A Course on Differential Privacy, Fall 2020

Gautam Kamath
Assistant Professor, University of Waterloo
@TheGautamKamath
www.gautamkamath.com

This course: Preparation to conduct research in differential privacy.

Lecture 1

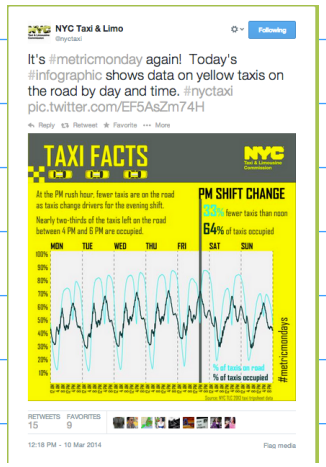
Some attempts at data privacy

Example 1: NYC Taxi and Limo Commission

2014

- Whong
- Freedom of Information

- fares + trips - 19GB




```
6B111958A39B24140C973B262EA9FEA5,D3B035A03C8A34DA17488129DA581EE7,VTS,5,,2013-12-03
15:46:00,2013-12-03 16:47:00,1,3660,22.71,-73.813927,40.698135,-74.093307,40.829346
```

```
medallion, hack_license, vendor_id, rate_code, store_and_fwd_flag, pickup_datetime,
dropoff_datetime, passenger_count, trip_time_in_secs, trip_distance, pickup_longitude,
pickup_latitude, dropoff_longitude, dropoff_latitude
```

[-] **ImJasonH** 3 points 6 years ago

Adding a filter for probably-erroneous data:

```
WHERE FLOAT(total_amount) < 1000
```

Gives an interesting result:

```
amt         date          hack_license
7682.76     2013-09-30    CFCD208495D565EF66E7DFF9F98764DA
6226.22     2013-09-28    CFCD208495D565EF66E7DFF9F98764DA
6218.94     2013-09-24    CFCD208495D565EF66E7DFF9F98764DA
5590.90     2013-09-25    CFCD208495D565EF66E7DFF9F98764DA
5406.15     2013-12-30    CFCD208495D565EF66E7DFF9F98764DA
5369.85     2013-12-27    CFCD208495D565EF66E7DFF9F98764DA
5358.72     2013-12-23    CFCD208495D565EF66E7DFF9F98764DA
5168.46     2013-09-19    CFCD208495D565EF66E7DFF9F98764DA
4837.05     2013-12-19    CFCD208495D565EF66E7DFF9F98764DA
4733.18     2013-12-13    CFCD208495D565EF66E7DFF9F98764DA
```

This guy's good! Maybe too good? :)

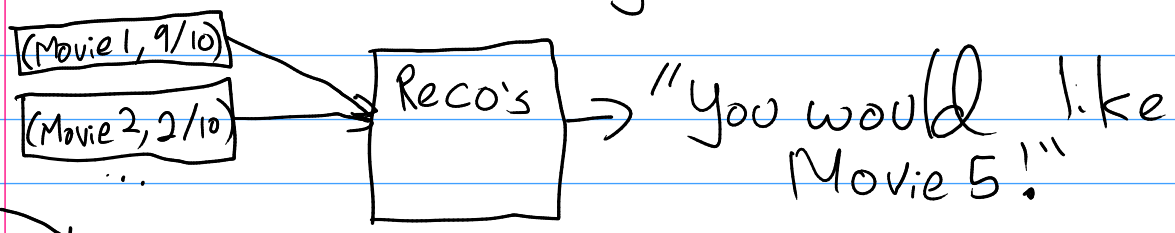
perma-link embed save parent give award

≈20M combinations

- Data linkage attack
- Random ids?

Example 2: Netflix Prize

- Recommendation engine 2006-2009



anon. →

(user ID, movie ID, rating, date)

Narayanan Shmatikov '08

Netflix data

anon.

(ID, movie, rating, date)

IMDb

public

(name, movie, rating, date)

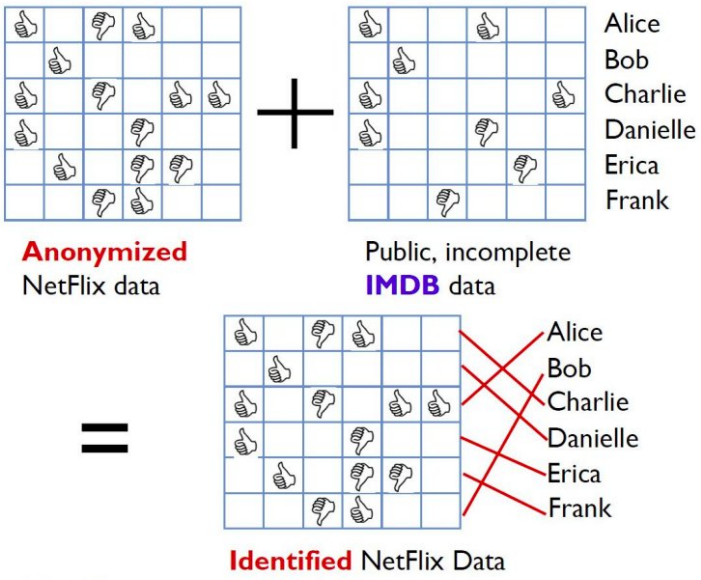


Image credit: Arvind Narayanan

Example 3: Memorization in Neural Networks

Carlini et al. '19

text corpus $Y \rightarrow$ model f_θ

Given $(x_1, \ldots, x_n)$

$P_\theta(x_1, \ldots, x_n) = -\log_2 Pr(x_1, \ldots, x_n | f_\theta) = \sum(-\log_2 Pr(x_i | f_\theta(x_1, \ldots, x_{i-1})))$

↑ Log-perplexity

Low perplexity $\leftrightarrow$ high prob.

"Mary had a little lamb" ↑prob, ↓perp.

"Correct horse battery stapler" ↓prob, ↑perp.

"My social security number is 078-05-1120"

1. Memorize?
2. Find secrets?

"My SSN is 867-53-0900"

R = "My SSN is xxx-xy-xxxx"

Sort by perp.

rank: order in list

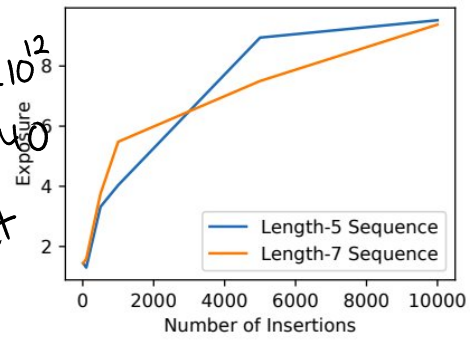
$r \in R$ exposure(r): $\log_2 |R| - \log_2 rank(r)$

Large exposure ↓ Least likely

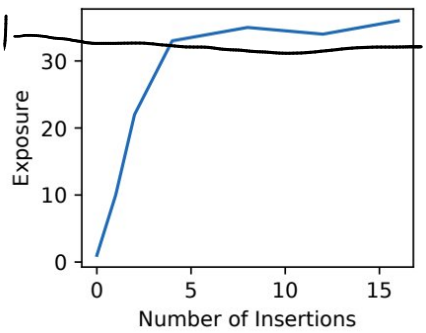
Most likely

$|R| \approx 10^{12}$

exposure ≈ 40 to extract



$\log_2 |R|$



Dropout, Regularization ✗

Differential Privacy ✓

Example 4: Membership Inference in Genomic Studies

Homer et al. 2008

NIH
- maf., χ^2-stats, p-values
- Restricted

Privacy ⇒ open science

Example 5: Massachusetts Group Insurance Commission

Sweeney (~~name~~, ~~SSN~~, ZIP, d.o.b, sex, condition)

87% of US identified

voter rolls $20

(name, ZIP, DoB, sex, ~~party~~)

k-anonymity [Samarati-Sweeney '98]

(~~name~~, ZIP, DoB, sex, cond)

pseudo identifiers

sensitive

A dataset is *k-anonymous* if, for any setting of the pseudo-identifiers, there are at least $k-1$ other points with the same settings of the pseudo-identifiers.

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

35-years old

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

28-years old