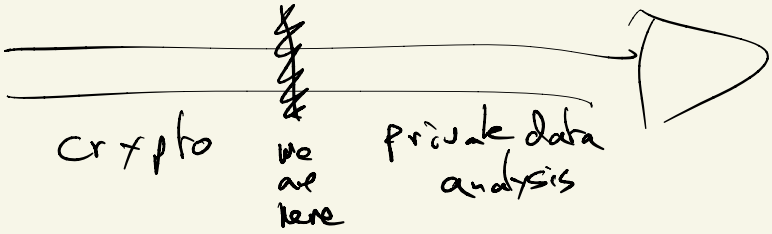
Privacy Enhancing Technologies

Data Anonymization

Florian Tramèr

We are here:



<u>One def of privacy</u>: we can compute

$y = f(x_1, \ldots, x_n)$

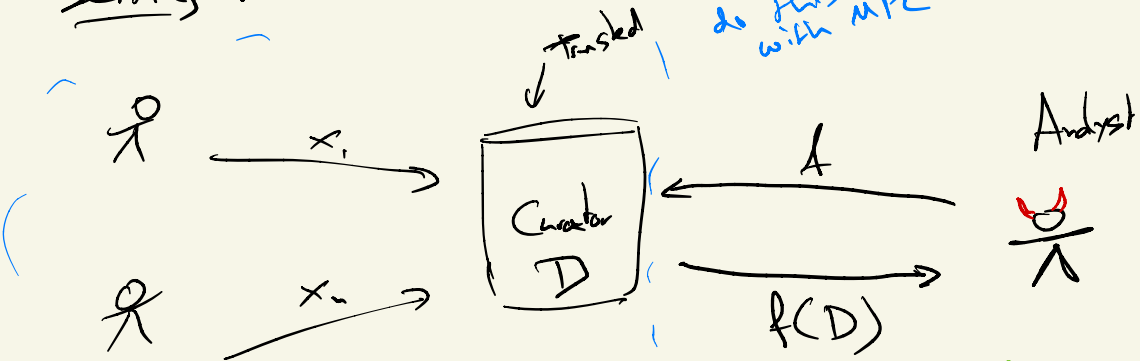
while leaking <u>nothing</u> other than y.

<u>Question</u>: what if leaking y is bad?

dummy example: let $f(x_1, \ldots, x_n) = x_1$

even with malicious MPC / FHE / ZK / ... Party P_1 doesn't like this function ...

Setting:

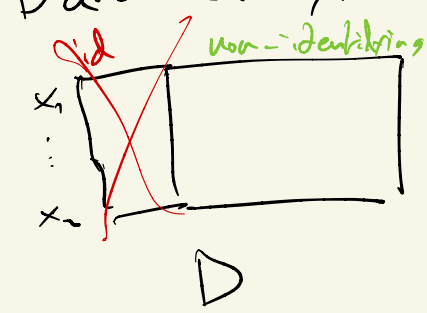


$D = \{x_1, \ldots, x_n\}$ — leaks info about D — utility ?

Extremes:

- Curator always outputs $\perp$ → not useful / → very private
- Analyst asks for $f(D) = D$ → most useful / → not private

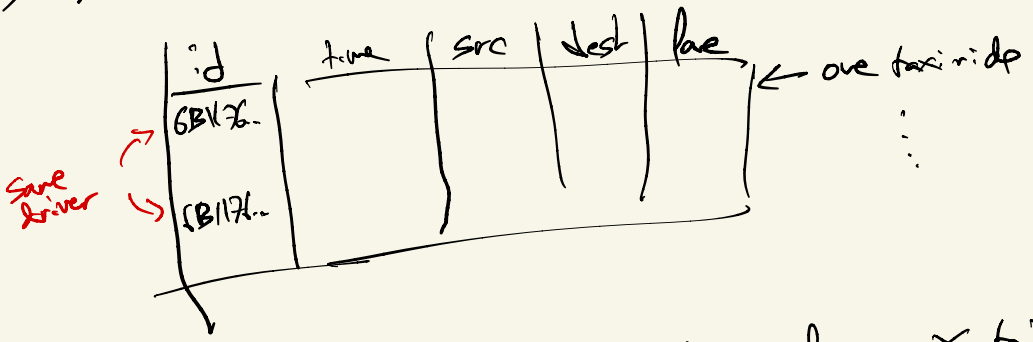
Data anonymization



$\Rightarrow$ release

How anonymization fails in practice

1) NYC Taxi data



issues: * if I take a taxi from X to Y and pay 23$, I can then find this row in the data, and figure out all other rides that this driver took

* someone looked at the data and found id CFCD ... 4DA with way too many rides !?!?!

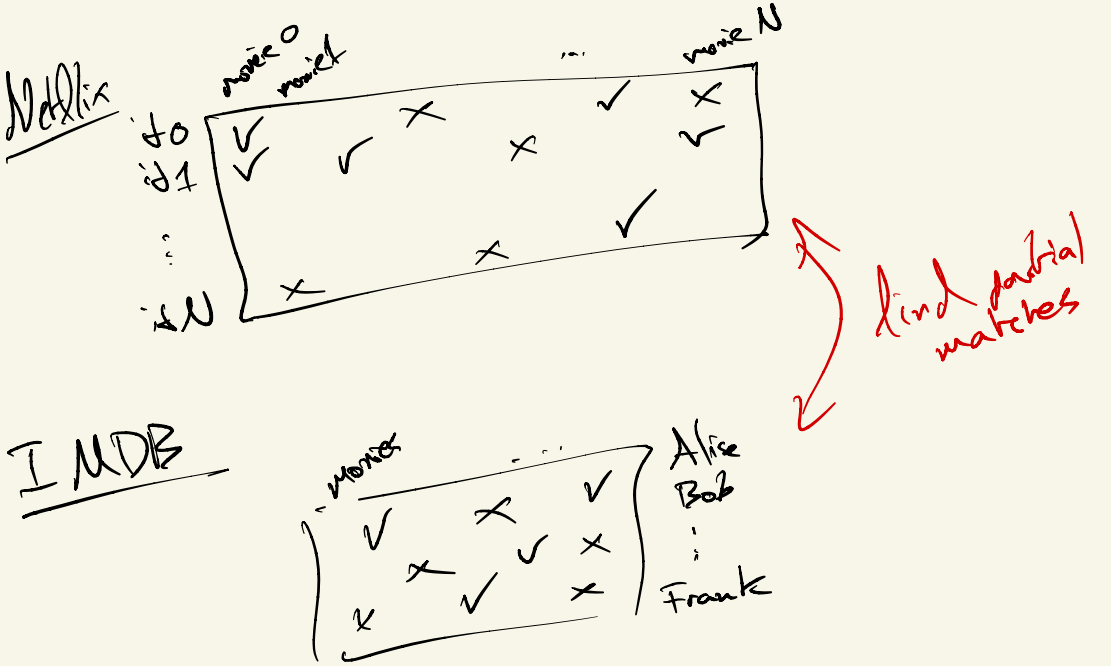
→ MD5("0")

This is not hiding?

if I know a driver has id 12345, just look for H(12345) in the data...

2) Netflix prize

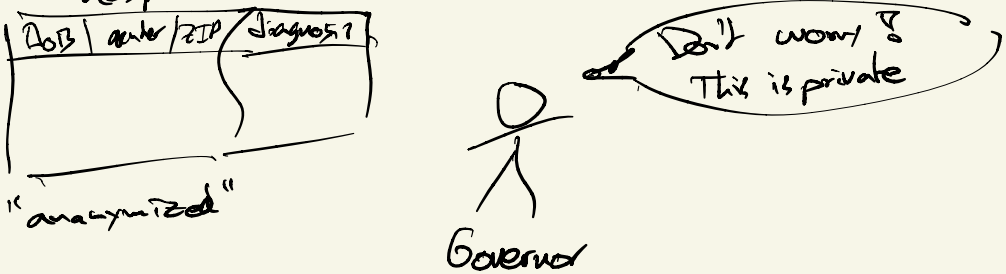
Netflix released a dataset of users' movie ratings



---

3. Massachusetts insurance

hospital records for state employees



Latanya Sweeney: $365 \cdot 78 \cdot 2 \approx 60k$ combinations for 1st-2nd release

a US ZIP code has 25k people on avg

⚠ 87% of the US population can be uniquely identified from DoB, gender, Zip

---

K-anonymity (Sweeney 2002):

idea: in a dataset, each person's features should be indistinguishable from at least k-1 other people in the data

Example

2-anonymous dataset

identifiers	quasi-identifiers				sensitive target
Name	Age	Gender	Weight	Zip-code	Disease
x	20-55	F	50-60	12***	Heart
x	20-55	F	50-60	12***	Flu
⋮	30	*	60-100	*	Cancer
	30	*	60-100	*	Cancer
x					

Attacks against k-anonymity

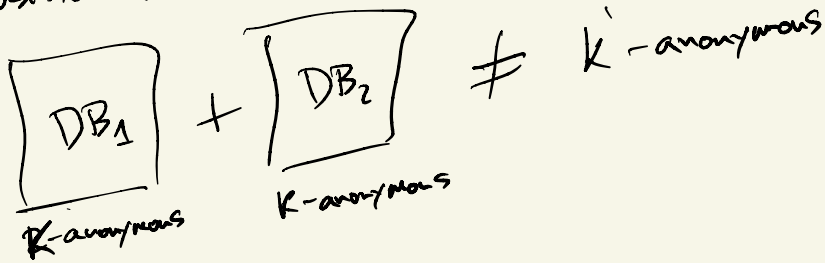
- Homogeneity attack

Name	Age	Gender	Weight	Zip code	Disease
x	20-55	F	50-60	12***	Heart
x	20-55	F	50-60	12***	Flu
⋮	30	*	60-80	*	Cancer
	30	*	60-80	*	Cancer
x					

I know that John is 30, male, 80kg, lives in ... What condition does John have?

- Background knowledge: e.g. I know my friend is in the DB and is 20, F, and does not have a Flu

- composition attack:



- Downcoding attack: infer what the data was based on k-anon output and knowledge of the algo

ex:

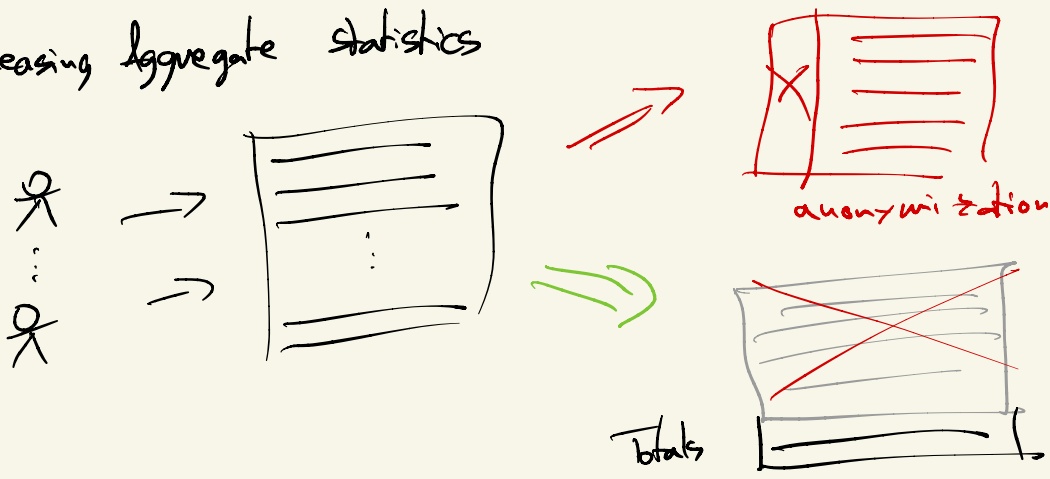
20 - 40	F	12***	Heart
20 - 40	F	12***	Flu

↳ at least one of these patients is > 35 years old

Conclusions : releasing "anonymized" data is dangerous

↳ often doesn't provide meaningful privacy

Releasing aggregate statistics



Even releasing only aggregates can be dangerous!

(US) Census

	Count	Age	
		mean	median
Total pop	7	30	38
F	4	30	33.5
M	3	30	44
white	3		
Black & F	3		
Black & M	— ?		

↓↓

This is a 1...

Ex of information leakage from simple statistics

	Count	Median age	Avg age
M	3	30	44

let's say the ages are $A \leq B \leq C$ for $A,B,C \subseteq [1,120]$

$A < 30$, $B = 30$, $C > 30$

↳ we know $B = 30$

↳ we know $A + C = 62$, $A < 30$, $C > 30$

Reconstruction attacks: How many aggregate stats can we release before privacy breaks down?

Formal model: Database of n rows

Public attributes: Name	Age	Gender	Private attr.: Disease?
Alice	25	F	1
Bob	30	M	0
...			
Zack	18	M	1

(Known to the adv)

$\Rightarrow$

Identifier	Secret
z_1	$s_1 \in \{0,1\}$
z_2	s_2
$\vdots$	$\vdots$
z_n	s_n

Adversary gets to make "statistical queries" to this DB

ex: . How many women under 25 have the disease?

. How many people called Bob have the disease?

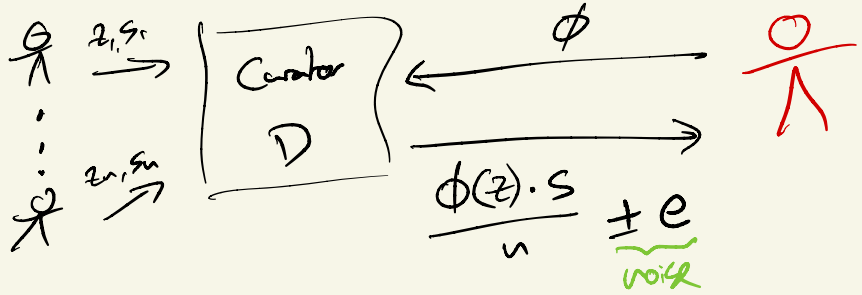
SQL: SELECT SUM(Disease) where ... (public att.)

Formally: the adv makes "counting queries" of the form

$$f_\phi(D) = \frac{1}{n}\sum_{i=1}^{n} \phi(z_i)\cdot s_i \quad , \text{ where } \phi: \mathcal{Z} \to \{0,1\}$$

$$= \frac{1}{n}\,\phi(\vec{z})\cdot\vec{s}$$

- The curator answer queries with some noise



- e is the noise, $0 \le e \le E$, arbitrary
- $Ans(\phi) = f_\phi(D) \pm e$

Observation: if there is no noise ($E = 0$), then the adv can recover the entire database ($\vec{s}$) by making n queries

How: $\phi_1(z) = \begin{cases} 1 & \text{if } z = z_1 \\ 0 & \text{otherwise} \end{cases}$

$\vdots$

$\phi_n(z) = \begin{cases} 1 & \text{if } z = z_n \\ 0 & \text{otherwise} \end{cases}$

A "defense": only answer queries if ϕ holds for at least $n/2$ rows

$\Rightarrow$ A "differentiating" attack

$$\phi_1(z) = [1, 1, 1, \ldots, 1] \quad \checkmark$$

$$\phi_2(z) = [1, 1, 1, \ldots, 1, 0] \quad \checkmark$$

$$A_{\phi_1}(D) - A_{\phi_2}(z) = s_n/n$$ ☹

Observation: with no noise, we can't do much

What if we have noise?

Def "Blatant non privacy": A curator's algorithm is blatantly non private if the adv can with high probability $(1-o(1))$, reconstruct a database $\tilde{s} \in \{0,1\}^n$ such that s and $\tilde{s}$ differ in $o(n)$ coordinates

---

Theorem 1 (Dinur & Nissim 2003):

If the adv can make 2^n subset queries on a DB of size n, then they can output a guess $\tilde{s} \in \{0,1\}^n$ such that s and $\tilde{s}$ differ in at most 4% of n positions

ex 1 : . we set $E = \frac{1}{401}$

- any query that counts less than $1/100$ th of the data, gets no utility
- an adv can reconstruct $\tilde{s}$ that differs from s in $4En$ positions ($n/100$) $\Rightarrow$ 99% of the rows

ex 2: if we set $E = o(1)$ (for example $E = \frac{1}{n}$) then the curator is blatantly non private

---

Proof of Thm:

the Adv's attack: 1) for each $v \in \{0,1\}^n$, make a query with $\phi_v(\vec{z}) = v$ and obtain answer $Ans(\phi_v) = \frac{\phi_v(z) \cdot s}{n} \pm e$

2) Output a guess $\hat{s}$ that is consistent with the queries

$$\left| \frac{\phi_v(z) \cdot \tilde{s}}{n} - Ans(\phi_v) \right| \leq E \quad \forall v \in \{0,1\}^n$$

1) this algorithm always outputs a database $\tilde{s}$

why? because the true database s is of course consistent with all answers

2) the guess output by the adversary is close to the true vector s

$$\Rightarrow \text{dist}(s, \hat{s}) = \sum_{i=1}^{n} \mathbb{1}\{s_i \neq \tilde{s}_i\} \leq 4En$$

proof by contradiction: suppose $\text{dist}(s, \hat{s}) > 4En$

we split these errors into two parts

$$S_0 = \{i \mid s_i = 0, \hat{s}_i = 1\} \qquad S_1 = \{i \mid s_i = 1, \hat{s}_i = 0\}$$

we have that $|S_0| + |S_1| = \text{dist}(s, \hat{s}) > 4En$

w.l.o.g assume that $|S_0| > 2En$

Denote by v the "indicator vector" for S_0

↳ $v_i = 1$ iff $i \in S_0$

- we have $f_{\phi_v}(D) = \frac{\phi_v(z) \cdot s}{n} = 0$

  and so $\text{Acc}(\phi_v) \leq E$

- But, $\frac{\phi_v(z) \cdot \tilde{s}}{n} = \frac{|S_0|}{n} > 2E$

$\hat{s}$ is not consistent!

Thm^2: If adv can make n queries and the noise is bounded by $E = o(1/n)$ then the curator is blatantly non private

Corollary: if an adv can make n queries you need noise $E = \Omega(1/n)$ for privacy to be possible

Proof sketch: Adv makes n random queries (where ϕ_v has $v \xleftarrow{\$} \{0,1\}^n$) and pick a database that is consistent with the answers $\Rightarrow$ can be done in poly(n) time