CS208: Applied Privacy for Data Science
Reidentification & Reconstruction Attacks

James Honaker & Salil Vadhan

School of Engineering & Applied Sciences
Harvard University

February 4, 2019

CRCS Center for Research on Computation and Society
at Harvard John A. Paulson School of Engineering and Applied Sciences

Cohen & Nissim

Linear Program Reconstruction in Practice

- Use queries of sums over random subsets to reconstruct individual data.
- Importantly, the members of the subset are reported in each sum.
- Received the Aircloak Bounty ($5000) for reidentifying challenge data in the *Diffix* commercial system.

Regression Based Reconstruction

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_N x_{N,i} + \epsilon_i$$

Here:

N	is the Number of people in the database
i	is query index
y_i	is i-th query release
$x_{h,i}$	is a $\{0, 1\}$-indicator of whether person h was included in query i
β_h	is h's sensitive data
ϵ_i	is the noise added to the i-th query

Regression Based Reconstruction

$$
\begin{aligned}
y_i &= \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_N x_{N,i} + \epsilon_i \\
7 &= 1{\cdot}1 + 0{\cdot}1 + 1{\cdot}0 + 0{\cdot}0 + \ldots + 0{\cdot}1 + 2 \\
4 &= 1{\cdot}0 + 0{\cdot}1 + 1{\cdot}1 + 0{\cdot}1 + \ldots + 0{\cdot}1 + (-1) \\
6 &= 1{\cdot}0 + 0{\cdot}0 + 1{\cdot}0 + 0{\cdot}1 + \ldots + 0{\cdot}0 + 1
\end{aligned}
$$

Here:

- N is the Number of people in the database
- i is query index
- y_i is i-th query release
- $x_{h,i}$ is a $\{0,1\}$-indicator of whether person h was included in query i
- β_h is h's sensitive data
- ϵ_i is the noise added to the i-th query

Regression Based Reconstruction

$$
\begin{aligned}
y_i &= \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_N x_{N,i} + \epsilon_i \\
7 &= 1{\cdot}1 + 0{\cdot}1 + 1{\cdot}0 + 0{\cdot}0 + \ldots + 0{\cdot}1 + 2 \\
4 &= 1{\cdot}0 + 0{\cdot}1 + 1{\cdot}1 + 0{\cdot}1 + \ldots + 0{\cdot}1 + (-1) \\
6 &= 1{\cdot}0 + 0{\cdot}0 + 1{\cdot}0 + 0{\cdot}1 + \ldots + 0{\cdot}0 + 1
\end{aligned}
$$

Here:

- N is the Number of people in the database
- i is query index
- y_i is i-th query release
- $x_{h,i}$ is a $\{0,1\}$-indicator of whether person h was included in query i
- β_h is h's sensitive data
- ϵ_i is the noise added to the i-th query

Regression Based Reconstruction

Find $\hat{\beta}_1, \ldots, \hat{\beta}_n$ s.t.:

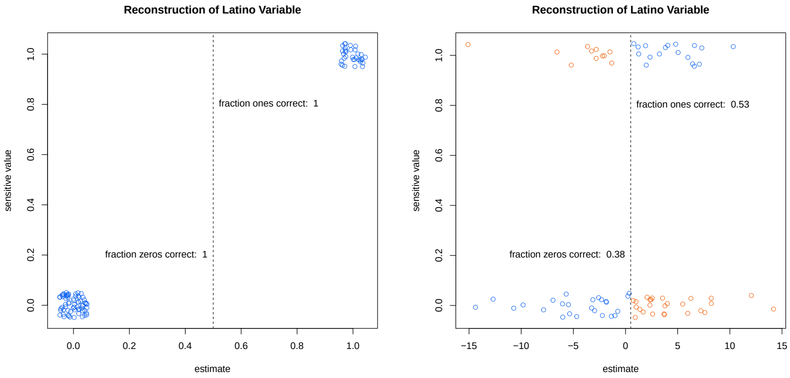
$$\hat{\beta} = \texttt{argmin}[\sum_i (y_i - \hat{y}_i)^2]$$

$$\text{where} \qquad \hat{y}_i = \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \ldots + \hat{\beta}_N x_{N,i}$$

In R see:
`lm()`
In Python see for example:
`linear_model.LinearRegression()`
from scikit-learn.

Example

From `regressionAttack.r`:



Garfinkel *et al.*

Understanding Database Reconstruction Attacks on Public Data

- Demonstrates feasibility of Database Reconstruction Attacks (DRAs) on small Census blocks by using their released aggregate statistics.
- 1.5M census blocks with between 1 and 7 residents
- Each released statistic provides a constraint, and some blocks have only one possible dataset that satisfy all constraints.

Example Dataset

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

		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

SAT Solvers

SAT Solvers state the feasibility of a solution to a series of logical formulae, and find a solution if one exists.

PicoSAT has R and Python bindings.
In R:

```
install.packages("rpicosat")
```

In Python:

```
pip install pycosat
```

Conjunctive Normal Form

Conjunctive Normal Form (CNF) is expressed as *conjunctions of disjunctions*, that is, clauses entirely composed of OR $\vee$, each of which are bound together by AND $\wedge$.
Negations of literals are allowed.

Examples:

- $(A \vee B) \wedge (\neg C \vee D)$
- $(A \vee B) \wedge (C \vee D \vee E \vee F)$
- $(A \vee B) \wedge (C)$
- (A)

Construction:

- $A \rightarrow B$ is expressed as $(\neg A \vee B)$
- $A \leftrightarrow B$ is expressed as $(\neg A \vee B) \wedge (A \vee \neg B)$

Dataset

	Actual	
	Sex	**Married**
1	1	0
2	1	1
3	0	1
4	0	1
5	0	0

Dataset

	Actual			Labels	
	Sex	**Married**		**Sex**	**Married**
1	1	0		A	F
2	1	1		B	G
3	0	1		C	H
4	0	1		D	I
5	0	0		E	J

Dataset

	Actual		Labels	
	Sex	**Married**	**Sex**	**Married**
1	1	0	A	F
2	1	1	B	G
3	0	1	C	H
4	0	1	D	I
5	0	0	E	J

Release:

$$\sum Sex = 2,$$
$$\sum Married = 3,$$
$$\sum Sex \cdot Married = 1.$$

Challenge

Transform into CNF:

$(A \wedge F) \vee (B \wedge G) \vee (C \wedge H) \vee (D \wedge I) \vee (E \wedge J)$

What about:

$(A \wedge F) \oplus (B \wedge G) \oplus (C \wedge H) \oplus (D \wedge I) \oplus (E \wedge J)$

Challenge

Transform into CNF:
$(A \wedge F) \vee (B \wedge G) \vee (C \wedge H) \vee (D \wedge I) \vee (E \wedge J)$

What about:
$(A \wedge F) \oplus (B \wedge G) \oplus (C \wedge H) \oplus (D \wedge I) \oplus (E \wedge J)$

(requires 2^4 clauses)