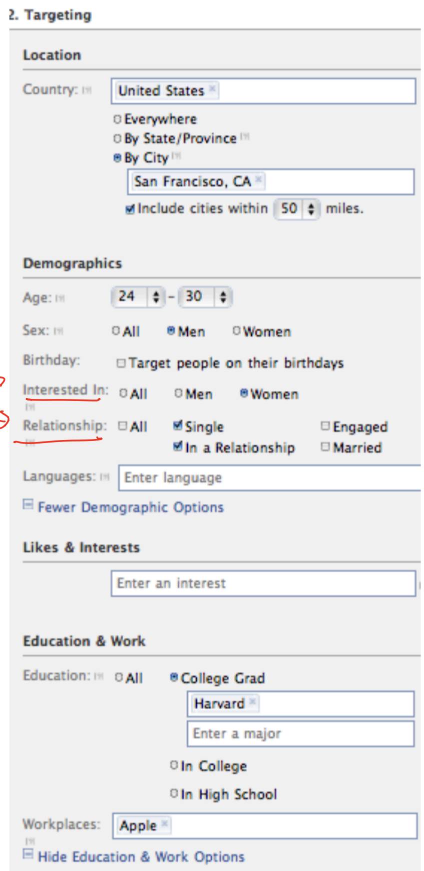
Lecture 2. Reconstruction Attacks.

- De-identified data ✗ ; releasing "Aggregate" Statistics?
- Warmup : Difference Attacks
- Reconstruction examples
- Reconstruction Formulation
  Linear Attacks [Dinur & Nissim 03]



Facebook ad campaign targeting interface.

Ref: Korolova,
"Privacy violation Using Microtargeted Ads: A Case Study"

Warmup : Difference Attacks

Q: How many people were born in 1992 and live in Zipcode 15206 and have a heart disease?

A: ~~1~~ less than 5.

Q: How many faculty members @ CMU joined before 9/1/2020 and have had a heart disease?

A: 37

Q: How many faculty members @ CMU joined before 9/2/2020 and have had a heart disease?

A: 38

Reconstruction in the US Census.

- 3 Males
- Ages $A \le B \le C$
- $1 \le A \le B \le C \le 125$
- Median $= 30$.

  $B = 30$

  $\{\boxed{A \le 30}\}$ $\quad C \ge 30$.
- Mean $= 44$.

  $\frac{A+B+C}{3} = 44$.

$\Rightarrow A + C = 102$.

(A, C) has 30 possibilities.

Before: $(125)^3$ possibilities

TABLE 1: **FICTIONAL STATISTICAL DATA FOR A FICTIONAL BLOCK**

STATISTIC	GROUP	AGE COUNT	AGE MEDIAN	AGE 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

Garfinkel, Abowd, Martindale 2018.

TABLE 2: **POSSIBLE AGES FOR A MEDIAN OF 30 AND MEAN OF 44**

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

Reconstruction in the US Census. 2010.

Variable	Range
Block	6,207,027 inhabited blocks
Sex	2 (Female/Male)
Age	103 (0-99 single age year categories, 100-104, 105-109, 110+)
Race	63 allowable race combinations
Ethnicity	2 (Hispanic/Not)
Relationship	17 values

Survey

Publication	Released counts
PL94-171 Redistricting	2,771,998,263
Balance of Summary File 1	2,806,899,669
Total Statistics in PL94-171 and Balance of SF1:	5,578,897,932
Published Statistics/person	18
Recall: Collected variables/person:	6
Published Statistics/collected variable	**18 ÷ 6 ffi 3**

5.5 billion simultaneous equations

on $\boxed{\text{1.8 billion unknown integers}}$

Reconstruction Formulation

Dataset X

Statistics $f_1, \dots, f_k$

answers

$$a_1 \approx f_1(X)$$
$$a_2 \approx f_2(X)$$
$$\vdots$$
$$a_k \approx f_k(X)$$

$\approx$ "approx"

Reconstruction Problem: Given "constraints" $\{f_i(X) \approx a_i\}$, find a dataset $\tilde{X}$ that is consistent w/ the constants.

Linear Reconstrudion Attack

- Introduced by Dinur& Nissim in 2003   evelopment of Differential Privacy. 06

Data Set X

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

Z: identifiers S: Secret bit

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

← Format.

Release count statistics: # people satisfy some property

- How many people $\underbrace{\text{are older than 40}}_{\text{property } \varphi(z_j)}$ & have secret bit =1?

$$f(X) = \sum_{j=1}^{n} \varphi(z_j)\, s_j \quad \text{for some } \varphi : Z \mapsto \{0,1\}$$

$$\Updownarrow$$

$$f(X) = \underbrace{(\varphi(z_1), \varphi(z_2), \ldots, \varphi(z_n))}_{\text{"property" bit vector}} \cdot \underbrace{(s_1, \ldots, s_n)}_{\text{secret bit vector.}}$$

↑ dot product

inner product

Releasing k linear Statistics

$$\text{Released Statistics} \quad \begin{bmatrix} f_1(X) \\ \vdots \\ f_k(X) \end{bmatrix} = \underbrace{\begin{bmatrix} \varphi_1(z_1) & \cdots & \varphi_1(z_n) \\ \vdots & \ddots & \vdots \\ \varphi_k(z_1) & \cdots & \varphi_k(z_n) \end{bmatrix}}_{F} \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix} \leftarrow \text{Seret bits}$$

$$f_i(X) = F_i \cdot s$$

Examples :

$\varphi_1(z_j) = 1$: z_j is older than 40

$\varphi_2(z_j) = 1$: z_j is older than 40 and male

$\varphi_3(z_j) = 1$: z_j is older than 20 and male

First Reconstruction Attack

"You can't release all count statistics with non-trivial accuracy."

n = number of people

Queries : $k = 2^n$

For every $v \in \{0,1\}^n$, $F_v = v$

Reconstruction :

Suppose the answers $(a_v)_{v \in \{0,1\}^n}$, $\forall\, v \in \{0,1\}^n$, $|\underbrace{F_v \cdot s}_{\text{True answer}} - \underbrace{a_v}_{\text{Released answer.}}| \le \alpha n$

$\alpha = 5\%$

Choose $\tilde{s} \in \{0,1\}^n$, $\forall\, v$, $\boxed{|F_v \cdot \tilde{s} - a_v| \le \alpha \cdot n}$

Constraints.

Theorem. $\| s - \tilde{s} \|_1 \le 4\alpha n$

↳ Reconstruct 80% of the bits.
$= 1 - 20\%$

Theorem. If all 2^n counts are within αn error, then $s, \tilde{s}$ disagree on $\le 4\alpha n$ bits.

Proof Intuition.

$s = [1011 \text{ ———— }]$

$\tilde{s} = [\underbrace{0100}_{\text{property } \varphi_j \text{ that captures the diff.}} \text{ ———— }]$

$\varphi_j \downarrow$ Statistic f

Assumption

a_f: Released answer.

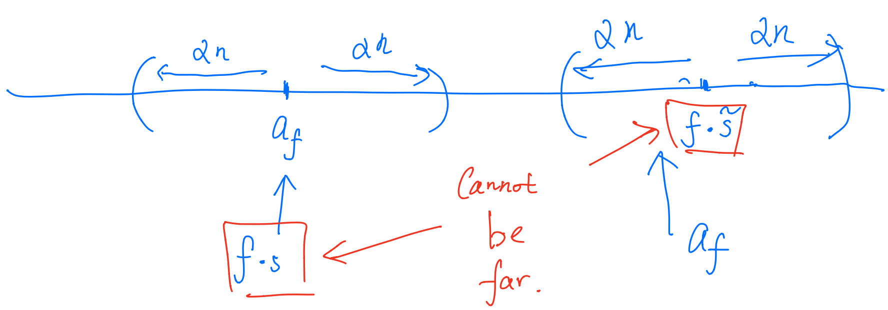
$f \cdot s$: true answer.

$|a_f - f \cdot s| \le \alpha n$

Reconstruction:

Find $\tilde{s}$ such that

$|a_f - f \cdot \tilde{s}| \le \alpha n$



Theorem. If all 2^n counts are within αn error, then $s, \tilde{s}$ disagree on $\underline{\leq 4\alpha n \text{ bits.}}$

Proof Sketch.

Two sets: $S_{01} = \{ j : s_j = 0 \ \& \ \tilde{s}_j = 1 \}$

$S_{10} = \{ j : s_j = 1 \ \& \ \tilde{s}_j = 0 \}$

Proof by Contradiction

If $\| s - \tilde{s} \|_1 > 4\alpha n$ $\quad$ ℓ_1 norm $\sum_j |s_j - \tilde{s}_j|$

$\Rightarrow |S_{01}| > 2\alpha n$ or $|S_{10}| > 2\alpha n$ $\quad = \sum_{j \in S_{01}} \underbrace{|s_j - \tilde{s}_j|}_{1} + \sum_{j \in S_{10}} \underbrace{|s_j - \tilde{s}_j|}_{1}$

$\Rightarrow$ Then there exists $v \in \{0,1\}^n$ such that $|v \cdot (s - \tilde{s})| > 2\alpha n$

$\Rightarrow |v \cdot \tilde{s} - a_v| > \underbrace{2\alpha n - |v \cdot s - a_v|}_{\text{Triangle Inequality}} > \alpha n$

Triangle Inequality $\rightarrow$ $|v \cdot \tilde{s} - a_v| > |v \cdot (s - \tilde{s})| - |v \cdot s - a_v| > 2\alpha n - |v \cdot s - a_v|$

$\Rightarrow$ Contradiction (Since $|v \cdot \tilde{s} - a_v| \leq \alpha n$ in our reconstruction)

No Class 9/6.

No Recitation this Friday

Reading for next Weds.

Reconstruction Using Fewer Queries

Released Statistics $\ll 2^n$?

Attack : Choose $k = 20n$ random $\varphi_i : Z \mapsto \{0,1\}$, $\forall i \in [k]$.

$\Rightarrow k$ random vectors/queries $F_i \in \{0,1\}^n$

Suppose that answers : $\forall i \in [k]$, $|F_i \cdot s - a_i| \leq \alpha n$

Find $\tilde{s} \in \{0,1\}^n$ such that: $\forall i \in [k]$, $|F_i \cdot \tilde{s} - a_i| \lesssim \alpha n$

Theorem. $\|s - \tilde{s}\|_1 \leq 256\, \alpha^2 n^2$.