CS208: Applied Privacy for Data Science
Reidentification & Reconstruction Attacks

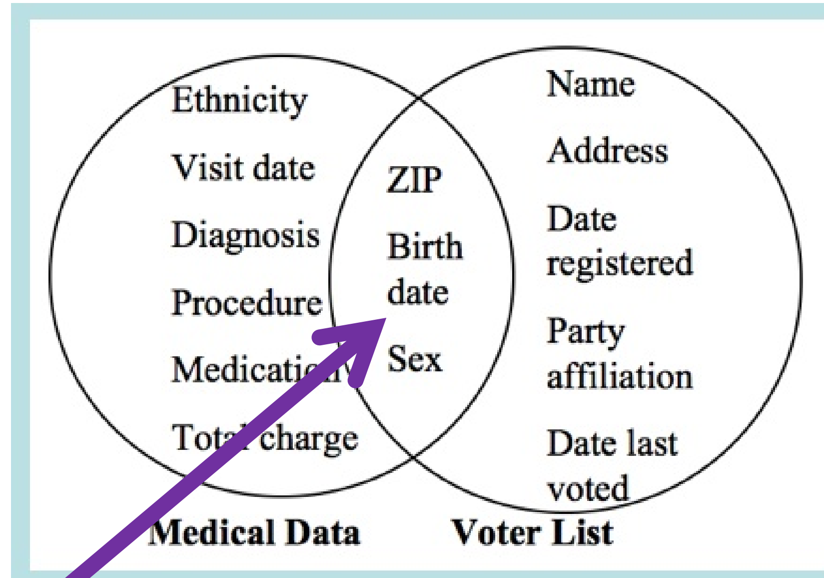
James Honaker & Salil Vadhan
School of Engineering & Applied Sciences
Harvard University

February 1, 2019

CRCS Center for Research on Computation and Society

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]

Q: What's your response to the Personal Genome Project re-identification?

Some Possible Responses

- Privacy is dead, informed consent is enough
- Informed consent is a fiction
- Value of the research trumps privacy
- Public sharing not needed for research purposes

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 $A(x) = (S_1, \ldots, S_n)$ where $x_i \in S_i \subseteq \mathcal{X}$ 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

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

K-Anonymity [Sweeney `02]

- Def (generalization): A generalization mechanism A satisfies k-anonymity (across all fields) if for every dataset $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ the output $A(x) = (S_1, \ldots, S_n)$ has the property that every set S that occurs at all occurs at least k times.

- Example: a 4-anonymous output

Zip code	Age	Nationality
130**	<30	*
130**	<30	*
130**	<30	*
130**	<30	*
130**	≥40	*
130**	≥40	*
130**	≥40	*
130**	≥40	*
130**	3*	*
130**	3*	*
130**	3*	*
130**	3*	*

Intuition: your privacy is protected if I can't isolate you.

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 $S_i = T_i \times U_i$, where each T_i occurs at least k times.

- Example:

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?

Failure of Composition

[Ganti-Kasiviswanathan-Smith `08]

Suppose two k-anonymous datasets are released,
and we know the quasi-identifiers in someone in both…

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

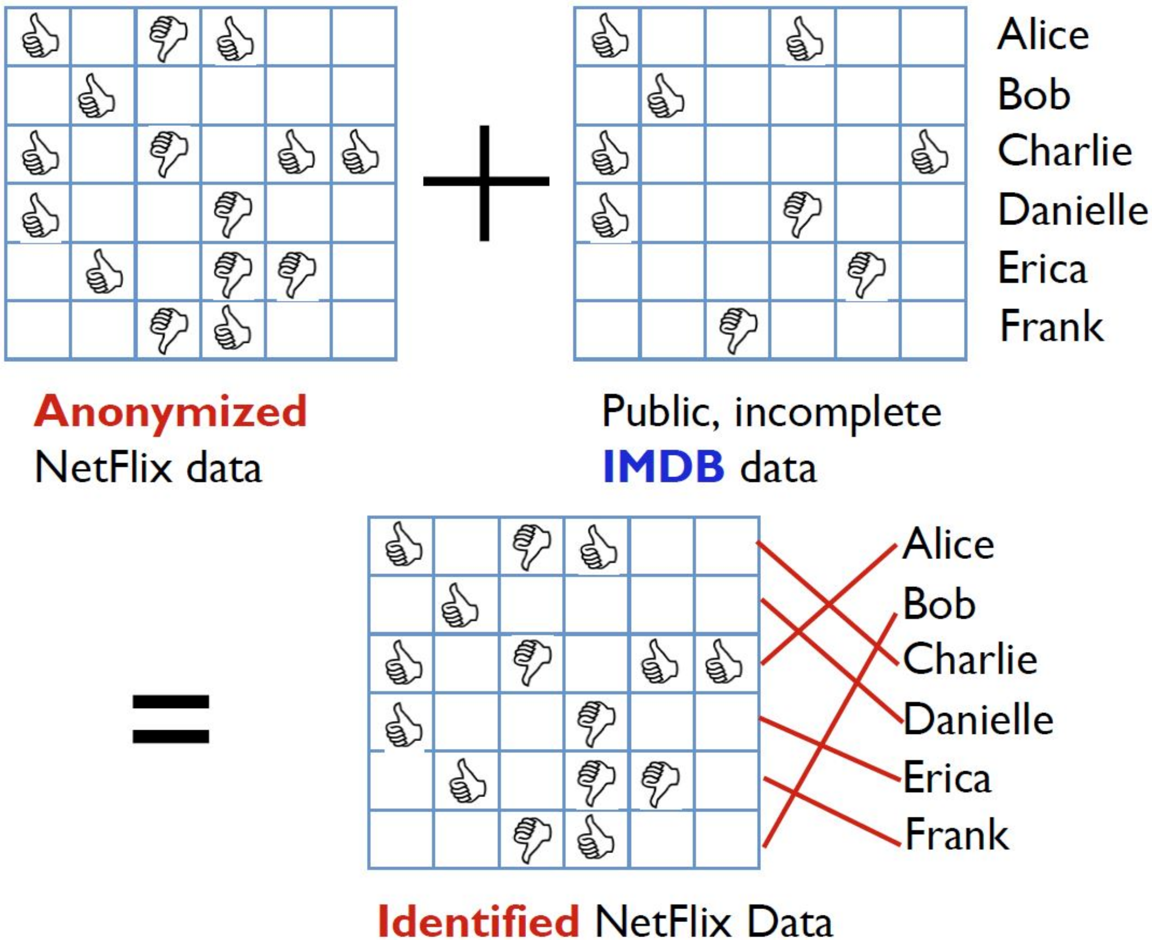
Zip code	Age	Nationality	Condition
130**	<35	*	AIDS
130**	<35	*	Tuberculosis
130**	<35	*	Flu
130**	<35	*	Tuberculosis
130**	<35	*	Cancer
130**	<35	*	Cancer
130**	≥ 35	*	Cancer
130**	≥ 35	*	Cancer
130**	≥ 35	*	Cancer
130**	≥ 35	*	Tuberculosis
130**	≥ 35	*	Viral Infection
130**	≥ 35	*	Viral Infection

k-anonymity across all fields

- 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.

Netflix Challenge Re-Identification

[Narayanan-Shmatikov `08]



On average, four movies uniquely identify user

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 σ 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:

$$\mathrm{score}(aux, r') = \overbrace{\sum_{a \in \mathrm{supp}(aux)}}^{\text{IMDB movies rated by user}} \overbrace{\frac{1}{\log|\{r' \in \hat{x} : a \in \mathrm{supp}(r')\}|}}^{\text{Downweight movies watched by many Netflix users}} \cdot \overbrace{\mathrm{sim}(aux_a, r'_a)}^{\text{Similarity of rating \& date}}$$

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"

Attacks on Aggregate Statistics

- 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 $n/10$?

ID	US?
1	1
2	0
3	0
4	1
$\vdots$	$\vdots$
n	1

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: $\left(S_1, a_1 = q_{S_1}(x)\right), \left(S_2, a_2 = q_{S_2}(x)\right), \dots, \left(S_m, a_m = \ q_{S_m}(x)\right)$

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

- Proof?

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. $\left|a_j - q_{S_j}(x)\right| \leq E = o\left(\sqrt{n}\right)$, whp adversary can reconstruct $1 - o(1)$ fraction of the bits x_i.
- Proof idea: $A(S_1, a_1, \dots, S_m, a_n)$

  = any $y \in \{0,1\}^n$ s.t. $\forall j \ \left|a_j - q_{S_j}(y)\right| \leq E$.

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

Integer Programming Implementation

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

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

Linear Programming Implementation

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

1. Find a vector $y \in \mathbb{R}^n$ such that:

   – $0 \leq y_i \leq 1$ for all $i = 1, \dots, n$

   – $-E \leq a_j - \sum_{i \in S_j} y_i \leq E$ for all $j = 1, \dots, m$

2. Output $\mathrm{round}(y)$. [coordinate-wise rounding]

Linear Programming Implementation for Average Error

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

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

Least-Squares Implementation for MSE

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

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

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

2. Output $\mathrm{round}(y)$.

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

Overall Message

- Every statistic released yields a (hard or soft) constraint on the dataset.
- Releasing too many statistics with too much accuracy necessarily determines almost the entire dataset.
- This works in theory and in practice (see readings, ps1).
- We need a quantitative theory that tells us “how much is too much” → differential privacy!

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: How could we defend against reconstruction attacks if we allow error $\Omega(\sqrt{n})$?

On the Level of Accuracy

Q: How could we defend against reconstruction attacks if we allow error $\Omega(\sqrt{n})$?

1. Always release $a_j = (\sum_{i=1}^{n} x_i)/2$.
   For random S_j has expected error $O(\sqrt{n})$ per query and expected maximum error $O\left(\sqrt{n \cdot \log m}\right)$.
2. Always release $a_j = (n/t) \cdot \left(\sum_{i \in T \cap S_j} x_i\right)/2$ where T is a random set of t rows chosen once.
   For arbitrary S has expected error $O(n/\sqrt{t})$ per query and expected maximum error $O\left(n\sqrt{\log m}/\sqrt{t}\right)$.
3. Add random noise, e.g. $a_j = \left(\sum_{i \in S_j} x_i\right) + e_j$ where $e_j \sim \mathcal{N}(0, \sigma^2)$ for an appropriate $\sigma = \Omega(\sqrt{n})$.
   For arbitrary S has expected error $O(\sigma)$ per query and expected maximum error $O\left(\sigma\sqrt{\log m}\right)$.

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?
1	1
2	0
3	0
4	1
⋮	⋮
n	1

- 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?