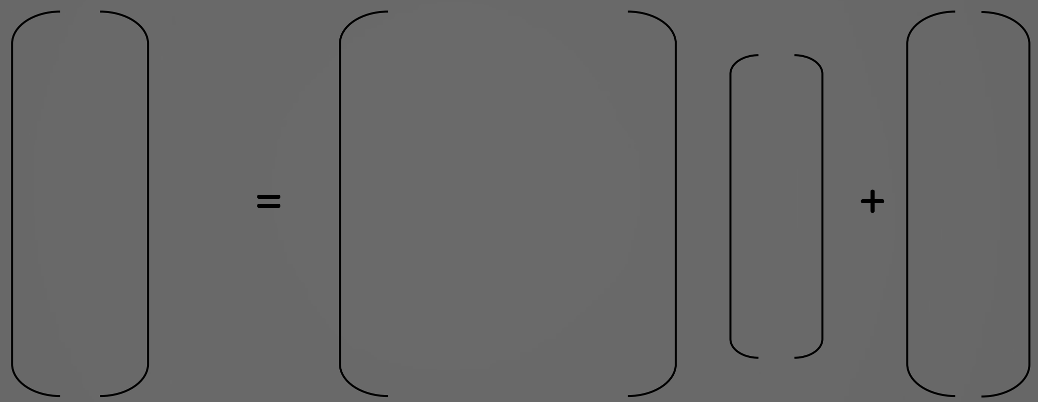
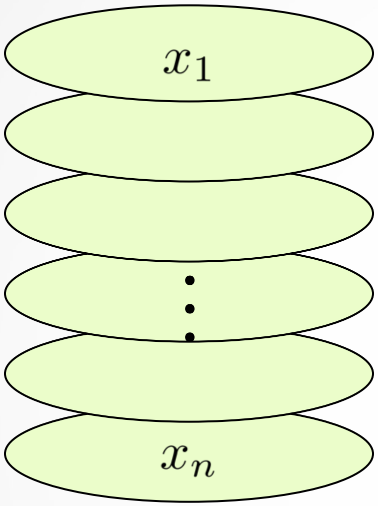
The Power of Linear Reconstruction Attacks

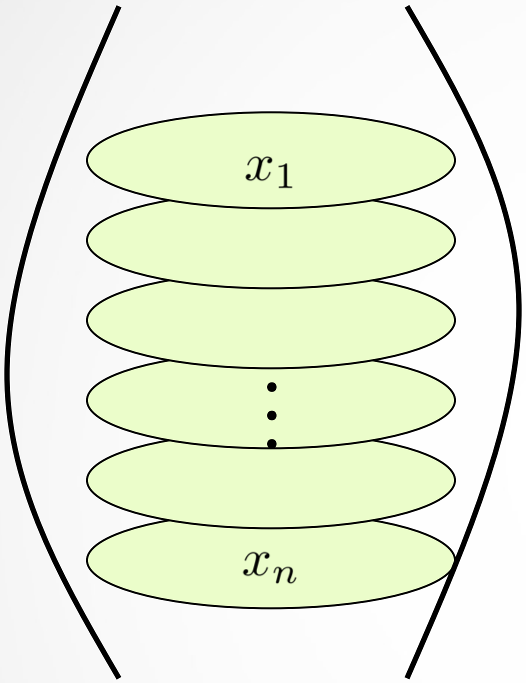


Shiva Kasiviswanathan (General Electric Research)

Joint work with
Mark Rudelson (University of Michigan)
Adam Smith (Penn State University)



Database D

f  privately

Database D

f could be the

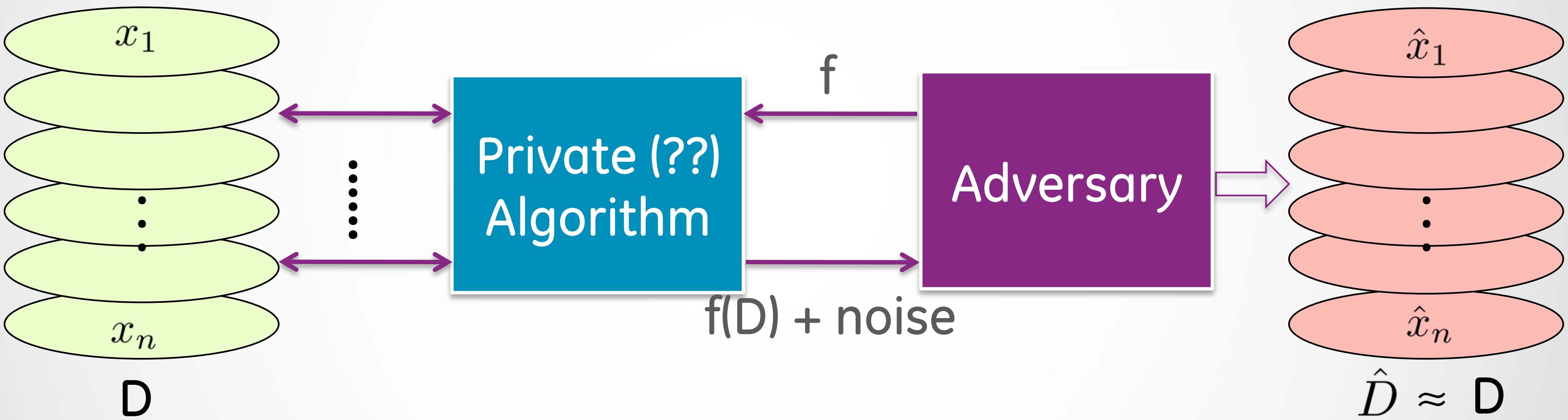
1. average function
2. correlation function
3. classifier

……

Informally: How much distortion is needed in f(D), to guarantee the privacy of D’s entries?

What is a Reconstruction Attack?

Reconstruction Attacks [DN'03,DMT'07,DY'08,KRSU'10,D'12,KRS'13]



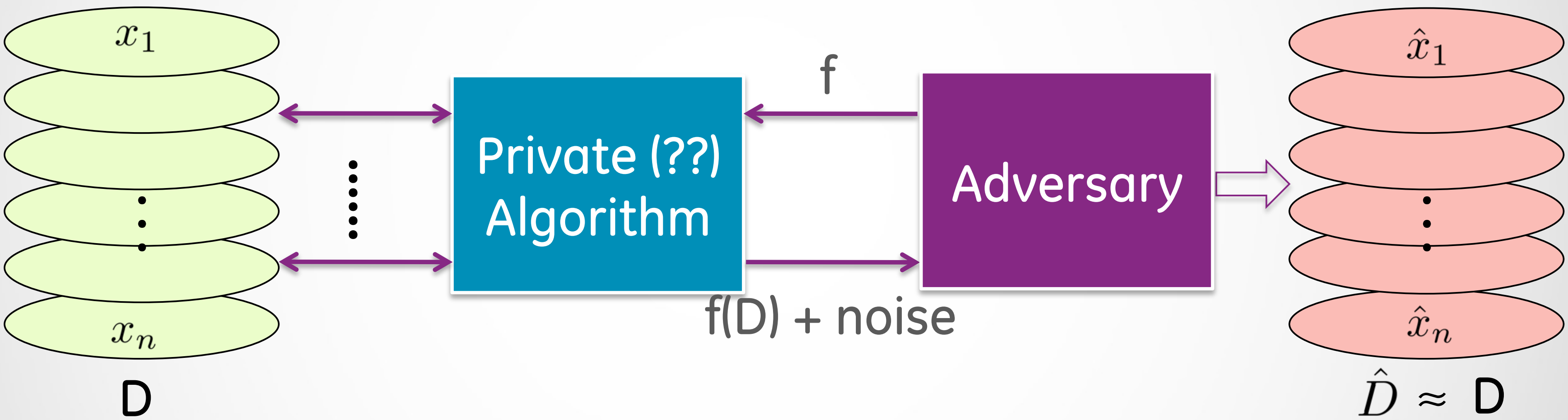
Reconstruction attack implies a lower bound on distortion for any reasonable notion of privacy

Talk Summary

- ❑ Linear reconstruction attacks work surprisingly in many settings
  - Marginal tables
  - Decision tree classification rate
  - Linear and Logistic regression parameters
  - M-estimators
  - .....
- ❑ Analysis of the attacks under distributional assumptions on data

Privacy Requires Distortion

Reconstruction Attacks [DN'03,DMT'07,DY'08,KRSU'10,D'12,KRS'13]

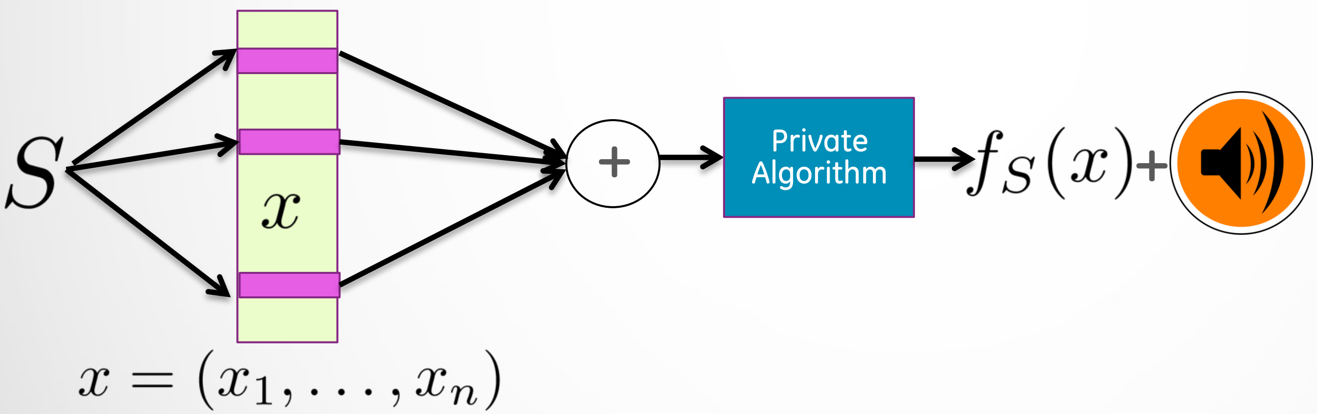


[DN'03]: Answering "too many" subset sum queries "too accurately" allows an adversary to reconstruct database almost entirely

Reconstruction Attacks [DN'03]

Concrete Setting: n users, each with secret $x_i \in \{0, 1\}$

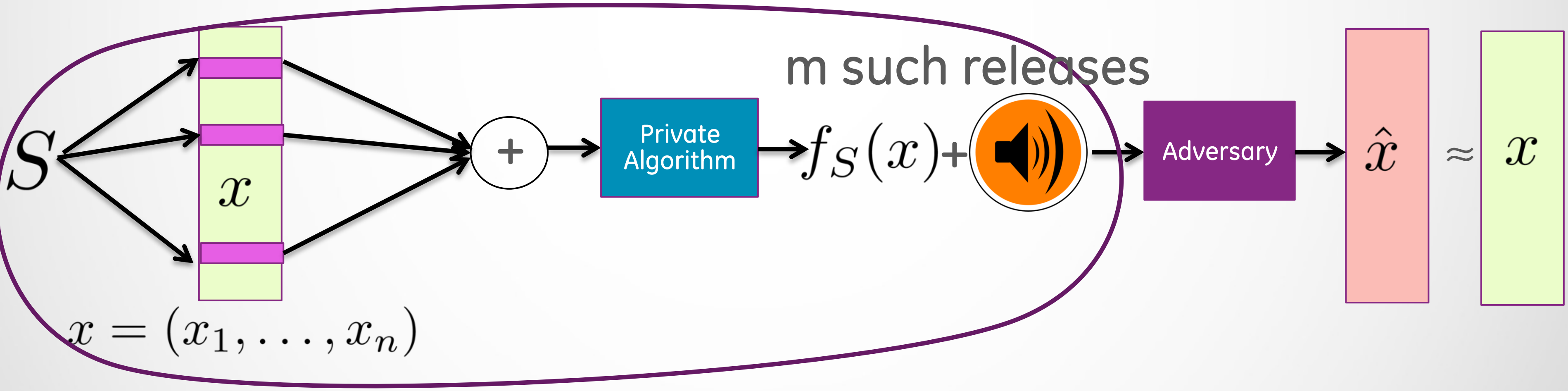
Inner-product Query: for $S \in \{-1, 1\}^n$, let $f_S(x) = \langle S, x \rangle$



Reconstruction Attacks [DN'03]

Concrete Setting: n users, each with secret $x_i \in \{0,1\}$

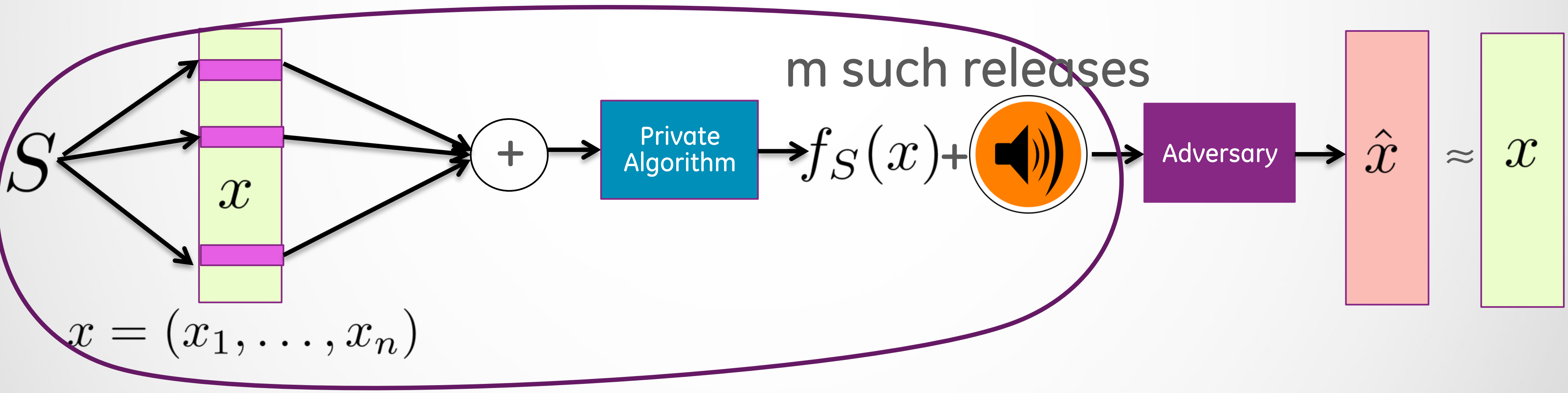
Inner-product Query: for $S \in \{-1,1\}^n$, let $f_S(x) = \langle S, x \rangle$



Reconstruction Attacks [DN'03]

Concrete Setting: n users, each with secret $x_i \in \{0,1\}$

Inner-product Query: for $S \in \{-1,1\}^n$, let $f_S(x) = \langle S, x \rangle$

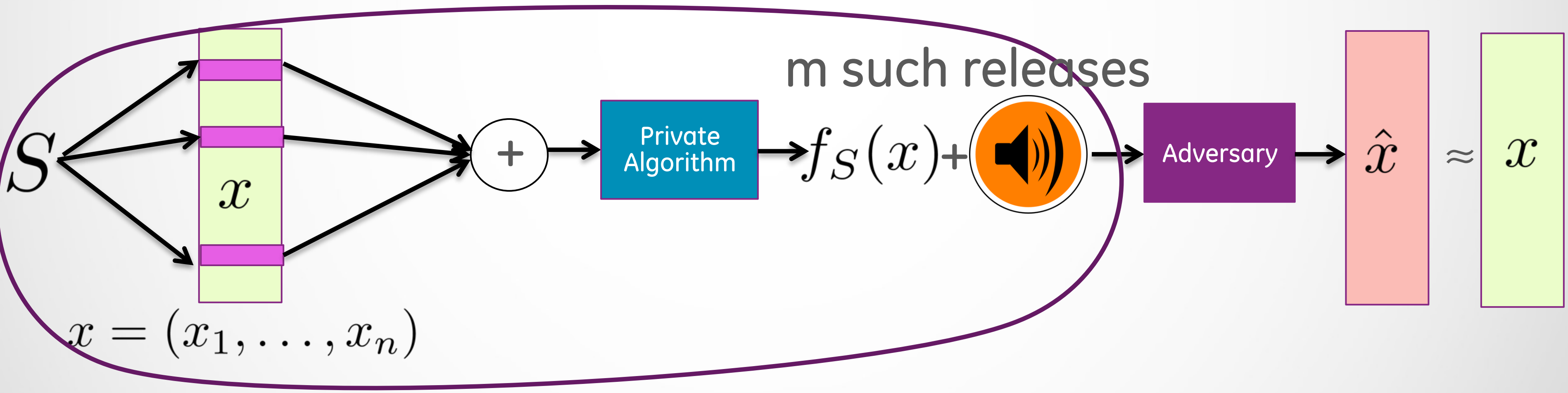


Theorem [DN'03] (Informal): If $m \approx n$ releases each with $o(\sqrt{n})$ noise then there exists an adversary with $d_{\text{Hamming}}(\hat{x}, x) = o(n)$.

Reconstruction Attacks [DN'03]

Concrete Setting: n users, each with secret $x_i \in \{0,1\}$

Inner-product Query: for $S \in \{-1,1\}^n$, let $f_S(x) = \langle S, x \rangle$



- Which queries $S_1, \ldots, S_m$ allow reconstruction?
- Number of queries?
- Running time?

Our Results:

Using linear reconstruction attacks to obtain privacy lower bounds for natural, symmetric queries

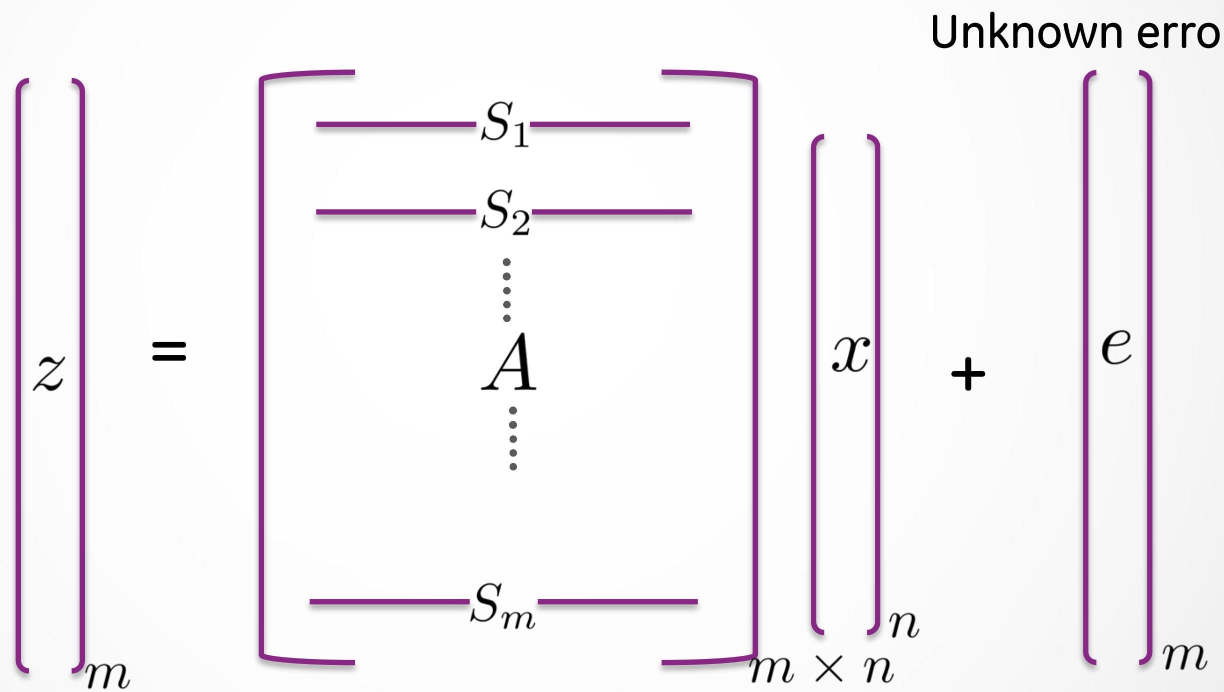
- [KRSU'10] marginal (contingency) tables
- Each person's data is a row in a table
- k-way marginal: distribution of some k attributes

- [KRS'12] regression analysis, decision tree classifiers, boolean functions

Linear Reconstruction Problem [DMT'07,DY'08]

Let A be a real-valued matrix and e be an unknown error vector

Problem: Given $z \approx Ax (z = Ax + e)$ construct $\hat{x} \approx x$.



Natural approach: $\hat{x} = \mathrm{argmin}_x \|z - Ax\|_p$

- p=2: gives least squares method
- p=1: gives LP decoding method

Least Squares Attack (L_2-attack) [DY'08]

Solving $\min_x \|z - Ax\|_2$

Let $A = U \times \Sigma \times V^\top$ be the singular value decomposition of A

Define $A_{\text{inv}} = V \times \Sigma^{-1} \times U^\top$(pseudo-inverse of A)

Attack: Define $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_n)$ where

$$\hat{x}_i = \begin{cases} 1 \text{ if the } i^{\text{th}} \text{ element of } A_{\text{inv}} z \geq \frac{1}{2} \\ 0 \text{ otherwise} \end{cases}$$

Least Squares Attack (L_2-attack) [DY'08]

Solving $\min_x \|z - Ax\|_2$

Let $A = U \times \Sigma \times V^\top$ be the singular value decomposition of A

Define $A_{\text{inv}} = V \times \Sigma^{-1} \times U^\top$ (pseudo-inverse of A)

Attack: Define $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_n)$ where

$$\hat{x}_i = \begin{cases} 1 \text{ if the } i^{\text{th}} \text{ element of } A_{\text{inv}} z \geq \frac{1}{2} \\ 0 \text{ otherwise} \end{cases}$$

$\Omega(\sqrt{m})$

Proof idea:

If the least singular value of A is "sufficiently big", then $\hat{x}$ is close to x

Both L_1- and L_2-attacks well understood

	Error vector e	**Fraction of Recovered** x	**Condition on** A	**Pluses**	**Minuses**
Least Squares Method	All entries $\leq \sqrt{n}$	1 - o(1)	Least singular value $\geq \sqrt{m}$		
LP Decoding Method	At least $1-\gamma$ frac. entries $\leq \sqrt{n}$	1 - o(1)	Least singular value $\geq \sqrt{m}$ and Euclidean section property	can tolerate bigger error vector	stronger condition on A, and costlier running time

Input Setting

D =

	Over 60	Smoking	Exercise	High Blood Pressure
Alice	0	1	0	1
Bob	1	1	0	1
Charlie	1	0	1	0
Dave	1	0	0	1

Non-sensitive

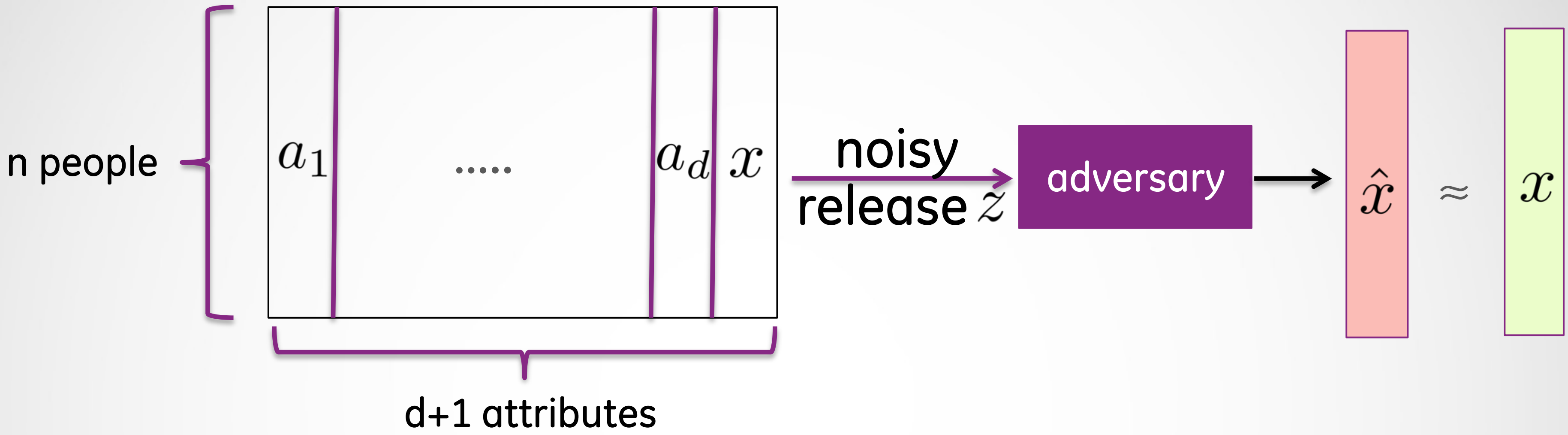
available to the adversary

Sensitive

secret

Database D: Table of values for n individuals on d+1 attributes

Reconstruction from Marginals [KRSU'10]

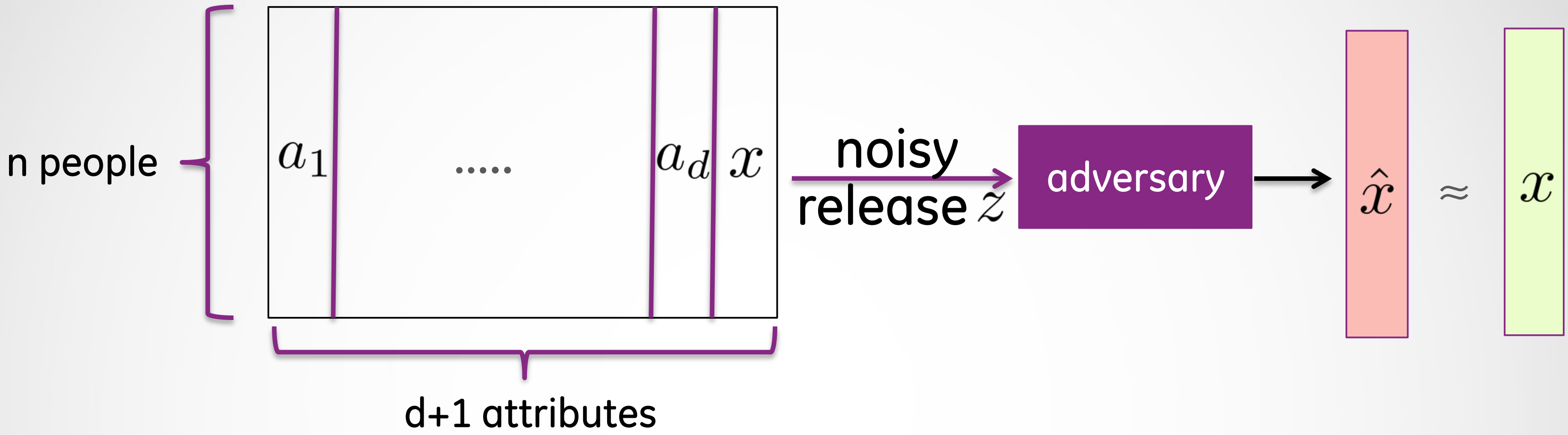


Releasing 2-way marginals

2-way marginals include $\langle a_1, x\rangle, \langle a_2, x\rangle, \ldots, \langle a_d, x\rangle$

Solve $\operatorname{argmin}_x \left\| \begin{pmatrix} \\ z \\ \\ \end{pmatrix} - \begin{pmatrix} \text{---} a_1^\top \text{---} \\ \text{---} a_2^\top \text{---} \\ A \\ \text{---} a_d^\top \text{---} \end{pmatrix} \begin{pmatrix} \\ x \\ \\ \end{pmatrix} \right\|_p \longrightarrow \hat{x}$

Reconstruction from Marginals [KRSU'10]



Releasing 3-way marginals

3-way marginals include $\langle a_1 \odot a_2, x \rangle, \langle a_1 \odot a_3, x \rangle, \ldots, \langle a_{d-1} \odot a_d, x \rangle$

$\odot = \text{Hadarmard product (entry-wise product)}$

$$\text{Solve } \operatorname{argmin}_x \left\| \begin{pmatrix} \\ z \\ \\ \end{pmatrix} - \begin{pmatrix} (a_1 \odot a_2)^\top \\ (a_1 \odot a_3)^\top \\ A \\ (a_{d-1} \odot a_d)^\top \end{pmatrix} \begin{pmatrix} \\ x \\ \\ \end{pmatrix} \right\|_p \longrightarrow \hat{x}$$

Analysis

Idea: Assume non-sensitive information are i.i.d.

Spectrum of Correlated Random Matrices

Key lemma for 3-way marginals:

Let each of the a_i be an i.i.d. (0-1) random vector with $d \geq \sqrt{n}$.

$$A = \underbrace{\begin{pmatrix} (a_1 \odot a_2)^\top \\ (a_1 \odot a_3)^\top \\ \vdots \\ (a_{d-1} \odot a_d)^\top \end{pmatrix}}_{\binom{d}{2} \times n}$$

Then w.h.p. the least singular value of matrix A is $\Omega(d)$.

Analysis

Idea: Assume non-sensitive information are i.i.d.

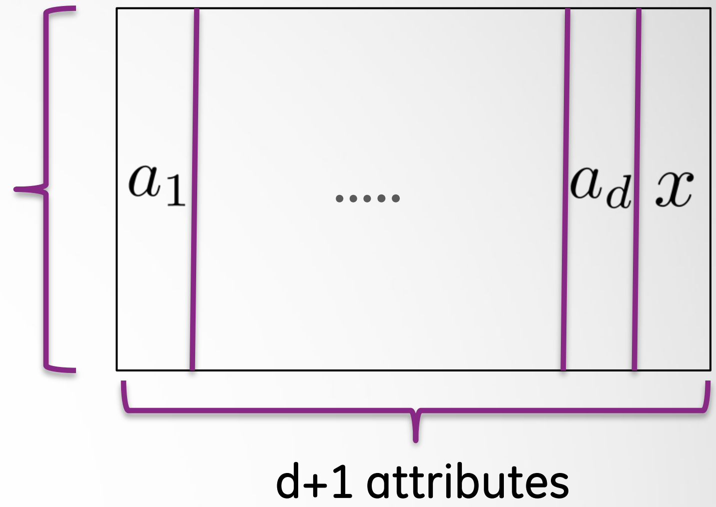
Spectrum of Correlated Random Matrices

Key lemma for k+1-way marginals:

Let each of the a_i be an i.i.d. (0-1) random vector with $d \geq n^{\frac{1}{k}}$.

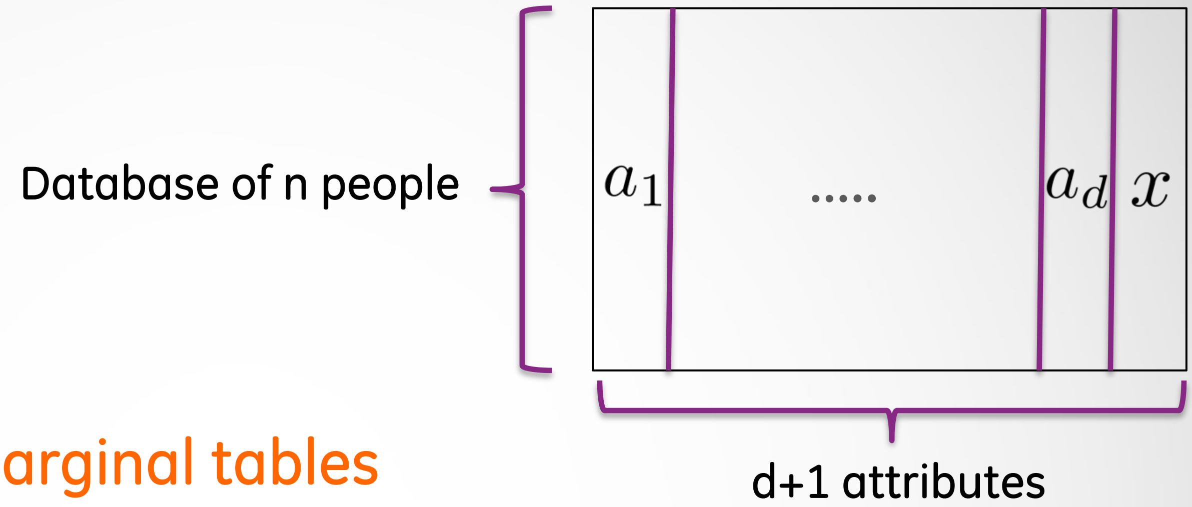
$$A = \underbrace{\begin{pmatrix} (a_1 \odot a_2 \cdot\cdot \odot a_k)^\top \\ (a_1 \odot a_3 \cdot\cdot \odot a_{k+1})^\top \\ \vdots \\ (a_{d-k} \odot a_{d-k+1} \cdot\cdot \odot a_d)^\top \end{pmatrix}}_{\binom{d}{k} \times n}$$

Then w.h.p. the least singular value of matrix A is $\Omega(d^{\frac{k}{2}})$.



Releasing k+1-way marginal tables

Theorem [KRSU'10]: If an algorithm always releases (k+1)-way marginals with $\min\{o(d^{\frac{k}{2}}), o(\sqrt{n})\}$ noise per entry then there exists an adversary that w.h.p. can construct $\hat{x}$ with $d_{\text{Hamming}}(\hat{x}, x) = o(n)$.

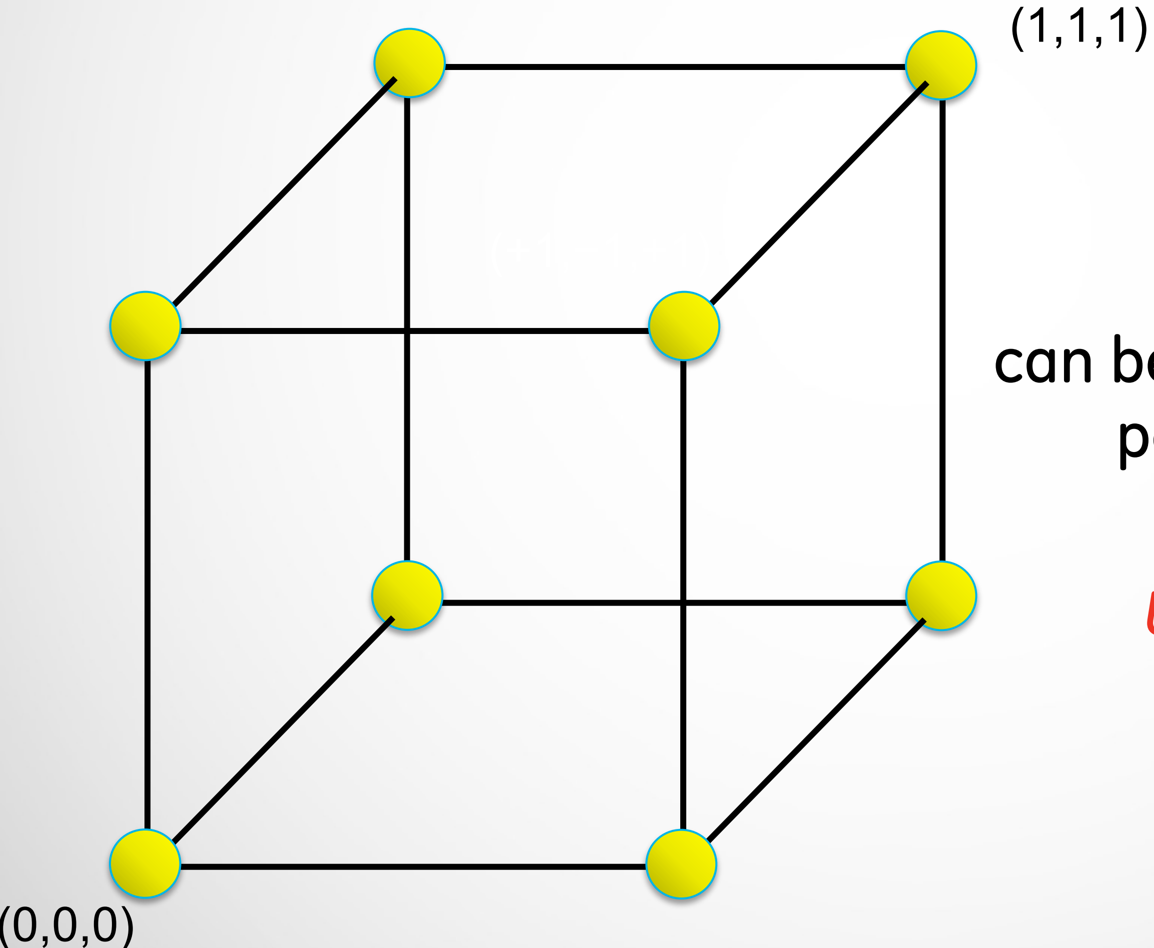


Releasing k+1-way marginal tables

Theorem [KRSU'10]: If an algorithm always releases (k+1)-way marginals with $\min\{o(d^{\frac{k}{2}}), o(\sqrt{n})\}$ noise per entry then there exists an adversary that w.h.p. can construct $\hat{x}$ with $d_{\text{Hamming}}(\hat{x}, x) = o(n)$.

Theorem [De'12]: Stronger result with L_1 attack

Extension to Boolean Functions



Fact: Every function

$$f : \{0,1\}^k \to \{0,1\}$$

can be expressed as a multilinear polynomial of degree ≤ k

Use Fourier Decomposition

Non-Degenerate Function: A boolean function on k variables is non-degenerate if it can be represented as a multilinear polynomial of degree exactly k

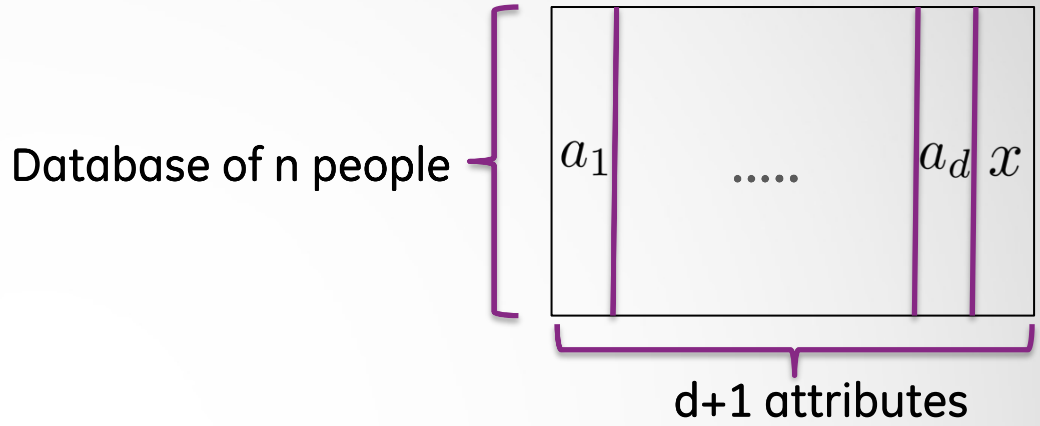
Examples include:

AND, OR, XOR, MAJ, depth k decision trees

Examples:

- AND function: $x_1 \times \ldots \times x_k$
- OR function: $1 - (1 - x_1) \times \ldots \times (1 - x_k)$

Evaluating Boolean Functions (k = 3)

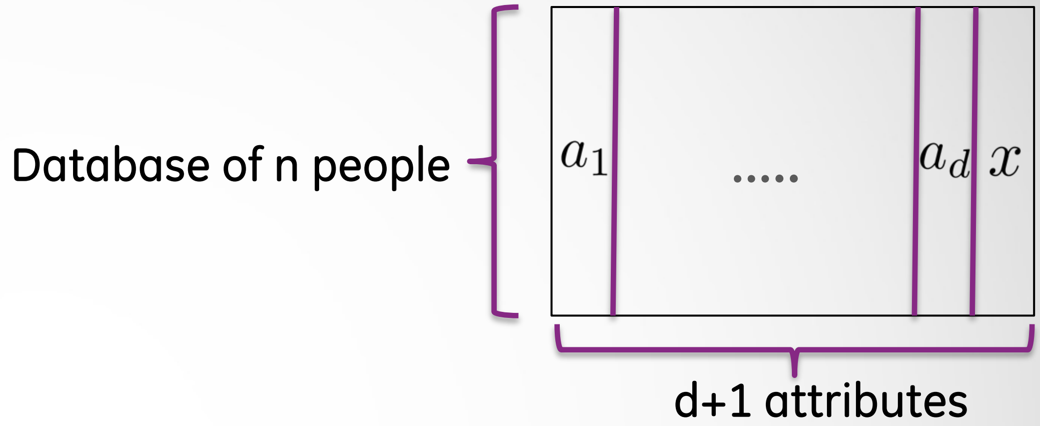


Remember 3-way marginal between columns a_1,a_2, and x is

$$\langle a_1 \odot a_2, x \rangle = \sum_{i=1}^{n} a_{1_i} \times a_{2_i} \times x_i$$

Adversary gets distorted $\langle a_1 \odot a_2, x \rangle, \langle a_1 \odot a_3, x \rangle, \ldots, \langle a_{d-1} \odot a_d, x \rangle$

Evaluating Boolean Functions (k = 3)



Remember 3-way marginal between columns a_1,a_2, and x is

$$\langle a_1 \odot a_2, x \rangle = \sum_{i=1}^{n} a_{1_i} \times a_{2_i} \times x_i$$

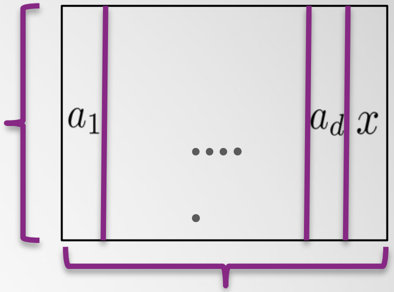
Adversary gets distorted $\langle a_1 \odot a_2, x \rangle, \langle a_1 \odot a_3, x \rangle, \ldots, \langle a_{d-1} \odot a_d, x \rangle$

For a general function $f : \{0,1\}^3 \to \{0,1\}$, let

$$F(a_1, a_2, x) = \sum_{i=1}^{n} f(a_{1_i}, a_{2_i}, x_i)$$

Adversary gets distorted $F(a_1, a_2, x), F(a_1, a_3, x), \ldots, F(a_{d-1}, a_d, x)$

Database of n people

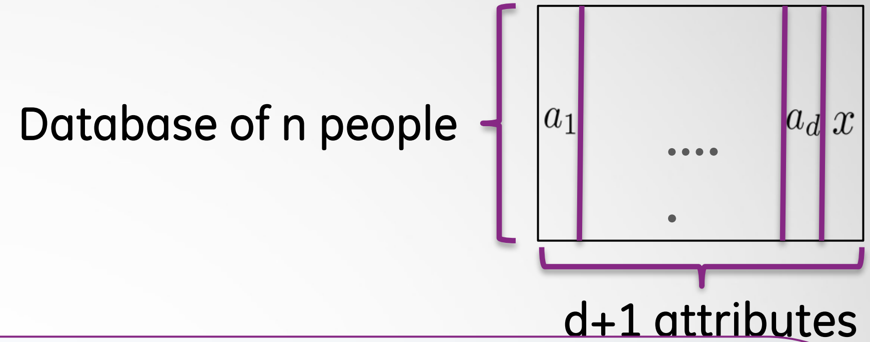


d+1 attributes

Theorem [KRS10]: Let $f : \{0,1\}^3 \to \{0,1\}$ be a non-degenerate function. Consider an algorithm releasing F evaluated on every pair of columns from $\{a_1, \ldots, a_d\}$ with x.

L_2-attack: If for every database D, the algorithm adds $\min\{o(\sqrt{d}, o(\sqrt{n})\}$ noise to each release

There exists an adversary that can w.h.p. construct $\hat{x}$ with $d_{\text{Hamming}}(\hat{x}, x) = o(n)$.



Theorem [KRS'10]: Let $f : \{0,1\}^3 \rightarrow \{0,1\}$ be a non-degenerate function. Consider an algorithm releasing F evaluated on every pair of columns from $\{a_1, \ldots, a_d\}$ with x.

L_2-attack: If for every database D, the algorithm adds $\min\{o(\sqrt{d}, o(\sqrt{n})\}$ noise to each release

There exists an adversary that can w.h.p. construct $\hat{x}$ with $d_{\text{Hamming}}(\hat{x}, x) = o(n)$.

Also generalizes to boolean function with more variables

Lower Bounds for Privately Releasing M-estimators

M-estimators (Emp. Risk. Min.)

Let $x_1, \ldots, x_n \in \mathcal{R}^k$ be n data points

Loss func: Let $\ell(\theta; x_i)$ measure the "fit" of the parameter $\theta \in \mathcal{R}^k$ to x_i

The M-estimator $\hat{\theta}$ is

$$= \text{argmin}_\theta \sum_{i=1}^{n} \ell(\theta; x_i)$$

e.g.,

$$\hat{\theta} = \text{argmin}_\theta \sum \|x_i - \theta\|_1$$

$$\hat{\theta} = \text{argmin}_\theta \sum \|x_i - \theta\|_2^2$$

M-estimators (Emp. Risk. Min.)

Let $x_1, \ldots, x_n \in \mathcal{R}^k$ be n data points

Loss func: Let $\ell(\theta; x_i)$ measure the "fit" of the parameter $\theta \in \mathcal{R}^k$ to x_i

The M-estimator $\hat{\theta}$ is

$$= \operatorname{argmin}_\theta \sum_{i=1}^{n} \ell(\theta; x_i)$$

e.g.,

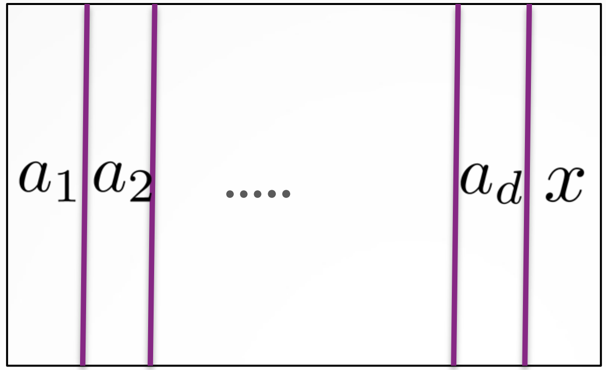
$$\hat{\theta} = \operatorname{argmin}_\theta \sum \|x_i - \theta\|_1$$

$$\hat{\theta} = \operatorname{argmin}_\theta \sum \|x_i - \theta\|_2^2$$

If loss function ℓ is differentiable, then $\hat{\theta}$ can be obtained by

$$\frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ell(\theta; x_i) = 0$$

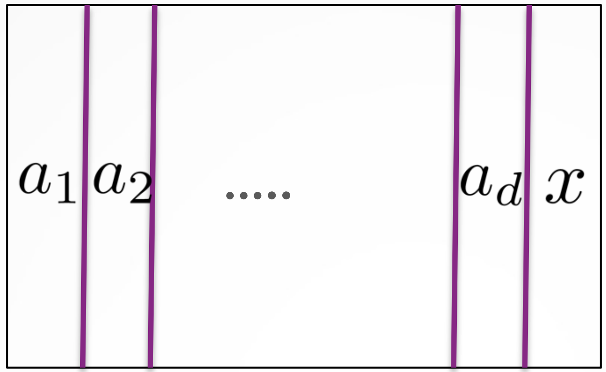
Look at Logistic Regression (k = 1)



The logistic regression parameter θ_1 between column a_1 and x is

$$\begin{pmatrix} \log\left(\frac{\zeta_1}{1-\zeta_1}\right) \\ \vdots \\ \log\left(\frac{\zeta_n}{1-\zeta_n}\right) \end{pmatrix} = a_1\theta_1 \text{ where } \zeta_i = \Pr[x_i = 1]$$

Look at Logistic Regression (k = 1)



To estimate θ_1:
1) Take MLE
2) Set grad. of MLE = 0

MLE estimate $\hat{\theta}_1$ of θ_1 is:

$$\begin{pmatrix} \text{---} \; a_1^\top \; \text{---} \end{pmatrix} \begin{bmatrix} \\ x \\ \\ \end{bmatrix} + \begin{pmatrix} \text{---} \; a_1^\top \; \text{---} \end{pmatrix} \begin{pmatrix} \frac{1+\exp(\hat{\theta}_1 a_{1_1})}{\exp(\hat{\theta}_1 a_{1_1})} \\ \vdots \\ \frac{1+\exp(\hat{\theta}_1 a_{1_n})}{\exp(\hat{\theta}_1 a_{1_n})} \end{pmatrix} = 0$$

Logistic Regression: Linear Reconstruction

from $\hat{\theta}_1$: $$\begin{pmatrix} a_1^\top \\ \\ \\ \end{pmatrix} \begin{bmatrix} \\ x \\ \\ \end{bmatrix} + \begin{bmatrix} a_1^\top & & \end{bmatrix} \begin{pmatrix} \frac{1+\exp(\hat{\theta}_1 a_{1_1})}{\exp(\hat{\theta}_1 a_{1_1})} \\ \vdots \\ \frac{1+\exp(\hat{\theta}_1 a_{1_n})}{\exp(\hat{\theta}_1 a_{1_n})} \end{pmatrix} = 0$$

Logistic Regression: Linear Reconstruction

from $\hat{\theta}_1$:
from $\hat{\theta}_2$:

$$\begin{pmatrix} a_1^\top \\ a_2^\top \\ \\ \\ \end{pmatrix} \begin{bmatrix} \\ x \\ \\ \end{bmatrix} + \begin{bmatrix} a_1^\top a_2^\top & & \end{bmatrix} \begin{bmatrix} \frac{1+\exp(\hat{\theta}_1 a_{1_1})}{\exp(\hat{\theta}_1 a_{1_1})} \\ \vdots \\ \frac{1+\exp(\hat{\theta}_1 a_{1_n})}{\exp(\hat{\theta}_1 a_{1_n})} \\ \frac{1+\exp(\hat{\theta}_2 a_{2_1})}{\exp(\hat{\theta}_2 a_{2_1})} \\ \vdots \\ \frac{1+\exp(\hat{\theta}_2 a_{2_n})}{\exp(\hat{\theta}_2 a_{2_n})} \end{bmatrix} = 0$$

Logistic Regression: Linear Reconstruction

$$
\begin{array}{r}
\text{from } \hat{\theta}_1: \\
\text{from } \hat{\theta}_2: \\
\text{from } \hat{\theta}_3: \\
\\
\end{array}
\begin{pmatrix} a_1^\top \\ a_2^\top \\ a_3^\top \\ \vdots \end{pmatrix}
\begin{bmatrix} x \end{bmatrix}
+
\begin{bmatrix} a_1^\top \, a_2^\top \, a_3^\top \, \ldots \end{bmatrix}
\begin{pmatrix}
\begin{Bmatrix}
\frac{1+\exp(\hat{\theta}_1 a_{1_1})}{\exp(\hat{\theta}_1 a_{1_1})} \\
\vdots \\
\frac{1+\exp(\hat{\theta}_1 a_{1_n})}{\exp(\hat{\theta}_1 a_{1_n})}
\end{Bmatrix} \\
\begin{Bmatrix}
\frac{1+\exp(\hat{\theta}_2 a_{2_1})}{\exp(\hat{\theta}_2 a_{2_1})} \\
\vdots \\
\frac{1+\exp(\hat{\theta}_2 a_{2_n})}{\exp(\hat{\theta}_2 a_{2_n})}
\end{Bmatrix} \\
\begin{Bmatrix}
\frac{1+\exp(\hat{\theta}_3 a_{3_1})}{\exp(\hat{\theta}_3 a_{3_1})} \\
\vdots \\
\frac{1+\exp(\hat{\theta}_3 a_{3_n})}{\exp(\hat{\theta}_3 a_{3_n})}
\end{Bmatrix} \\
\vdots
\end{pmatrix}
= 0
$$

Logistic Regression: Linear Reconstruction

$$
\begin{array}{l}
\text{from } \hat{\theta}_1: \\
\text{from } \hat{\theta}_2: \\
\text{from } \hat{\theta}_3:
\end{array}
\quad
\underbrace{\begin{pmatrix} a_1^\top \\ a_2^\top \\ a_3^\top \\ \vdots \end{pmatrix}}_{A}
\begin{bmatrix} x \end{bmatrix}
+
\underbrace{\begin{bmatrix} a_1^\top & a_2^\top & a_3^\top & \cdots \end{bmatrix}
\begin{bmatrix}
\begin{Bmatrix} \frac{1+\exp(\hat{\theta}_1 a_{1_1})}{\exp(\hat{\theta}_1 a_{1_1})} \\ \vdots \\ \frac{1+\exp(\hat{\theta}_1 a_{1_n})}{\exp(\hat{\theta}_1 a_{1_n})} \end{Bmatrix} \\
\begin{Bmatrix} \frac{1+\exp(\hat{\theta}_2 a_{2_1})}{\exp(\hat{\theta}_2 a_{2_1})} \\ \vdots \\ \frac{1+\exp(\hat{\theta}_2 a_{2_n})}{\exp(\hat{\theta}_2 a_{2_n})} \end{Bmatrix} \\
\begin{Bmatrix} \frac{1+\exp(\hat{\theta}_3 a_{3_1})}{\exp(\hat{\theta}_3 a_{3_1})} \\ \vdots \\ \frac{1+\exp(\hat{\theta}_3 a_{3_n})}{\exp(\hat{\theta}_3 a_{3_n})} \end{Bmatrix} \\
\begin{Bmatrix} \vdots \end{Bmatrix}
\end{bmatrix}}_{b}
= 0
$$

Logistic Regression: Linear Reconstruction

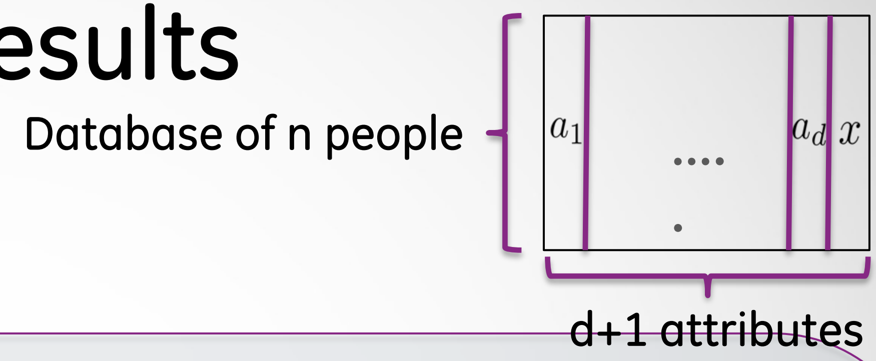
$$\begin{array}{l} \text{from } \hat{\theta}_1: \\ \text{from } \hat{\theta}_2: \\ \text{from } \hat{\theta}_3: \\ \\ \end{array} \underbrace{\begin{pmatrix} a_1^\top \\ a_2^\top \\ a_3^\top \\ \vdots \end{pmatrix}}_{A} \begin{bmatrix} x \end{bmatrix} + \underbrace{\begin{bmatrix} a_1^\top \, a_2^\top \, a_3^\top \, \ldots \end{bmatrix} \begin{bmatrix} \begin{Bmatrix} \frac{1+\exp(\hat{\theta}_1 a_{1_1})}{\exp(\hat{\theta}_1 a_{1_1})} \\ \vdots \\ \frac{1+\exp(\hat{\theta}_1 a_{1_n})}{\exp(\hat{\theta}_1 a_{1_n})} \end{Bmatrix} \\ \begin{Bmatrix} \frac{1+\exp(\hat{\theta}_2 a_{2_1})}{\exp(\hat{\theta}_2 a_{2_1})} \\ \vdots \\ \frac{1+\exp(\hat{\theta}_2 a_{2_n})}{\exp(\hat{\theta}_2 a_{2_n})} \end{Bmatrix} \\ \begin{Bmatrix} \frac{1+\exp(\hat{\theta}_3 a_{3_1})}{\exp(\hat{\theta}_3 a_{3_1})} \\ \vdots \\ \frac{1+\exp(\hat{\theta}_3 a_{3_n})}{\exp(\hat{\theta}_3 a_{3_n})} \end{Bmatrix} \\ \begin{Bmatrix} \vdots \end{Bmatrix} \end{bmatrix}}_{b} = 0$$

Linear system of the form: $Ax + b = 0$

Slight issue is that adversary gets noisy M-estimators $\hat{\theta}_1, \ldots, \hat{\theta}_d$ and not the noisy version of vector $Ax + b = 0$

But, this can be come by using Lipchitz-ness of the function

Logistic Regression Results



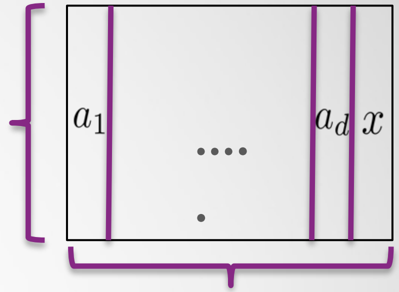
Theorem [KRS10]: Consider an algorithm that releases that releases the parameters of the logistic regression model each between column in $\{a_1, \ldots, a_d\}$ with x. Let $d \geq 2n$.

L$_2$-attack: If for every database D, the algorithm adds $o(1/\sqrt{n})$ noise to each parameter

There exists an adversary  that can w.h.p. construct $\hat{x}$ with $d_{\text{Hamming}}(\hat{x}, x) = o(n)$.

Logistic Regression Results

Database of n people



d+1 attributes

Theorem [KRS10]: Consider an algorithm that releases that releases the parameters of the logistic regression model each between column in $\{a_1, \ldots, a_d\}$ with x. Let $d \geq 2n$.

L_2-attack: If for every database D, the algorithm adds $o(1/\sqrt{n})$ noise to each parameter

There exists an adversary  that can w.h.p. construct $\hat{x}$ with $d_{\text{Hamming}}(\hat{x}, x) = o(n)$.

Attack & analysis works for any differentiable M-estimators

Wrapping Up

We use linear reconstruction attack to obtain privacy lower bounds for two natural and broad classes of functions

Boolean functions: Marginals, Decision tree error rates

Differentiable M-estimators: Linear and Logistic regression parameters

These bounds are tight under this loose notion of privacy

Open Questions

1) Lower bounds for non-differentiable M-estimators (like median)
2) Non-linear attacks??