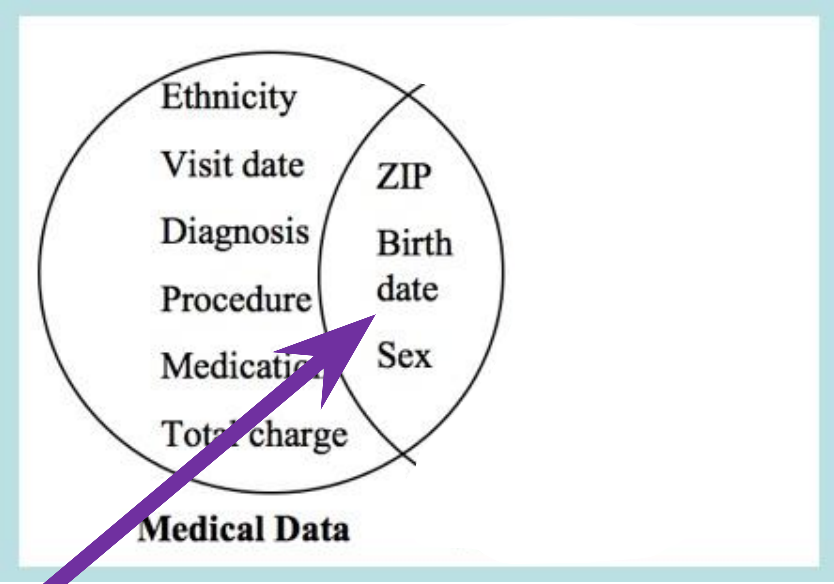
January 29. 2025

# Announcements

- Fill out first-class survey if you haven't already: https://shorturl.at/jSosl
- Post questions to Ed rather than emailing us individually.
  Keep an eye on Ed for announcements!
- Let us know ASAP if you can't access course platforms (esp. Ed, Perusall).
- Office hours the rest of this week:
  - Salil Fri 10:30am-12pm (SEC 3.327)
  - Priyanka Wed 2:30pm-4:30pm (SEC 2.101)
  - Zach Thu 3pm-4pm (SEC 3.314)
- Probability/algorithms/stats review sessions this week:
  - Jason Wed 3pm-4pm, Science Center 304
  - Zach Thu 9:45-11:00am, SEC 4.308+Zoom+recording (possibly including programming)

# Reidentification via Linkage

| ~~Name~~ | Sex | Blood | | HIV? |
|---|---|---|---|---|
| ~~Chen~~ | F | B | | Y |
| ~~Jones~~ | M | A | | N |
| ~~Smith~~ | M | O | | N |
| ~~Ross~~ | M | O | | Y |
| ~~Lu~~ | F | A | | N |
| ~~Shah~~ | M | B | | Y |




[Sweeney `97]

Uniquely identify > 60% of the US population [Sweeney `00, Golle `06]

# Deidentification via Generalization

- Def (generalization): A generalization mechanism is an algorithm $A$ that takes a dataset $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and outputs $A(x) = (T_1, \ldots, T_n)$ where $x_i \in T_i \subseteq \mathcal{X}$ for all $i$.

- Example:

| Name | Sex | Blood | | HIV? |
|---|---|---|---|---|
| * | F | B | | Y |
| * | M | A | | N |
| * | M | O | | N |
| * | M | O | | Y |
| * | F | A | | N |
| * | M | B | | Y |

$$T_i = \{\text{all strings}\} \times \{x_{i2}\} \times \cdots \times \{x_{im}\}$$

# K-Anonymity [Sweeney `02]

- Def (generalization): A generalization mechanism $A$ satisfies $k$-anonymity (across all fields) if for every dataset $x = (x_1, \dots, x_n) \in \mathcal{X}^n$ the output $A(x) = (T_1, \dots, T_n)$ has the property that every set $T$ that occurs at all occurs at least $k$ times.

- Example: 3-anonymizing a dataset

$x =$

| ZIP | Income | COVID |
|---|---|---|
| 91010 | \$125k | Yes |
| 91011 | \$105k | No |
| 91012 | \$80k | No |
| 20037 | \$50k | No |
| 20037 | \$20k | No |
| 20037 | \$25k | Yes |

$\xrightarrow{A}$

| ZIP | Income | COVID |
|---|---|---|
| 9101⋆ | \$75–150k | ⋆ |
| 9101⋆ | \$75–150k | ⋆ |
| 9101⋆ | \$75–150k | ⋆ |
| 20037 | \$0–75k | ⋆ |
| 20037 | \$0–75k | ⋆ |
| 20037 | \$0–75k | ⋆ |

$= A(x)$

# Quasi-Identifiers

- Typically, $k$-anonymity only applied on "quasi-identifiers" – attributes that might be linked with an external dataset. i.e. $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$, where $\mathcal{Y}$ is domain of quasi-identifiers, and $T_i = U_i \times V_i$, where each $U_i$ occurs at least $k$ times.

- 

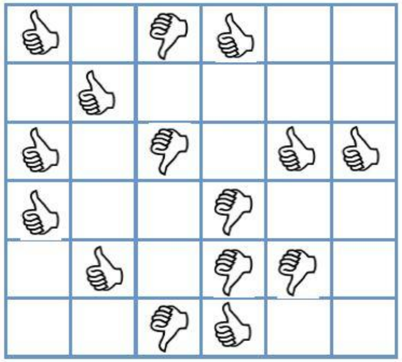
| Zip code | Age | Nationality | Condition |
|---|---|---|---|
| 130** | <30 | * | AIDS |
| 130** | <30 | * | Heart Disease |
| 130** | <30 | * | Viral Infection |
| 130** | <30 | * | Viral Infection |
| 130** | ≥40 | * | Cancer |
| 130** | ≥40 | * | Heart Disease |
| 130** | ≥40 | * | Viral Infection |
| 130** | ≥40 | * | Viral Infection |
| 130** | 3* | * | Cancer |
| 130** | 3* | * | Cancer |
| 130** | 3* | * | Cancer |
| 130** | 3* | * | Cancer |

Q: what could go wrong?

Q: What if no quasi-identifiers?

# Netflix Challenge Re-identification

[Narayanan & Shmatikov `08]



**Anonymized**
NetFlix data

Q: Why would Netflix release such a dataset?

Image credit: Arvind Narayanan

# Narayanan-Shmatikov Set-Up

- Dataset: $x$ = set of records $r$ (e.g. Netflix ratings)

- Adversary's inputs:
  - $\hat{x}$ = subset of records from $x$, possibly distorted slightly
  - $aux$ = auxiliary information about a record $r \in D$ (e.g. a particular user's IMDB ratings)

- Adversary's goal: output either
  - $r'$ = record that is "close" to $r$, or
  - $\perp$ = failed to find a match

# Narayanan-Shmatikov Algorithm

1. Calculate $\mathrm{score}(aux, r')$ for each $r' \in \hat{x}$, as well as the standard deviation $\sigma$ of the calculated scores.
2. Let $r_1{}'$ and $r_2{}'$ be the records with the largest and second-largest scores.
3. If $\mathrm{score}(aux, r_1{}') - \mathrm{score}(aux, r_2{}') > \phi \cdot \sigma$, output $r_1{}'$, else output $\perp$.

An instantiation:

IMDB movies rated by user

Similarity of rating & date

Downweight movies watched by many Netflix users

$$\mathrm{score}(aux, r') = \sum_{a \in \mathrm{supp}(aux)} \mathrm{sim}(aux_a, r'_a)$$

eccentricity $\phi = 1.5$

# Narayanan-Shmatikov Results

- For the $1m Netflix Challenge, a dataset of ~.5 million subscribers' ratings (less than 1/10 of all subscribers) was released (total of ~$100m ratings over 6 years).
- Out of 50 sampled IMBD users, two standouts were found, with eccentricities of 28 and 15.
- Reveals all movies watched from only those publicly rated on IMDB.
- Class action lawsuit, cancelling of Netflix Challenge II.

Message: any attribute can be a "quasi-identifier"

# k-anonymity across all attributes?

- Utility concerns?
  - Significant bias even when applied on quasi-identifiers, cf. [Daries et al. `14]

- Privacy concerns?
  - Consider mechanism $A(x)$: if Salil is in $x$ and has tuberculosis, generalize starting with rightmost attribute. Else generalize starting on left.
  - Message: privacy is not only a property of the output, but of the input-output relationships.

# Downcoding Attacks [Cohen `21]

$\mathbf{X} =$

| ZIP | Income | COVID |
|---|---|---|
| 91010 | \$125k | Yes |
| 91011 | \$105k | No |
| 91012 | \$80k | No |
| 20037 | \$50k | No |
| 20037 | \$20k | No |
| 20037 | \$25k | Yes |

- Downcoding undoes generalization
- $X$ is the original dataset → $Y$ is a 3-anonymized version
- $Z$ is a **downcoding**: It *generalizes* $X$ and *refines* $Y$

Slide credit: Aloni Cohen

# Cohen's Result

**Theorem (informal):** There are settings in which **every** minimal, hierarchical k-anonymizer (even enforced on all attributes) enables strong downcoding attacks.

Setting

- A (relatively natural) data **distribution** and **hierarchy**, which depend on k

Strength

- **How many** records are refined? $\Omega(N)$ ($> 3\%$ for $k \leq 15$)
- **How much** are records refined? $3D/8$ (38% of attributes)
- **How often**? w.p. $1 - o(1)$ over a random dataset

# Composition Attacks

- [Ganti-Kasiviswanathan-Smith `08]:
  Two k-anonymous generalizations of the same dataset can be combined to be not k-anonymous.

- [Cohen `21]:
  Reidentification on Harvard-MIT EdX Dataset [Daries et al. `14]
  - 5-anonymity enforced separately (a) on course combination, and (b) on demographics + 1 course

# EdX Quasi-identifiers

| User 17 | Year of Birth | Gender | Country | Course 1 | Course 2 | Course 3 | |
|---|---|---|---|---|---|---|---|
| | 2000 | F | India | Yes | No | Yes | Enrolled |
| | | | | 5 | | 8 | # Posts |
| | | | | Yes | | No | Certificate |

{Year of Birth, Gender, Country, Course(i).Enrolled, Course(i).Posts}
for i = 1, . . ., 16

| User 17 | Year of Birth | Gender | Country | Course 1 | Course 2 | Course 3 | |
|---|---|---|---|---|---|---|---|
| | 2000 | F | India | Yes | No | Yes | Enrolled |
| | | | | 5 | | 8 | # Posts |
| | | | | Yes | | No | Certificate |

{Course(1).Enrolled, Course(2).Enrolled, . . ., Course(16).Enrolled

Slide credit: Aloni Cohen

# Failure of Composition

| User 17 | YoB | Gender | Country | Course 1 | Course 2 | Course 3 | |
|---|---|---|---|---|---|---|---|
| | 2000 | F | India | Yes | No | Yes | **Enrolled** |
| | | | | 5 | | 8 | **# Posts** |
| | | | | Yes | | No | **Certificate** |

If you combine the QIs:

- 7.1% uniques (34,000)
- 15.3% not 5-anonymous

Reidentification carried out using LinkedIn profiles → dataset heavily redacted

Slide credit: Aloni Cohen

# Reading & Discussion for Next Time

- Q: How should we respond to the failure of de-identification?
- Not assigned: writings claiming that de-identification works (see cs208 annotated bibliography)
- Next: what if we only release aggregate statistics?

# Attacks on Aggregate Statistics

| ID | US? |
|---|---|
| | |
| | |
| | |
| | |
| | |
| | |

- Stylized set-up:
  - Dataset $x \in \{0,1\}^n$.
  - (Known) person $i$ has sensitive bit $x_i$.
  - Adversary gets $q_S(x) = \sum_{i \in S} x_i$ for various $S \subseteq [n]$.

- How to attack if adversary can query chosen sets $S$?

- What if we restrict to sets of size at least $n/10$?

This attack has been used on Israeli Census Bureau!
(see [Ziv `13])

# Attacks on Exact Releases

- What if adversary cannot choose subsets, but $q_S(x)$ is released for "innocuous" sets $S$?

- Example: uniformly random $S_1, S_2, \dots, S_m \subseteq [n]$ are chosen, and adversary receives:
$\left(S_1, a_1 = q_{S_1}(x)\right), \left(S_2, a_2 = q_{S_2}(x)\right), \dots, (S_m, a_m = q_{S_m}(x))$

- Claim: for $m = n$, with prob. $1 - o(1)$ adversary can reconstruct entire dataset!

- Proof?

# Example for $n = 5$

$S_1 = \{1,2,3\}$, $a_1 = 2$, $S_2 = \{1,3,4\}$, $a_2 = 1$, $S_3 = \{4,5\}$, $a_3 = 1$,
$S_4 = \{2,3,4,5\}$, $a_4 = 3$, $S_5 = \{1,2,4,5\}$, $\mathrm{a}_5 = 2$

Unknowns: $x_1, x_2, \ldots, x_5$

Equations:

1. $x_1 + x_2 + x_3 = 2$
2. $x_1 + x_3 + x_4 = 1$
3. $x_4 + x_5 = 1$
4. $x_2 + x_3 + x_4 + x_5 = 3$
5. $x_1 + x_2 + x_4 + x_5 = 2$

Unique Solution:

$x_1 = 0$
$x_2 = 1$
$x_3 = 1$
$x_4 = 0$
$x_5 = 1$

# Attacks on Approximate Statistics

- What if we release statistics $a_i \approx q_{S_i}(x)$?

- Thm [Dinur-Nissim `03]: given $m = n$ uniformly random sets $S_j$ and answers $a_j$ s.t. $\left|a_j - q_{S_j}(x)\right| \leq E = o(\sqrt{n})$, whp adversary can reconstruct $1 - o(1)$ fraction of the bits $x_i$.

- Proof idea: $A(S_1, a_1, \ldots, S_m, a_n)$ = any $\hat{x} \in \{0,1\}^n$ s.t.
  $\forall j \ \left|a_j - q_{S_j}(\hat{x})\right| \leq E$.

  (Show that whp, for all $\hat{x}$ that differs from $x$ in a constant fraction of bits, $\exists j$ such that $\left|q_{S_j}(\hat{x}) - q_{S_j}(x)\right| > 2E$.)

# Integer Programming Implementation

$A(S_1, a_1, \ldots, S_m, a_n)$:

1. Find a vector $\hat{x} \in \mathbb{Z}^n$ such that:
   – $0 \le \hat{x}_i \le 1$ for all $i = 1, \ldots, n$
   – $-E \le a_j - \sum_{i \in S_j} \hat{x}_i \le E$ for all $j = 1, \ldots, m$
2. Output $\hat{x}$.

Problem: Can be computationally expensive ("NP-hard", exponential time in worst case)

# Faster: Linear Programming Implementation

$A(S_1, a_1, \dots, S_m, a_n)$:

1. Find a vector $\hat{x} \in \mathbb{R}^n$ such that:
   - $0 \leq \hat{x}_i \leq 1$ for all $i = 1, \dots, n$
   - $-E \leq a_j - \sum_{i \in S_j} \hat{x}_i \leq E$ for all $j = 1, \dots, m$
2. Output $\hat{x}$

# Linear Programming Implementation for Average Error

$A(S_1, a_1, \dots, S_m, a_n)$:

1. Find vectors $\hat{x} \in \mathbb{R}^n$ and $E \in \mathbb{R}^m$
   - Minimizing $\sum_{j=1}^{m} E_j$ and such that
   - $0 \le \hat{x}_i \le 1$ for all $i = 1, \dots, n$
   - $-E_j \le a_j - \sum_{i \in S_j} \hat{x}_i \le E_j$ for all $j = 1, \dots, m$
2. Output $\mathrm{round}(\hat{x})$.

# Least-Squares Implementation for MSE

•

$A(S_1, a_1, \dots, S_m, a_n)$:

1. Find vector $\hat{x} \in \mathbb{R}^n$ minimizing

$$\sum_{j=1}^{m} \left( a_j - \sum_{i \in S_j} \hat{x}_i \right)^2 = \|a - M_S \hat{x}\|^2$$

2. Output $\mathrm{round}(\hat{x})$.

Also works for random $S_j$'s, and is much faster than LP!

# On the Level of Accuracy

- The theorems require the error per statistic to be $o(\sqrt{n})$. This is necessary for reconstructing almost all of $x$.

- Q: What is significant about the threshold of $\sqrt{n}$?
  - If dataset is a random sample of size $n$ from a larger population, the standard deviation of a count query is $O(\sqrt{n})$.
  - Reconstruction attacks $\Rightarrow$ if we want to release many ($> n$) arbitrary or random counts, then we need introduce error at least as large as the sampling error to protect privacy.

# How to Make Subset Sum Queries?

- Stylized set-up:
  - Dataset $x \in \{0,1\}^n$.
  - (Known) person $i$ has sensitive bit $x_i$.
  - Adversary gets $a_S \approx q_S(x) = \sum_{i \in S} x_i$ for various $S \subseteq [n]$.

| ID | US? |
|---|---|
| | |
| | |
| | |
| | |
| | |
| | |

- Q: How to attack if the subjects aren't numbered w/ ID's?
  - If we know the set of people but not their IDs? (e.g. current Harvard students)
  - If we only know the size $n$ of the dataset?

# Overall Message

- Every statistic released yields a (hard or soft) constraint on the dataset.
  - Sometimes have nonlinear or logical constraints $\Rightarrow$ use fancier solvers (e.g. SAT or SMT solvers)
- Releasing too many statistics with too much accuracy necessarily determines almost the entire dataset.
- This works in theory and in practice (see readings, ps2).
- We need a quantitative theory that tells us "how much is too much" $\rightarrow$ differential privacy!