CS208: Applied Privacy for Data Science
Overview & Attacks on Privacy

School of Engineering & Applied Sciences
Harvard University

January 25, 2022

Course Staff

- Instructors
  - Salil Vadhan
  - James Honaker
  - Wanrong Zhang
- Teaching Fellows
  - Daniel Alabi
  - Jayshree Sarathy
  - Mike Shoemate (half time)
  - Connor Wagaman (half time)

To do before Thursday

- Use a name placard in class
- Fill out class background survey
- Check that you can access our platforms: Ed, Perusall, Panopto
- Read the guidelines for reading & commenting
- Watch the video (posted on Panopto) from the preview session if you haven’t already done so
- Comment on and read the readings assigned for Thurs
- Review updated syllabus for covid & auditor policies
- Look out for PS1 (due Wed 2/2), Section & OH this week. (Future psets will be due on Fridays, PS2 due 2/11).

Data Privacy: The Problem

Given a dataset with sensitive information, such as:

- Census data
- Health records
- Social network activity
- Telecommunications data

- Academic research
- Informing policy
- Identifying subjects for drug trial
- Searching for terrorists
- Market analysis
- …

How can we:

- enable “desirable uses” of the data
- while protecting the “privacy” of the data subjects?

????

Privacy Models from Theoretical CS

Model	Utility	Privacy	Who Holds Data?
Differential Privacy	statistical analysis of dataset	individual-specific info	trusted curator
Secure Multiparty Computation	any query desired	everything other than result of query	original users (or semi-trusted delegates)
Fully Homomorphic (or Functional) Encryption	any query desired	everything (except possibly result of query)	untrusted server

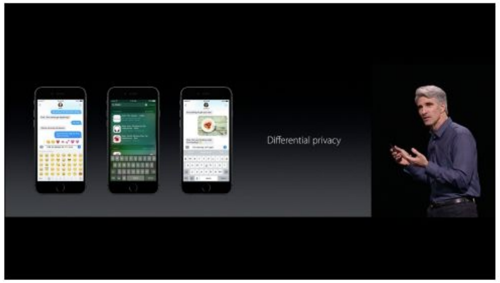
DP Theory

Differential privacy research has

- many intriguing theoretical challenges
- rich connections w/other parts of CS theory & mathematics

e.g. cryptography, learning theory, game theory & mechanism design, convex geometry, pseudorandomness, optimization, approximability, communication complexity, statistics, …

Differential Privacy Deployed



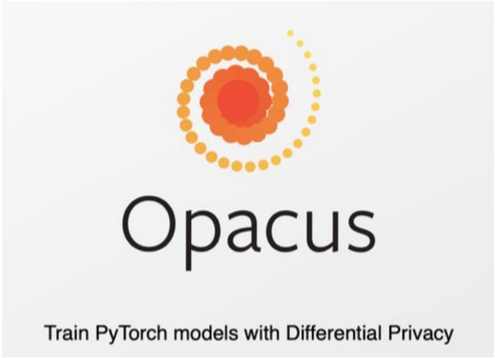
Apple

Google

Microsoft

Census Bureau

Uber



Meta

Harvard Privacy Tools Project

http://privacytools.seas.harvard.edu/

NSF

U.S. DEPARTMENT OF COMMERCE BUREAU OF THE CENSUS

Alfred P. Sloan FOUNDATION

Google

Computer Science, Law, Social Science, Statistics

IQSS The Institute for Quantitative Social Science HARVARD UNIVERSITY

CRCS Center for Research on Computation and Society

BERKMAN KLEIN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

DATA PRIVACY LAB

OpenDP

http://opendp.org/

A community effort to build a trustworthy and open-source suite of differential privacy tools that can be easily adopted by custodians of sensitive data to make it available for research and exploration in the public interest.

Why?

- Channel the collective advances on science & practice of DP
- Enable wider adoption of DP
- Address high-demand, compelling use cases
- Provide a starting point for custom DP solutions
- Identify important research directions for the field

Class Goals

By the end of the course, we hope that you will all be able to:

- Identify and demonstrate risks to privacy in data science settings,
- Correctly match differential privacy technology with an application,
- Safely implement privacy solutions, and experimentally validate the performance and utility of algorithms,
- Understand differential privacy at a level sufficient to engage in research about best practices in implementation, apply the material in practice, and/or connect it to other areas,
- Analyze the ethical and policy implications of differential privacy deployments,
- Formulate and carry out an interesting, short-term independent research project, and present the work in both written and oral form.

Course Elements

- Asynchronous readings or videos to comment on
- Lecture/discussion and practicum class meetings (live-streamed for students in isolation & recorded to create open-access materials)
- Problem sets, approx. weekly. Mix of analytical and experimental problems.
- Weekly section and office hours
- Final project

Grading: approx. ⅓ participation, ⅓ problem sets, ⅓ project

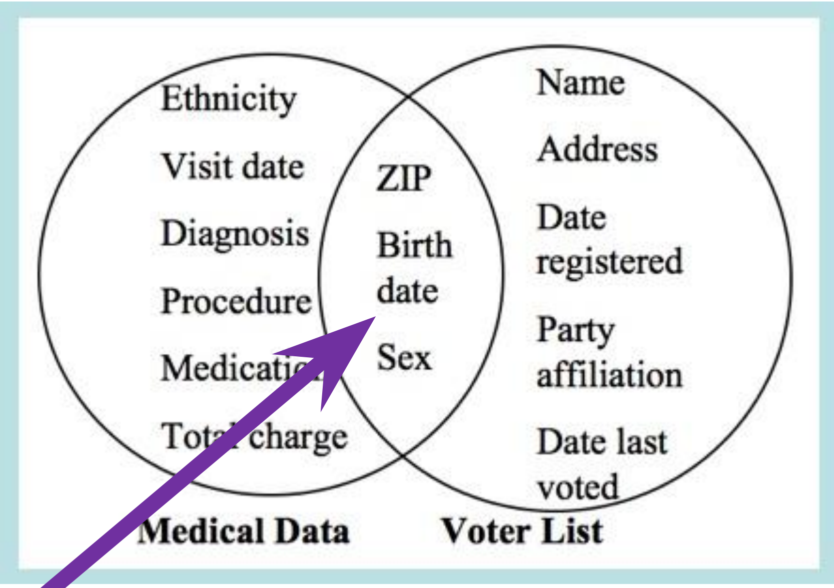
Ethics, Law, and Society

- Analyze differential privacy deployments from various perspectives
  - **Ethics**: How does differential privacy alter ethical considerations around collecting sensitive data for <u>public interest</u> purposes?
  - **Law and policy**: What is the relationship between differential privacy and existing <u>regulatory standards</u> for privacy protection?
  - **Science & Technology Studies**: How does differential privacy reflect and shape <u>power dynamics</u> among data subjects, data holders, and researchers?

- Identify critiques, gaps, points of tension, and possible solutions

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 $A(x) = (S_1, \ldots, S_n)$ 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

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 $A(x) = (S_1, \dots, S_n)$ 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

•

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?

What if no quasi-identifiers?
Netflix Challenge Re-identification

[Narayanan & Shmatikov `08]

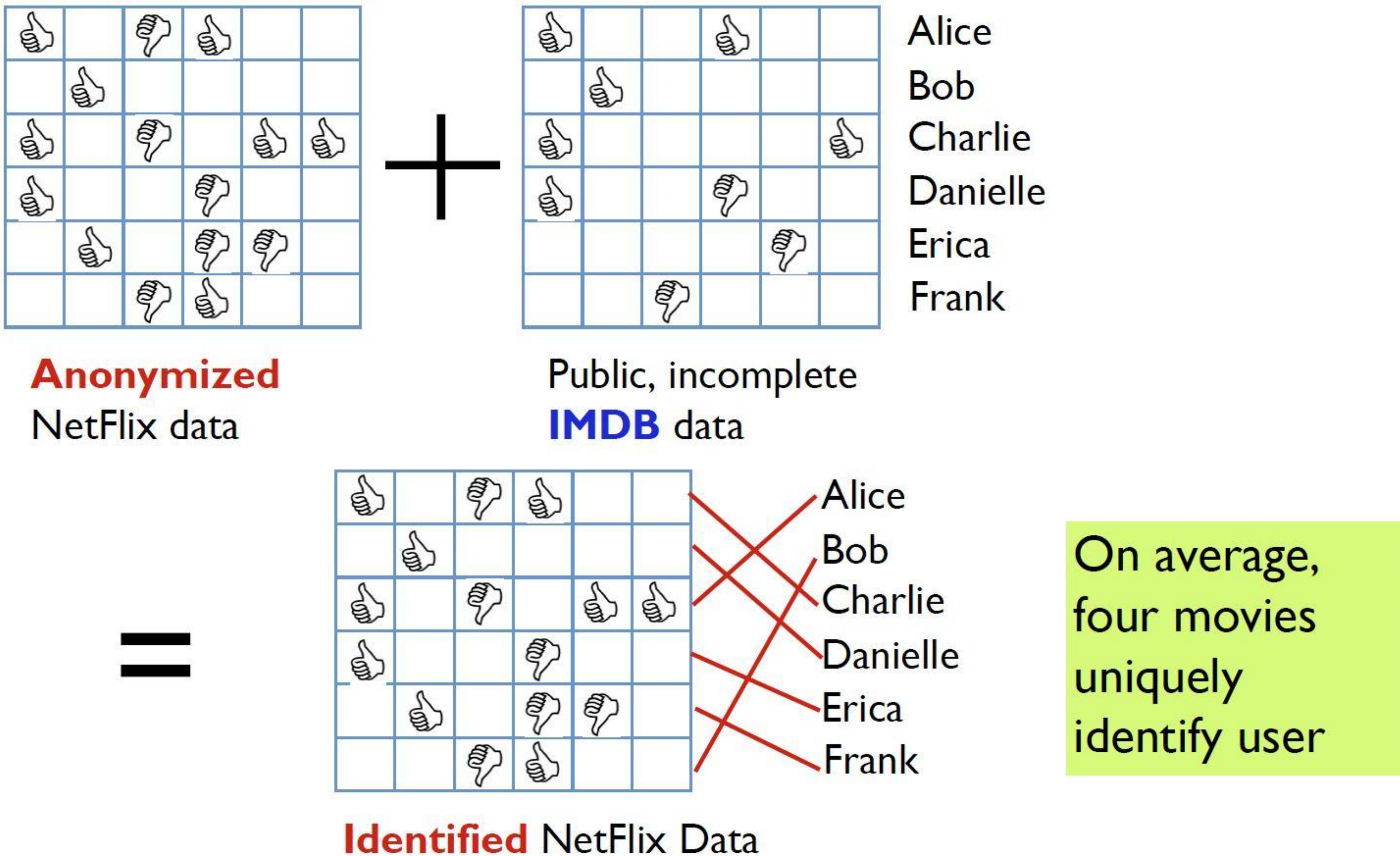


Image credit: Arvind Narayanan

Narayanan-Shmatikov Set-Up

•

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"

k-anonymity across all attributes

-

Downcoding Attacks [Cohen `21]

X =

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

Y =

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

Z =

ZIP	Income	COVID
91010	$125–150k	⋆
9101⋆	$100–125k	⋆
9101⋆	$75–150k	⋆
20037	$0–75k	No
20037	$0–75k	⋆
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

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

- Practice [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
  - Auxiliary information: LinkedIn profiles

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

Slide credit: Aloni Cohen

Reading & Discussion for Next Time

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

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 $S \subseteq [n]$.

ID	US?

- 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])