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

f could be the
1. average function
2. correlation function
3. classifier
......
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 $f_S(x) = \langle S, x \rangle$

\[ x = (x_1, \ldots, x_n) \]
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 $f_S(x) = \langle S, x \rangle$

$x = (x_1, \ldots, x_n)$
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 \( f_S(x) = \langle S, x \rangle \)

\[
\begin{align*}
S & \xrightarrow{\oplus} x \\
x = (x_1, \ldots, x_n) & \xrightarrow{\text{Private Algorithm}} f_S(x) \xrightarrow{\text{Adversary}} \hat{x} \approx x
\end{align*}
\]

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 $\hat{x} \approx x$.

\[
\begin{bmatrix}
S_1 \\
S_2 \\
\vdots \\
S_m
\end{bmatrix}
\begin{bmatrix}
z \\
\end{bmatrix}
= 
\begin{bmatrix}
A \\
\vdots \\
\end{bmatrix}
\begin{bmatrix}
x \\
\end{bmatrix}
+ 
\begin{bmatrix}
e \\
\end{bmatrix}
\]

Unknown error vector

Natural approach: $\hat{x} = \arg\min_x \| z - Ax \|_p$

- $p=2$: 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

$$\hat{x}_i = \begin{cases} 1 & \text{if the } i^{th} \text{ element of } A_{\text{inv}}z \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$
Least Squares Attack (L₂-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

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

**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

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

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

<table>
<thead>
<tr>
<th></th>
<th>Over 60</th>
<th>Smoking</th>
<th>Exercise</th>
<th>High Blood Pressure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Bob</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Charlie</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Dave</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Available to the adversary**: Non-sensitive
- **Sensitive**

---

---
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 $\arg\min_x \| z - A x \|_p$
Reconstruction from Marginals [KRSU’10]

n people

\[ a_1 \quad \text{.....} \quad a_d \quad x \]

\[ \text{noisy release } \hat{z} \to \text{adversary} \]

\[ \hat{x} \approx x \]

\[ d+1 \text{ attributes} \]

Releasing 3-way marginals

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

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

\[ \text{Solve } \arg \min_x \| z - \begin{pmatrix} (a_1 \odot a_2)^\top \\ (a_1 \odot a_3)^\top \\ (a_{d-1} \odot a_d)^\top \end{pmatrix} A x \|_p \to \hat{x} \]
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 $d \geq \sqrt{n}$.

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

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 $d \geq n \frac{1}{k}$.

\[
\begin{pmatrix}
(a_1 \odot a_2 \cdots \odot a_k)^T \\
(a_1 \odot a_3 \cdots \odot a_{k+1})^T \\
(a_{d-k} \odot a_{d-k+1} \cdots \odot a_d)^T
\end{pmatrix}
\begin{pmatrix}
d \\
\binom{k}{d}
\end{pmatrix}
\times n
\]

Then w.h.p. the least singular value of matrix $A$ is $\Omega(d^{\frac{k}{2}})$. 
Theorem [KRSU’10]: If an algorithm always releases \((k+1)\)-way marginals with \(\min\left\{o\left(d^{\frac{k}{2}}\right), o\left(\sqrt{n}\right)\right\}\) noise per entry then there exists an adversary \(\mathcal{A}\) that w.h.p. can construct \(\hat{x}\) with 
\[
d_{\text{Hamming}}(\hat{x}, x) = o(n).
\]
Theorem [KRSU’10]: If an algorithm always releases \((k+1)\)-way marginals with \(\min\{o(d^{k/2}), o(\sqrt{n})\}\) noise per entry then there exists an adversary \(\mathcal{A}\) that w.h.p. can construct \(\hat{\mathcal{C}}\) 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 \( \leq k \)

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

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

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 $f : \{0, 1\}^3 \rightarrow \{0, 1\}$, let

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

Adversary gets distorted $F(a_1, a_2, x), F(a_1, a_3, x), \ldots, F(a_{d-1}, a_d, x)$
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 $\hat{x}$ that can w.h.p. construct $\hat{x}$ with 
$$d_{\text{Hamming}}(\hat{x}, x) = o(n).$$
**Theorem [KRS’10]:** Let $f : \{0, 1\}^3 \to \{0, 1\}$ 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
\[
\min\{o(\sqrt{d}, o(\sqrt{n}))\} \text{ noise to each release}
\]

There exists an adversary $\mathcal{A}$ 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

\[
\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \ell(\theta; x_i)
\]

**Example:**

\[
\hat{\theta} = \arg\min_{\theta} \sum ||x_i - \theta||_1
\]

\[
\hat{\theta} = \arg\min_{\theta} \sum ||x_i - \theta||_2^2
\]
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

\[
\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \ell(\theta; x_i)
\]

If loss function \( \ell \) is differentiable, then \( \hat{\theta} \) can be obtained by

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

**e.g.,**

\[
\hat{\theta} = \arg\min_{\theta} \sum ||x_i - \theta||_1
\]

\[
\hat{\theta} = \arg\min_{\theta} \sum ||x_i - \theta||_2
\]
Look at Logistic Regression (k = 1)

\[
\begin{array}{cccc}
  a_1 & a_2 & \ldots & a_d & x \\
\end{array}
\]

The logistic regression parameter \( \theta_1 \) between column \( a_1 \) and \( x \) is

\[
\begin{aligned}
  &\log \left( \frac{\zeta_1}{1 - \zeta_1} \right) \\
  &\vdots \\
  &\log \left( \frac{\zeta_n}{1 - \zeta_n} \right) \\
  &= a_1 \theta_1 \text{ where } \zeta_i = \Pr[x_i = 1]
\end{aligned}
\]
Look at Logistic Regression \((k = 1)\)

To estimate \(\theta_1\):
1) Take MLE
2) Set grad. of MLE = 0

MLE estimate \(\hat{\theta}_1\) of \(\theta_1\) is:

\[
\begin{pmatrix}
a_1^T \\
\vdots \\
a_d^T
\end{pmatrix}
\begin{pmatrix}
x
\end{pmatrix} + \begin{pmatrix}
a_1^T \\
\vdots \\
1 + \exp(\hat{\theta}_1 a_{1n}) \\
\end{pmatrix} = 0
\]
Logistic Regression: Linear Reconstruction

\[ \text{from } \hat{\theta}_1: \begin{bmatrix} a_1^\top \end{bmatrix} \begin{bmatrix} x \end{bmatrix} + \begin{bmatrix} a_1^\top \end{bmatrix} \begin{bmatrix} \frac{1 + \exp(\hat{\theta}_1 a_{11})}{\exp(\hat{\theta}_1 a_{11})} \\
\frac{1 + \exp(\hat{\theta}_1 a_{1n})}{\exp(\hat{\theta}_1 a_{1n})} \end{bmatrix} = 0 \]
Logistic Regression: Linear Reconstruction

\[
\begin{align*}
\text{from } \hat{\theta}_1: & \quad \begin{pmatrix} a_1^\top \\ a_2^\top \end{pmatrix} \left\{ \begin{array}{c} x \\ \end{array} \right\} + \begin{pmatrix} \frac{1 + \exp(\hat{\theta}_1 a_{11})}{\exp(\hat{\theta}_1 a_{11})} \\ \frac{1 + \exp(\hat{\theta}_1 a_{1n})}{\exp(\hat{\theta}_1 a_{1n})} \\ \frac{1 + \exp(\hat{\theta}_2 a_{21})}{\exp(\hat{\theta}_2 a_{21})} \\ \frac{1 + \exp(\hat{\theta}_2 a_{2n})}{\exp(\hat{\theta}_2 a_{2n})} \end{array} \right\} = 0
\end{align*}
\]
Logistic Regression: Linear Reconstruction

\[
\begin{align*}
\text{from } \hat{\theta}_1: & \quad \begin{pmatrix} a_1^\top \\ a_2^\top \\ a_3^\top \\ \vdots \end{pmatrix} \begin{pmatrix} x \\ \vdots \end{pmatrix} + \begin{pmatrix} \frac{1+\exp(\hat{\theta}_1 a_{11})}{\exp(\hat{\theta}_1 a_{11})} \\ \frac{1+\exp(\hat{\theta}_1 a_{1n})}{\exp(\hat{\theta}_1 a_{1n})} \\ \frac{1+\exp(\hat{\theta}_2 a_{21})}{\exp(\hat{\theta}_2 a_{21})} \\ \frac{1+\exp(\hat{\theta}_2 a_{2n})}{\exp(\hat{\theta}_2 a_{2n})} \\ \frac{1+\exp(\hat{\theta}_3 a_{31})}{\exp(\hat{\theta}_3 a_{31})} \\ \frac{1+\exp(\hat{\theta}_3 a_{3n})}{\exp(\hat{\theta}_3 a_{3n})} \\ \vdots \end{pmatrix} &= 0
\end{align*}
\]
Logistic Regression: Linear Reconstruction

\[
\begin{align*}
\text{from } \hat{\theta}_1: & \begin{bmatrix} a_1^\top \\ a_2^\top \\ a_3^\top \\ \vdots \\ A \end{bmatrix} \\
\text{from } \hat{\theta}_2: & \begin{bmatrix} a_1^\top \\ a_2^\top \\ a_3^\top \\ \vdots \\ A \end{bmatrix} \\
\text{from } \hat{\theta}_3: & \begin{bmatrix} a_1^\top \\ a_2^\top \\ a_3^\top \\ \vdots \\ A \end{bmatrix}
\end{align*}
\]

\[
\begin{bmatrix} a_1^\top \\ a_2^\top \\ a_3^\top \\ \vdots \\ A \end{bmatrix} x + \begin{bmatrix} 1+\exp(\hat{\theta}_1 a_{11}) \\ \exp(\hat{\theta}_1 a_{11}) \\ \exp(\hat{\theta}_1 a_{1n}) \\ \exp(\hat{\theta}_2 a_{21}) \\ 1+\exp(\hat{\theta}_2 a_{21}) \\ \exp(\hat{\theta}_2 a_{21}) \\ \exp(\hat{\theta}_2 a_{2n}) \\ 1+\exp(\hat{\theta}_3 a_{31}) \\ \exp(\hat{\theta}_3 a_{31}) \\ \exp(\hat{\theta}_3 a_{3n}) \\
\vdots 
\end{bmatrix} b = 0
\]
Logistic Regression: Linear Reconstruction

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
**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 \( \mathcal{A} \) that can w.h.p. construct \( \hat{x} \) with \( d_{\text{Hamming}}(\hat{x}, x) = o(n) \).
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 \( d \geq 2n \).

L2-attack: If for every database \( D \), the algorithm adds 
\[ o\left(\frac{1}{\sqrt{n}}\right) \] noise to each parameter

There exists an adversary \( \mathcal{A} \) 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
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??