THE UNIVERSITY OF MAINE

PERC_LAB
Home of Privacy Engineering – Regulatory Compliance Research

EACL 2023 Tutorial on Privacy-Preserving NLP

Block 1: Attacks

Membership Inference, Motivation & Threats

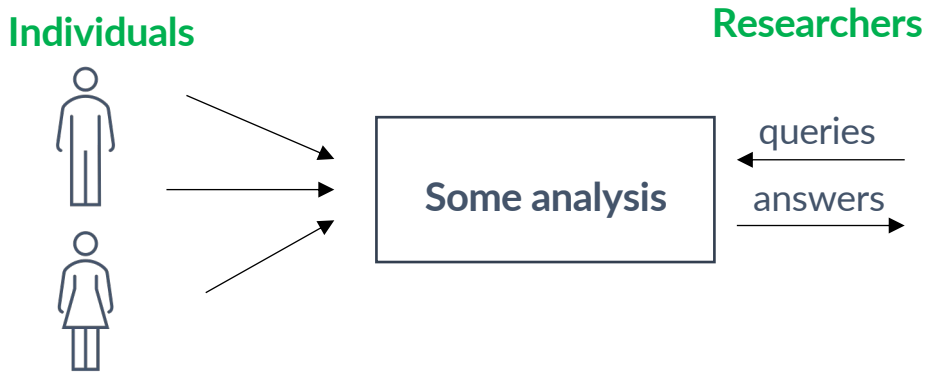
Dr. Sepideh Ghanavati

May 6th, 2023

https://www.sepidehghanavati.com/

Privacy in Statistical Database

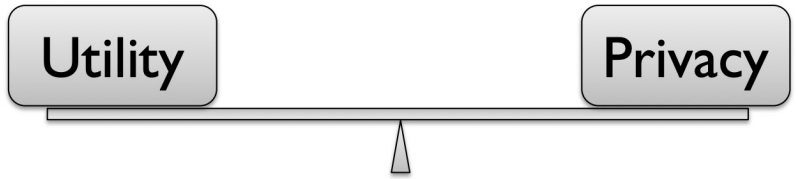
- Large collection of PI:
  - Census data
  - Medical/public health
  - Social networks
  - Education



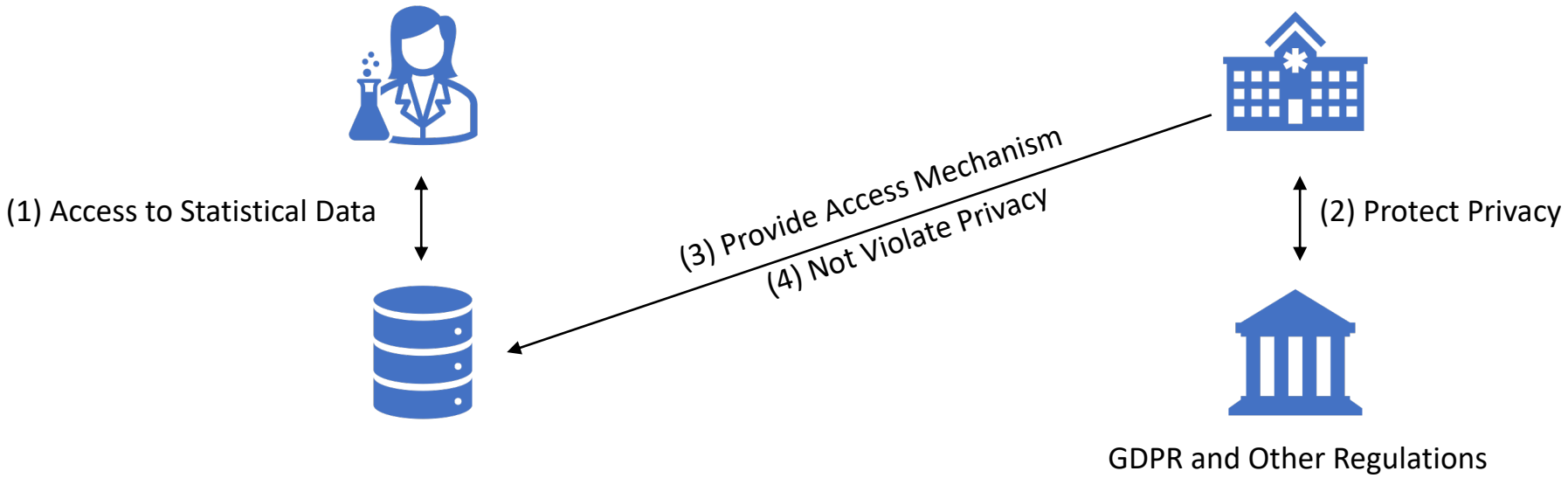
- Statistical analysis benefit society!
- They are valuable because they reveal so much about our lives.

Two Conflicting Goals

- **Utility:**
  - Release aggregate statistics.
- **Privacy:**
  - Individual information stays hidden.



Motivation


(1) Access to Statistical Data
(3) Provide Access Mechanism
(4) Not Violate Privacy
(2) Protect Privacy
GDPR and Other Regulations

Simple Solution

- Remove all *identifying* attributes from the database!
- Examples:
  - Patient's name, social security number, date of birth

Simple Solution

- Remove all *identifying* attributes from the database!
- Examples:
  - Patient's name, social security number, date of birth
- What is the problem?

Simple Solution

- Remove all *identifying* attributes from the database!
- Examples:
  - Patient's name, social security number, date of birth
- What is the problem?
  - Can de-identify users via *indirectly identifying* attributes! → **Quasi Identifiers**

Other Simple Solutions

Query restrictions:

- **Limit queries** – Queries are required to obey a special structure to prevent an adversary from gaining too much information about the specific dataset entries.
- **Query auditing** – Allowing/disallowing queries.

Other Simple Solutions

Query restrictions:

- **Limit queries** – Queries are required to obey a special structure to prevent an adversary from gaining too much information about the specific dataset entries.
- **Query auditing** – Allowing/disallowing queries.

Data/Output perturbation

- **Data perturbation** – Queries are answered according to a perturbed dataset.
- **Output perturbation** – The dataset first computes an exact answer but then releases a noisy answer.

Linkage Attacks

- Anonymizing data is non-trivial...
  - Massachusetts governor’s data identified from a public release of anonymized healthcare data by linking it to public voter registration lists.
  - Netflix data re-identified by linking it to IMDB data.
  - AOL search data re-identified without linking because it wasn’t fully anonymized.

Reconstruction Attack

Discussed by Dinur and Nissim in 2003

- How algorithms can be blatantly non-private!

Non-private dataset:

- If a computationally bounded adversary can reveal 1- ε fraction of the dataset, for all $\varepsilon > 0$.

Simple Example:

- Difference attacks

Difference Attack – Example I

Difference attack:

- Describes how aggregate information can reveal a lot about an individual!

An example:

- Releasing an account.

A simple solution:

- If there is no real aggregate and the answer is too small → Do not give the exact number and just give a range (example: < 5)

* Simplest possible example
  ↳ releasing an account

Question: How many employees were born on Jan 16th 84 and live in zipcode 04415 and have agoraphobia.

A: Let's say answer is 1.

⇒ In this case even questions that are not supposed to reveal info about the individual can reveal a lot.

Difference Attack – Example II

Difference attack shows that:

- Even aggregated statistics that seems to successfully aggregate a lot of individuals can single out one of them because the difference between two queries might narrow down to one person.

Example:

- **Query 1:** How many employees joined before 09/01/21 and suffered from agoraphobia?
  - Answer 1: 317
- **Query 2:** How many employees joined before 09/02/21 and suffered from agoraphobia?
- Answer 2: 318

Reconstruction Attack – Census Example I

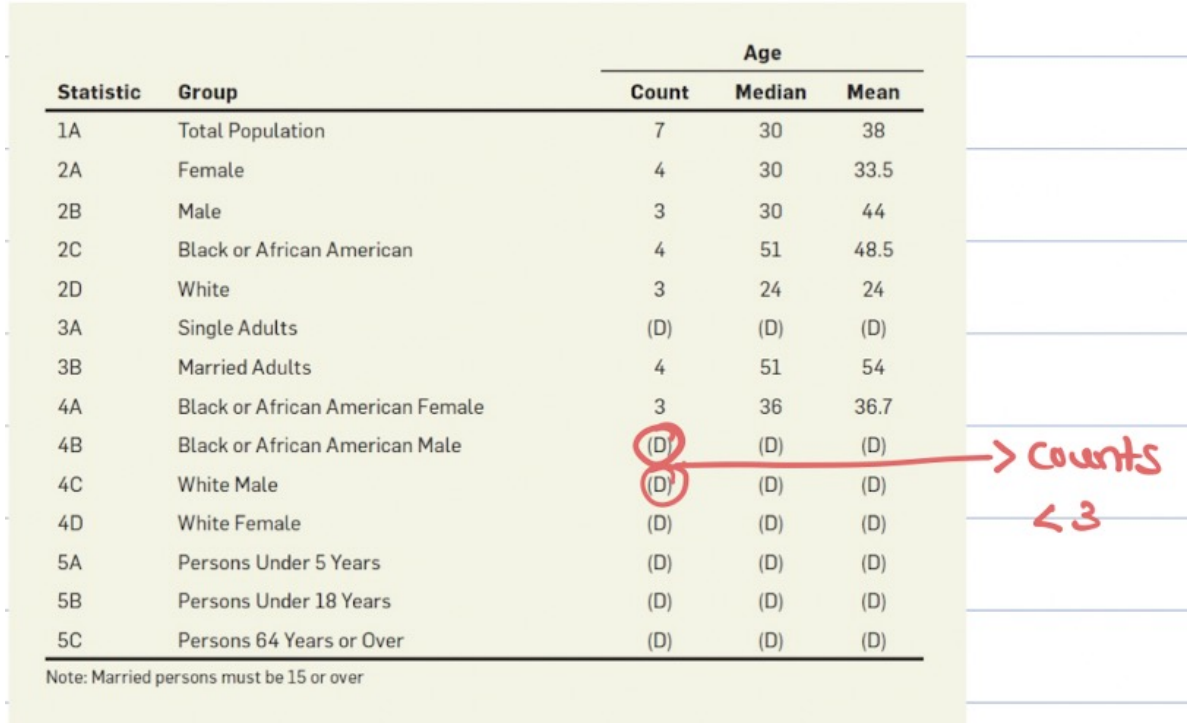
Let's talk about the census example…

- Federal law prohibits the releasing of individual responses, which is called *microdata*.
- *Microdata example:* age, race, sex

- Can we protect privacy?
  - If we do not let users select queries?
  - If we provide summary statistics of the dataset for the public and do not allow the public to choose suspicious queries to give answers.

		Age		
Statistic	**Group**	**Count**	**Median**	**Mean**
1A	Total Population	7	30	38
2A	Female	4	30	33.5
2B	Male	3	30	44
2C	Black or African American	4	51	48.5
2D	White	3	24	24
3A	Single Adults	(D)	(D)	(D)
3B	Married Adults	4	51	54
4A	Black or African American Female	3	36	36.7
4B	Black or African American Male	(D)	(D)	(D)
4C	White Male	(D)	(D)	(D)
4D	White Female	(D)	(D)	(D)
5A	Persons Under 5 Years	(D)	(D)	(D)
5B	Persons Under 18 Years	(D)	(D)	(D)
5C	Persons 64 Years or Over	(D)	(D)	(D)

Note: Married persons must be 15 or over



Source: By Simson Garfinkel, John M. Abowd, Christian Martindale, 2019.

Reconstruction Attack – Census Example II

- 3 males
- ages $A \leq B \leq C$
- $1 \leq A, B, C \leq 125$
  - ↳ assume no one is below 1
  - ↳ The oldest person in the US.

Statistic	Group	Age		
		Count	Median	Mean
1A	Total Population	7	30	38
2A	Female	4	30	33.5
2B	Male	3	30	44
2C	Black or African American	4	51	48.5
2D	White	3	24	24
3A	Single Adults	(D)	(D)	(D)
3B	Married Adults	4	51	54
4A	Black or African American Female	3	36	36.7
4B	Black or African American Male	(D)	(D)	(D)
4C	White Male	(D)	(D)	(D)
4D	White Female	(D)	(D)	(D)
5A	Persons Under 5 Years	(D)	(D)	(D)
5B	Persons Under 18 Years	(D)	(D)	(D)
5C	Persons 64 Years or Over	(D)	(D)	(D)

Note: Married persons must be 15 or over

Source: By Simson Garfinkel, John M. Abowd, Christian Martindale, 2019.

Reconstruction Attack – Census Example II

- 3 males
- ages $A \leq B \leq C$
- $1 \leq A, B, C \leq 125$
  - ↳ assume no one is below 1
  - ↳ the oldest person in the US.

⇒ possible combination: $\binom{125}{3}$

↳ 317,750 possibilities.

		Age		
Statistic	**Group**	**Count**	**Median**	**Mean**
1A	Total Population	7	30	38
2A	Female	4	30	33.5
2B	Male	3	30	44
2C	Black or African American	4	51	48.5
2D	White	3	24	24
3A	Single Adults	(D)	(D)	(D)
3B	Married Adults	4	51	54
4A	Black or African American Female	3	36	36.7
4B	Black or African American Male	(D)	(D)	(D)
4C	White Male	(D)	(D)	(D)
4D	White Female	(D)	(D)	(D)
5A	Persons Under 5 Years	(D)	(D)	(D)
5B	Persons Under 18 Years	(D)	(D)	(D)
5C	Persons 64 Years or Over	(D)	(D)	(D)

Note: Married persons must be 15 or over

Source: By Simson Garfinkel, John M. Abowd, Christian Martindale, 2019.

Reconstruction Attack – Census Example III

- median : 30 ⇒ B = 30

  A < 30

  C > 30

  ⌐> 30

- mean : 44 ⇒ $\frac{A+B+C}{3} = 44$

		Age		
Statistic	**Group**	**Count**	**Median**	**Mean**
1A	Total Population	7	30	38
2A	Female	4	30	33.5
2B	Male	3	30	44
2C	Black or African American	4	51	48.5
2D	White	3	24	24
3A	Single Adults	(D)	(D)	(D)
3B	Married Adults	4	51	54
4A	Black or African American Female	3	36	36.7
4B	Black or African American Male	(D)	(D)	(D)
4C	White Male	(D)	(D)	(D)
4D	White Female	(D)	(D)	(D)
5A	Persons Under 5 Years	(D)	(D)	(D)
5B	Persons Under 18 Years	(D)	(D)	(D)
5C	Persons 64 Years or Over	(D)	(D)	(D)

Note: Married persons must be 15 or over

Source: By Simson Garfinkel, John M. Abowd, Christian Martindale, 2019.

Reconstruction Attack – Census Example IV

- median : 30 ⇒ B = 30

  A ≤ 30

  C ≥ 30

  ↦ 30

- mean : 44 ⇒ $\frac{A+B+C}{3} = 44$

⇒ $A + C = 3 \cdot 34 = 102$

A	B	C	A	B	C	A	B	C
1	30	101	11	30	91	21	30	81
2	30	100	12	30	90	22	30	80
3	30	99	13	30	89	23	30	79
4	30	98	14	30	88	24	30	78
5	30	97	15	30	87	25	30	77
6	30	96	16	30	86	26	30	76
7	30	95	17	30	85	27	30	75
8	30	94	18	30	84	28	30	74
9	30	93	19	30	83	29	30	73
10	30	92	20	30	82	30	30	72

Source: By Simson Garfinkel, John M. Abowd, Christian Martindale, 2019.

Reconstruction Attack – Census Example V

- Statistic 1A:
  - Shows the universe of the constraint system.
  - There are 7 people.
  - Because the mean age is 38, we know that:

$$A1+A2+A3+A4+A5+A6+A7=7\times38$$

  - We can encode the constraints and use SAT solver to solve.

		Age		
Statistic	**Group**	**Count**	**Median**	**Mean**
1A	Total Population	7	30	38
2A	Female	4	30	33.5
2B	Male	3	30	44
2C	Black or African American	4	51	48.5
2D	White	3	24	24
3A	Single Adults	(D)	(D)	(D)
3B	Married Adults	4	51	54
4A	Black or African American Female	3	36	36.7
4B	Black or African American Male	(D)	(D)	(D)
4C	White Male	(D)	(D)	(D)
4D	White Female	(D)	(D)	(D)
5A	Persons Under 5 Years	(D)	(D)	(D)
5B	Persons Under 18 Years	(D)	(D)	(D)
5C	Persons 64 Years or Over	(D)	(D)	(D)

Note: Married persons must be 15 or over

Source: By Simson Garfinkel, John M. Abowd, Christian Martindale, 2019.

Reconstruction Attack I

- Assume we have a dataset X: *Microdata*

- statistics $f_1 \dots f_K$ ⇒ "the stats/queries/counts"
- answers ⇒ that are possibly noisy

$a_1 \approx f_1(X)$
$a_2 \approx f_2(X)$
$\vdots$
$a_K \approx f_3(X)$

⇒ Constraints ⇓ That narrow down the choices for X!

Reconstruction Attack II

- To conduct the attack:
  - Generate a set of constraints.
  - Find feasible points.

Given constraints $(f_i(x) \approx a_i)$ find a dataset $\tilde{x}$ that is consistent with the constraints.

→ This is called the *"constraint satisfaction problem"* → It is generally an NP-hard problem in the worst case **BUT**

→ With modern SAT solvers or with integer programming solvers, you can solve the problem easily.

In the census example, SAT solver can solve the constraints in 0.07 seconds.

Reconstruction Attack – Census Example V

Garfinkel et al. conducted a real attack.

They were able to construct a microdata dataset.

- 46% of data (~150 million Americans) were matched to the actual dataset.
- By adding some range, for example, +/- 1 years of age, they could match 71% of the population with the actual dataset.
- Together with a *linkage attack*, they could get an exact match with the name and sex of 50 million Americans.

Dinur – Nissim – Formal Definition of Statistical Dataset

- Define the statistical dataset as:

$$(d_1, \dots, dn) \in \{0,1\}^n$$

- A query, $s \subseteq [n]$, is answered by $A(s)$.
  - We allow the dataset to perturb its answers in such a way that hides values of specific bits, but still yields meaningful information about sums of bits.
- The (exact) answer to a query s is the sum of all database entries specified by s, i.e., $a_s = \sum_{i \in s} d_i$

Dinur – Nissim – Example of a Dataset

- A dataset with:
  - Rows – Datapoints
  - Columns – Features/dimensions
    - We assume that some of the columns are identifiers that are publicly available, such as (name, postal code, DoB, and sex) and we have one (*secret bit {0, 1}*)
  - Dataset: $d \in \{0, 1\}^n$

Name	Postal Code	Date of Birth	Sex	Has Disease?
Alice	K8V7R6	5/2/1984	F	1
Bob	V5K5J9	2/8/2001	M	0
Charlie	V1C7J	10/10/1954	M	1
David	R4K5T1	4/4/1944	M	0
Eve	G7N8Y3	1/1/1980	F	1

Source: Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy, 2014.

Dinur – Nissim – Reconstruction Attack I

- An answer $A(s)$ is within ε perturbation, if:

$$|a_s - A(s)| \leq \varepsilon$$

- We say that algorithm A is a within ε perturbation if for all queries $s \subseteq [n]$, the answer $A(s)$ is within ε perturbation.

Dinur – Nissim – Reconstruction Attack I – Example I

- Analyst – Trying to get an answer
- Curator – Respond to the answer – but *"private"*

Name	Postal Code	Date of Birth	Sex	Has Disease?
Alice	K8V7R6	5/2/1984	F	1
Bob	V5K5J9	2/8/2001	M	0
Charlie	V1C7J	10/10/1954	M	1
David	R4K5T1	4/4/1944	M	0
Eve	G7N8Y3	1/1/1980	F	1

Source: Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy, 2014.

Dinur – Nissim – Reconstruction Attack I – Example II

- Analyst – Trying to get an answer.
  - How many rows satisfy **(condition)** have **'has disease = 1'**?
  - Condition:
    - "Name = Alice OR Name = Bob OR Name = Eve"

Name	Postal Code	Date of Birth	Sex	Has Disease?
Alice	K8V7R6	5/2/1984	F	1
Bob	V5K5J9	2/8/2001	M	0
Charlie	V1C7J	10/10/1954	M	1
David	R4K5T1	4/4/1944	M	0
Eve	G7N8Y3	1/1/1980	F	1

Source: Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy, 2014.

Dinur – Nissim – Reconstruction Attack I – Example II

- Analyst – Trying to get an answer.
  - How many rows satisfy **(condition)** have **'has disease = 1'**?
  - Condition:
    - "Name = Alice OR Name = Bob OR Name = Eve"
  - True answer = 2

Name	Postal Code	Date of Birth	Sex	Has Disease?
Alice	K8V7R6	5/2/1984	F	1
Bob	V5K5J9	2/8/2001	M	0
Charlie	V1C7J	10/10/1954	M	1
David	R4K5T1	4/4/1944	M	0
Eve	G7N8Y3	1/1/1980	F	1

Source: Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy, 2014.

Dinur – Nissim Database Reconstruction Attack II

- Analyst – Trying to get an answer.
- The analyst can specify queries that are a subset of [n].
- *s* is a query vector as follows:

Queries: $S \subseteq [n]$, $S \in \{0,1\}^n$: $\begin{cases} S_i = 1 & \text{if } i \text{ is in the subset} \\ 0 & \text{else} \end{cases}$

A subset query: $S = [1,1,0,0,1]$

- True answer:
  - A(s) = d . S;
  - [1,1,0,0,1] . [1,0,1,0, 1] = 2

Name	Postal Code	Date of Birth	Sex	Has Disease?
Alice	K8V7R6	5/2/1984	F	1
Bob	V5K5J9	2/8/2001	M	0
Charlie	V1C7J	10/10/1954	M	1
David	R4K5T1	4/4/1944	M	0
Eve	G7N8Y3	1/1/1980	F	1

Source: Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy, 2014.

Dinur – Nissim Database Reconstruction Attack III

- Curator – Respond to the answer – but ***"private"***
  - Receives s, responds r(s)
  - Options: r(s) = A(s) → Give the analyst the correct/exact answer
    - Good for the analyst but bad for privacy!
  - Another way:
    - Add some noises (arbitrary)
    - Return r(s) s.t. | r(s) – A(s) | ≪ E

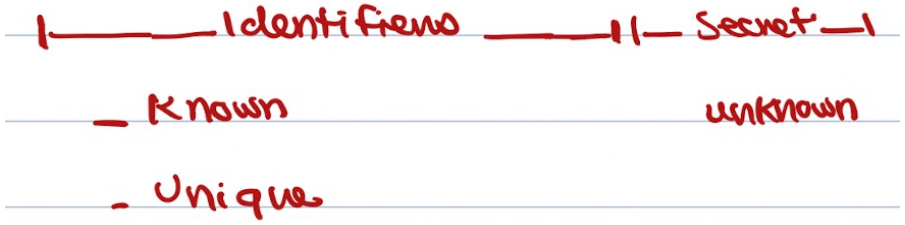
Name	Postal Code	Date of Birth	Sex	Has Disease?
Alice	K8V7R6	5/2/1984	F	1
Bob	V5K5J9	2/8/2001	M	0
Charlie	V1C7J	10/10/1954	M	1
David	R4K5T1	4/4/1944	M	0
Eve	G7N8Y3	1/1/1980	F	1

Source: Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy, 2014.

Dinur – Nissim Database Reconstruction Attack V

- Can the attacker who gets the identifiers and some statistics involving the secret, reconstruct the secret?
  - We know that the statistics is noisy.

Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1



Dinur – Nissim Database Reconstruction Attack V

- Can the attacker who gets the identifiers and some statistics involving the secret, reconstruct the secret?
  - We know that the statistics is noisy.

Name	Postal Code	Age	Sex	Has Disease?
Alice	02445	36	F	1
Bob	02446	18	M	0
Charlie	02118	66	M	1
⋮	⋮	⋮	⋮	⋮
Zora	02120	40	F	1

Identifiers	Secret
z_1	s_1
z_2	s_2
z_3	s_3
⋮	⋮
z_n	s_n

General Definition of Reconstruction Attack

Input: a set of query vectors $F_1, \ldots, F_k \in \{0,1\}^n$ and a set of answers $a_1, \ldots, a_k \in \mathbb{R}$
Output: a vector of secrets $\tilde{s} \in \{0,1\}^n$

Return $\tilde{s} \in \{0,1\}^n$ that *minimizes* the quantity $\max_{i \in [k]} |F_i \cdot \tilde{s} - a_v|$

General Definition of Reconstruction Attack

Input: a set of query vectors $F_1, \ldots, F_k \in \{0,1\}^n$ and a set of answers $a_1, \ldots, a_k \in \mathbb{R}$
Output: a vector of secrets $\tilde{s} \in \{0,1\}^n$

Return $\tilde{s} \in \{0,1\}^n$ that *minimizes* the quantity $\max_{i \in [k]} |F_i \cdot \tilde{s} - a_v|$

Claim:

If every query is answered to within error $\leq \alpha n$

i.e. $\max_{i \in [k]} |F_i \cdot s - a_i| \leq \alpha n$ $\quad (0 \leq \alpha \leq 1)$

then the reconstruction attack return $\tilde{s}$ that

$\max_{i \in [k]} |F_i \cdot s - a_v| \leq \alpha n$

Blatantly Non – Private Algorithm

Definition

- An algorithm is blatantly *non-private* if an adversary can construct a database $c \in \{0,1\}^n$ that it matches the true dataset d in all but $o(n)$ entries.
- The idea is that if an adversary construct the dataset which matches d on almost everything (e.g., 99% of datapoints), then we have a blatantly non-private algorithm → *blatant privacy violation*.

Two types of attacks

- There is no limit on the number of queries to ask → *Inefficient Attack*
- There is a limit on the number of queries to ask → *Efficient Attack*

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[Query Phase]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[Weeding Phase]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

Attack:

1. Analyst asks all possible 2^n subset queries: s=[1,0,0,..], s=[0,1,0,..], s=[0,0,1,..], and so on
2. For all $c \in \{1,0\}^n$:

   2.a Set S s.t. $|\sum c_i - r(s)| > E$

   if this equation is true: rule out c

   2.b Otherwise, output any c that is not ruled out and halt.

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[Query Phase]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[Weeding Phase]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

- It works for every secret.
- If the attackers get no statistics, they can guess only 50%.
- The attacker can't conduct a reconstruction attack without statistics.
- The attacker can't guess the data w/o having the data in the dataset. (i.e., won't know if a person has a disease or not if their data is not in the dataset.)

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[Query Phase]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[Weeding Phase]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

- It works for every secret.
- If the attackers get no statistics, they can guess only 50%.
- The attacker can't conduct a reconstruction attack without statistics.
- The attacker can't guess the data w/o having the data in the dataset. (i.e., won't know if a person has a disease or not if their data is not in the dataset.)

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[Query Phase]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[Weeding Phase]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

- It works for every secret.
- If the attackers get no statistics, they can guess only 50%.
- The attacker can't conduct a reconstruction attack without statistics.
- The attacker can't guess the data w/o having the data in the dataset. (i.e., won't know if a person has a disease or not if their data is not in the dataset.)

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[Query Phase]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[Weeding Phase]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

- It works for every secret.
- If the attackers get no statistics, they can guess only 50%.
- The attacker can't conduct a reconstruction attack without statistics.
- The attacker can't guess the data w/o having the data in the dataset. (i.e., won't know if a person has a disease or not if their data is not in the dataset.)

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[Query Phase]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[Weeding Phase]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

- It works for every secret.
- If the attackers get no statistics, they can guess only 50%.
- The attacker can't conduct a reconstruction attack without statistics.
- The attacker can't guess the data w/o having the data in the dataset. (i.e., won't know if a person has a disease or not if their data is not in the dataset.)

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[Query Phase]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[Weeding Phase]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

Proof

1. The algorithm always halts and outputs some *candidate* c, since the real database d is never weeded out.
2. We now show that the output candidate c satisfies $\text{dist}(d,c) \leq 4E = o(n)$.
3. Assume this is not the case, i.e., $\text{dist}(d,c) > 4E$.
4. Let $qo = \{i | d_i = 1, ci = 0\}$ and $q1 = \{i | d_i = 0, ci = 1\}$
5. Since $|q1| + |q0| = \text{dist}(d,c) > 4E$, at least one of the disjoint sets has size $2E + 1$ or more.
6. Assume $|q1| > 2E$.
7. We have that $\sum_{i \in qi} d_i = 0$. Hence, it must be that $\tilde{a}_{q1} \leq E$.
8. On the other hand, $\sum_{i \in qi} c_i = |q1| > 2E$
9. We get that $|\sum_{i \in qi} c_i - \tilde{a}_{q1}| > E$, which is contradicting the fact that c survives the weeding phase.

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[Query Phase]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[Weeding Phase]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

Proof

1. The algorithm always halts and outputs some *candidate* c, since the real database d is never weeded out.
2. We now show that the output candidate c satisfies $\text{dist}(d, c) \leq 4E = o(n)$.
3. Assume this is not the case, i.e., $\text{dist}(d, c) > 4E$.
4. Let $q0 = \{i | d_i = 1, ci = 0\}$ and $q1 = \{i | d_i = 0, ci = 1\}$
5. Since $|q1| + |q0| = \text{dist}(d, c) > 4E$, at least one of the disjoint sets has size $2E + 1$ or more.
6. Assume $|q1| > 2E$.
7. We have that $\sum_{i \in qi} d_i = 0$. Hence, it must be that $\tilde{a}_{q1} \leq E$.
8. On the other hand, $\sum_{i \in qi} c_i = |q1| > 2E$
9. We get that $|\sum_{i \in qi} c_i - \tilde{a}_{q1}| > E$, which is contradicting the fact that c survives the weeding phase.

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[Query Phase]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[Weeding Phase]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

Proof

1. The algorithm always halts and outputs some *candidate* c, since the real database d is never weeded out.
2. We now show that the output candidate c satisfies $\text{dist}(d,c) \leq 4E = o(n)$.
3. Assume this is not the case, i.e., $\text{dist}(d,c) > 4E$.
4. Let $q0 = \{i | d_i = 1, c_i = 0\}$ and $q1 = \{i | d_i = 0, c_i = 1\}$
5. Since $|q1| + |q0| = \text{dist}(d,c) > 4E$, at least one of the disjoint sets has size $2E + 1$ or more.
6. Assume $|q1| > 2E$.
7. We have that $\sum_{i \in qi} d_i = 0$. Hence, it must be that $\tilde{a}_{q1} \leq E$.
8. On the other hand, $\sum_{i \in qi} c_i = |q1| > 2E$
9. We get that $|\sum_{i \in qi} c_i - \tilde{a}_{q1}| > E$, which is contradicting the fact that c survives the weeding phase.

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[Query Phase]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[Weeding Phase]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

Proof

1. The algorithm always halts and outputs some *candidate* c, since the real database d is never weeded out.
2. We now show that the output candidate c satisfies $\text{dist}(d, c) \leq 4E = o(n)$.
3. Assume this is not the case, i.e., $\text{dist}(d, c) > 4E$.
4. Let $q0 = \{i | d_i = 1, c_i = 0\}$ and $q1 = \{i | d_i = 0, c_i = 1\}$
5. Since $|q1| + |q0| = \text{dist}(d, c) > 4E$, at least one of the disjoint sets has size $2E + 1$ or more.
6. Assume $|q1| > 2E$.
7. We have that $\sum_{i \in qi} d_i = 0$. Hence, it must be that $\tilde{a}_{q1} \leq E$.
8. On the other hand, $\sum_{i \in qi} c_i = |q1| > 2E$
9. We get that $|\sum_{i \in qi} c_i - \tilde{a}_{q1}| > E$, which is contradicting the fact that c survives the weeding phase.

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[Query Phase]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[Weeding Phase]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

Proof

1. The algorithm always halts and outputs some *candidate* c, since the real database d is never weeded out.
2. We now show that the output candidate c satisfies $\text{dist}(d, c) \leq 4E = o(n)$.
3. Assume this is not the case, i.e., $\text{dist}(d, c) > 4E$.
4. Let $q0 = \{i | d_i = 1, c_i = 0\}$ and $q1 = \{i | d_i = 0, c_i = 1\}$
5. Since $|q1| + |q0| = \text{dist}(d, c) > 4E$, at least one of the disjoint sets has size $2E + 1$ or more.
6. Assume $|q1| > 2E$.
7. We have that $\sum_{i \in qi} d_i = 0$. Hence, it must be that $\tilde{a}_{q1} \leq E$
8. On the other hand, $\sum_{i \in qi} c_i = |q1| > 2E$
9. We get that $|\sum_{i \in qi} c_i - \tilde{a}_{q1}| > E$, which is contradicting the fact that c survives the weeding phase.

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[Query Phase]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[Weeding Phase]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

Proof

1. The algorithm always halts and outputs some *candidate* c, since the real database d is never weeded out.
2. We now show that the output candidate c satisfies $\text{dist}(d,c) \leq 4E = o(n)$.
3. Assume this is not the case, i.e., $\text{dist}(d,c) > 4E$.
4. Let $q0 = \{i | d_i = 1, c_i = 0\}$ and $q1 = \{i | d_i = 0, c_i = 1\}$
5. Since $|q1| + |q0| = \text{dist}(d,c) > 4E$, at least one of the disjoint sets has size $2E + 1$ or more.
6. Assume $|q1| > 2E$.
7. We have that $\sum_{i \in qi} d_i = 0$. Hence, it must be that $\tilde{a}_{q1} \leq E$.
8. On the other hand, $\sum_{i \in qi} c_i = |q1| > 2E$
9. We get that $|\sum_{i \in qi} c_i - \tilde{a}_{q1}| > E$, which is contradicting the fact that c survives the weeding phase.

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[QUERY PHASE]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[WEEDING PHASE]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

Proof

1. The algorithm always halts and outputs some *candidate* c, since the real database d is never weeded out.
2. We now show that the output candidate c satisfies $\text{dist}(d,c) \leq 4E = o(n)$.
3. Assume this is not the case, i.e., $\text{dist}(d,c) > 4E$.
4. Let $q0 = \{i | d_i = 1, c_i = 0\}$ and $q1 = \{i | d_i = 0, c_i = 1\}$
5. Since $|q1| + |q0| = \text{dist}(d,c) > 4E$, at least one of the disjoint sets has size $2E + 1$ or more.
6. Assume $|q1| > 2E$.
7. We have that $\sum_{i \in qi} d_i = 0$. Hence, it must be that $\tilde{a}_{q1} \leq E$.
8. On the other hand, $\sum_{i \in qi} c_i = |q1| > 2E$
9. We get that $|\sum_{i \in qi} c_i - \tilde{a}_{q1}| > E$, which is contradicting the fact that c survives the weeding phase.

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[QUERY PHASE]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[WEEDING PHASE]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

Proof

1. The algorithm always halts and outputs some *candidate* c, since the real database d is never weeded out.
2. We now show that the output candidate c satisfies $\text{dist}(d,c) \leq 4E = o(n)$.
3. Assume this is not the case, i.e., $\text{dist}(d,c) > 4E$.
4. Let $q0 = \{i | d_i = 1, c_i = 0\}$ and $q1 = \{i | d_i = 0, c_i = 1\}$
5. Since $|q1| + |q0| = \text{dist}(d,c) > 4E$, at least one of the disjoint sets has size $2E + 1$ or more.
6. Assume $|q1| > 2E$.
7. We have that $\sum_{i \in qi} d_i = 0$. Hence, it must be that $\tilde{a}_{q1} \leq E$.
8. On the other hand, $\sum_{i \in qi} c_i = |q1| > 2E$
9. We get that $|\sum_{i \in qi} c_i - \tilde{a}_{q1}| > E$, which is contradicting the fact that c survives the weeding phase.

Inefficient Attack I

Theorem 1 (Dinur & Nissim 2003)

If the analyst is allowed to ask 2^n subset queries, and the curator adds noise with some bound E, then based on the results, the adversary can reconstruct the dataset in all but $4E$ positions.

[QUERY PHASE]
For all $q \subseteq [n]$: let $\tilde{a}_q \leftarrow \mathcal{A}(q)$.

[WEEDING PHASE]
For all $c \in \{0,1\}^n$: if $|\sum_{i \in q} c_i - \tilde{a}_q| \leq \mathcal{E}$ for all $q \subseteq [n]$ then output c and halt.

Proof

1. The algorithm always halts and outputs some *candidate* c, since the real database d is never weeded out.
2. We now show that the output candidate c satisfies $\text{dist}(d,c) \leq 4E = o(n)$.
3. Assume this is not the case, i.e., $\text{dist}(d,c) > 4E$.
4. Let $q0 = \{i | d_i = 1, c_i = 0\}$ and $q1 = \{i | d_i = 0, c_i = 1\}$
5. Since $|q1| + |q0| = \text{dist}(d,c) > 4E$, at least one of the disjoint sets has size $2E + 1$ or more.
6. Assume $|q1| > 2E$.
7. We have that $\sum_{i \in qi} d_i = 0$. Hence, it must be that $\tilde{a}_{q1} \leq E$.
8. On the other hand, $\sum_{i \in qi} c_i = |q1| > 2E$
9. We get that $|\sum_{i \in qi} c_i - \tilde{a}_{q1}| > E$, which is contradicting the fact that c survives the weeding phase.

Efficient Attack

Theorem 2 (Dinur & Nissim 2003). *If the analyst is allowed to ask O(n) subset queries, and the curator adds noise with some bound E = O($\alpha\sqrt{n}$), then based on the results, a computationally efficient adversary can reconstruct the database in all but O(α^2) positions.*

Attack:

1. Analyst asks random queries.
   S is chosen uniformly at random
   $\{0,1\}^n$
2. Find any database c consistent.
   ↳ use an LP (Linear Programming)

Summary of Reconstruction Attack

The attacker submits sufficiently random queries that link prior information to private data (which the attackers wants to learn)

The attacker receives noisy answers to the queries and writes them down as constraints for a linear program to solve for the private bits.

The attacker solves the linear program and rounds the result to recover most of the bits.

Summary of Reconstruction Attack

The attacker submits sufficiently random queries that link prior information to private data (which the attackers wants to learn)

The attacker receives noisy answers to the queries and writes them down as constraints for a linear program to solve for the private bits.

The attacker solves the linear program and rounds the result to recover most of the bits.

Summary of Reconstruction Attack

The attacker submits sufficiently random queries that link prior information to private data (which the attackers wants to learn)

The attacker receives noisy answers to the queries and writes them down as constraints for a linear program to solve for the private bits.

The attacker solves the linear program and rounds the result to recover most of the bits.

Recap – Reconstruction Theorems

We cannot hope to prevent reconstruction if we don’t put some limit on the number of statistics we reveal.

Linear Reconstruction Attack in Practice I

- Reconstruction attack on a commercial system called Diffix.
- Diffix is a system designed by a startup Aircloak for computing statistics on a private database.
- They mentioned that:
  - *“Anonymizing SQL interface [that] sits in front of your data and enables you to conduct ad hoc analytics— fully privacy preserving and GDPR-compliant.”*

Linear Reconstruction Attack in Practice II

- The goal of Diffix is to answer SQL queries such as:

```
SELECT count(*) FROM loans
WHERE loanStatus = 'C'
AND clientId BETWEEN 2000 and 3000
```

- While preventing the disclosure of individuals' records.
- Diffix uses a variety of heuristics for adding noise to the answers, but the noise doesn't satisfy formal privacy guarantees (i.e., differential privacy).
- Diffix includes three components to prevent averaging attacks by adding noises and attacks on syntactically distinct but semantically-equivalent queries.

Linear Reconstruction Attack in Practice III

- However, Cohen-Nissim in 2018 showed that Diffix is susceptible to reconstruction attacks.
- Consider random queries such as *"randomly"* choosing whether to add each `clientId` to the query as follows:

```
SELECT count(*) FROM loans
WHERE loanStatus = 'C'
AND (clientId = 2007
OR clientId = 2018
...
OR clientId = 2991)
```

- In Diffix, to avoid reconstruction attack, they have a system that adds noise to the answer of queries with the square root of *the number of terms* in the query (i.e., $\Omega(\sqrt{n})$)

Linear Reconstruction Attack in Practice IV

- Cohen-Nissim attack (2018) shows that if you form a query with a low number of *"conditions"* or *"terms"* such as:

```
SELECT COUNT(clientId) FROM loans
WHERE FLOOR(100 * ((clientId * 2)^0.7))
    = FLOOR(100 * ((clientId * 2)^0.7) + 0.5)
AND clientId BETWEEN 2000 and 3000
AND loanStatus = 'C'
```

- The reason is that these types of queries do not require many terms to specify, so Diffix answers the queries with a very small noise in the order of $\Omega(1)$ → Victim of full reconstruction attack.

- Diffix was changed so that it does not allow mathematical operations on attributes that are unique for each record, such as `clientId`.

Key Takeaways

- Merely releasing *"aggregate"* statistics can reveal information about individuals because of small counts and difference attacks.

- Releasing aggregate statistics imposes constraints on the underlying data, and releasing many such statistics, even with the noise, likely allows an attacker to reconstruct all or part of the dataset.

- Releasing even a linear number of sufficiently rich statistics with noise $O(\sqrt{n})$ will allow the reconstruction of the dataset.

Key Takeaways

- Merely releasing *"aggregate"* statistics can reveal information about individuals because of small counts and difference attacks.
- Releasing aggregate statistics imposes constraints on the underlying data, and releasing many such statistics, even with the noise, likely allows an attacker to reconstruct all or part of the dataset.
- Releasing even a linear number of sufficiently rich statistics with noise $O(\sqrt{n})$ will allow the reconstruction of the dataset.

Key Takeaways

- Merely releasing *"aggregate"* statistics can reveal information about individuals because of small counts and difference attacks.
- Releasing aggregate statistics imposes constraints on the underlying data, and releasing many such statistics, even with the noise, likely allows an attacker to reconstruct all or part of the dataset.
- Releasing even a linear number of sufficiently rich statistics with noise $O(\sqrt{n})$ will allow the reconstruction of the dataset.

Resources I

- Aloni Cohen and Kobbi Nissim. Linear program reconstruction in practice. The Journal of Privacy and Confidentiality, 10(1), 2020.
- Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '03, pages 202-210, New York, NY, USA, 2003. ACM.
- Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. Chapter 8, Foundations and Trends® in Theoretical Computer Science, 2014.
- Simson Garfinkel, John M Abowd, and Christian Martindale. Understanding database reconstruction attacks on public data. Communications of the ACM, 62(3):46-53, 2019.
- Cynthia Dwork, Frank McSherry, and Kunal Talwar. The price of privacy and the limits of lp decoding. In Proceedings of the 39th Annual ACM Symposium on the Theory of Computing, STOC '07, pages 85-94, New York, NY, USA, 2007. ACM.

Resources II

- John Abowd. Tweetorial: Reconstruction-abetted re-identification attacks and other traditional vulnerabilities, April 2019. https://twitter.com/john_abowd/status/1114942180278272000
- Theory of Reconstruction Attacks: https://differentialprivacy.org/reconstruction-theory/
- The 7th BIU Winter School: Reconstruction Attacks- Jon Ullman: https://www.youtube.com/watch?v=5_lTDPYOUJw