Reconstruction Attacks

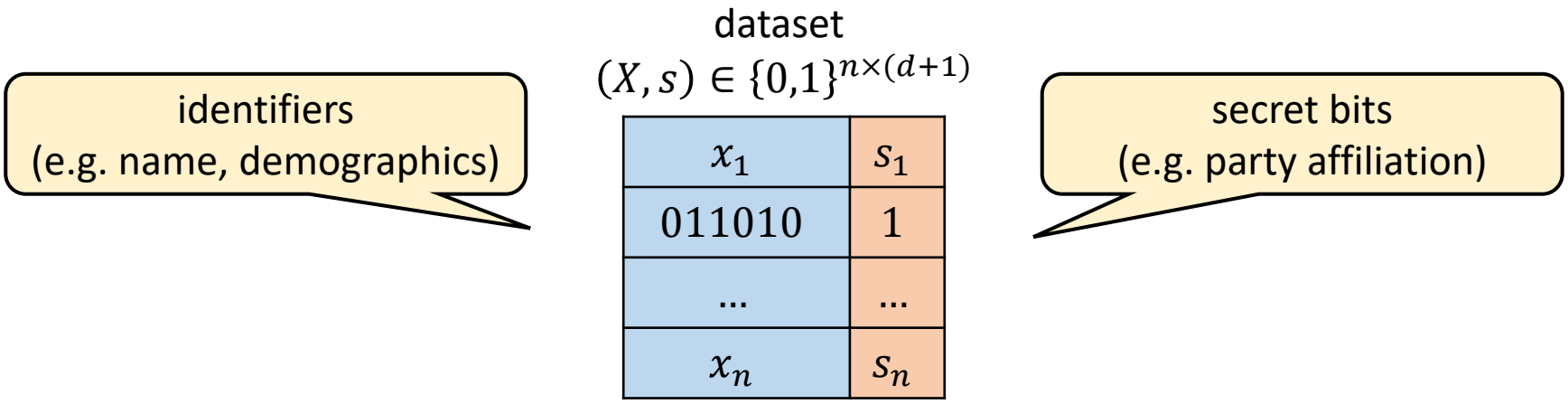
Jonathan Ullman, Northeastern University

Bar Ilan University Winter School: Differential Privacy. February 13, 2017.

Outline

- Reconstruction Attacks [Dinur-Nissim'03]
  - "Releasing overly accurate answers to too many statistics is blatantly non-private."
    - Establishes limits on the accuracy achieved by any private algorithm, not just differentially private ones.
  - Neat connections to linear algebra, discrepancy theory, and error correcting codes.

Modeling Reconstruction



Want to release statistics involving the secret vector.

- Correlation between each attribute j and the secret $\frac{1}{n}\sum_i x_{ij} s_i$
- Statistical queries $\frac{1}{n}\sum_i \phi(x_i) s_i$
- Parameters of a regression model that predicts s_i given x_i

Modeling Reconstruction

dataset

$(X, s) \in \{0,1\}^{n \times (d+1)}$

identifiers
(e.g. name, demographics)

secret bits
(e.g. party affiliation)

x_1	s_1
011010	1
…	…
x_n	s_n

These can all be translated to a linear function Q of the secret vector s.

- Correlation between each attribute j and the secret $\frac{1}{n}\sum_i x_{ij} s_i$ is exactly $\frac{X^T s}{n}$

$$X^T s = \begin{array}{|c|c|c|c|} \hline x_1^T & x_2^T & \dots & x_n^T \\ \hline \end{array} \begin{array}{|c|} \hline s_1 \\ \hline s_2 \\ \hline \dots \\ \hline s_n \\ \hline \end{array}$$

Modeling Reconstruction

dataset

$(X, s) \in \{0,1\}^{n \times (d+1)}$

identifiers
(e.g. name, demographics)

secret bits
(e.g. party affiliation)

x_1	s_1
011010	1
…	…
x_n	s_n

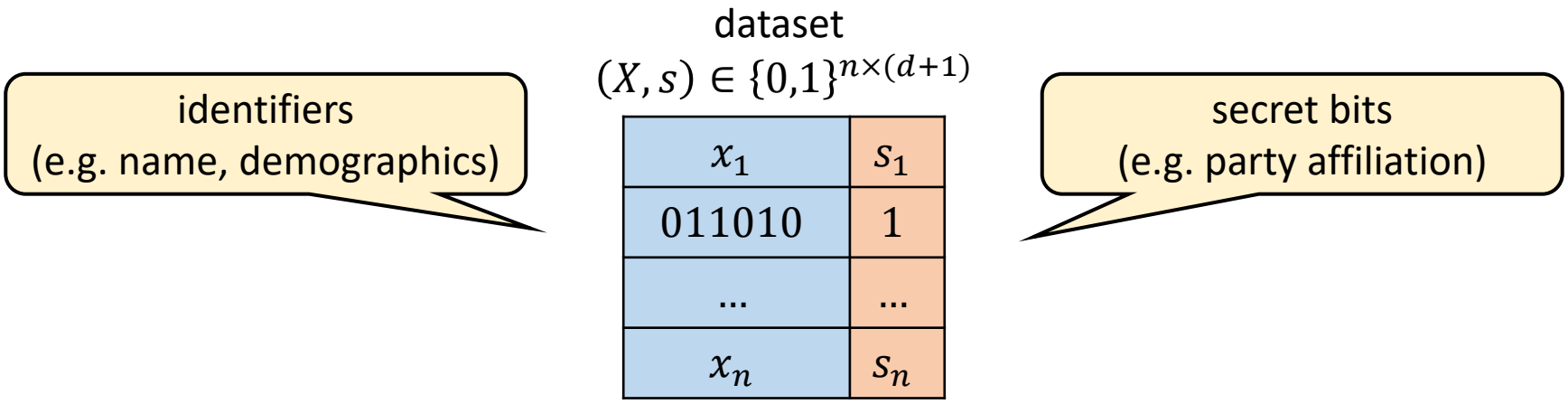
These can all be translated to a linear function Q of the secret vector s.

- Statistical query of the form $\frac{1}{n}\sum_i \phi(x_i)s_i$ is exactly:

$\frac{\phi(x_1)}{n}$	$\frac{\phi(x_2)}{n}$	…	$\frac{\phi(x_n)}{n}$

s_1
s_2
…
s_n

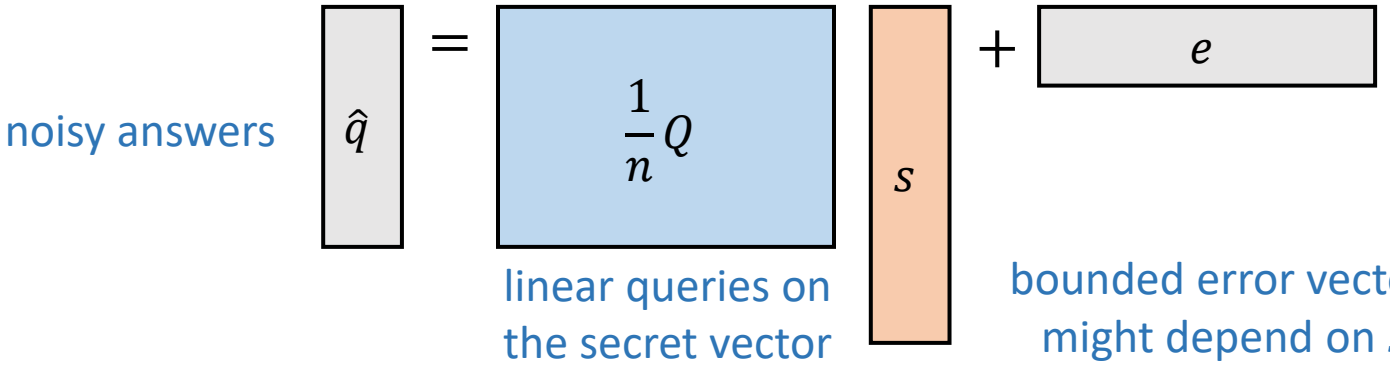
Modeling Reconstruction



These can all be translated to a linear function Q of the secret vector s.

- Parameters of a regression model that predicts s_i given x_i
  - Less immediate, but the optimality of the parameters imply certain linear functions of s

Modeling Reconstruction



So we want to understand the following problem:

Given a matrix $Q \in \{0,1\}^{k \times n}$ of k linear queries, and $\hat{q} = \frac{1}{n}Qs + e$, where $\|e\|_\infty \leq \alpha$ and $s \in \{0,1\}^n$, find $\hat{s}$ such that $\frac{1}{n}Ham(\hat{s}, s) \leq \frac{1}{10}$

If we can solve this problem, then no $\left(\frac{1}{10}, \frac{1}{10}\right)$-dp algorithm A_Q can satisfy $\Pr\left[\left\|A_Q(s) - \frac{1}{n}Qs\right\|_\infty \leq \alpha\right] \geq 9/10$ for all $s \in \{0,1\}^n$.

Exponentially Many Queries

Suppose we consider *all* $\{0,1\}$-valued queries:

$Q \in \{0,1\}^{2^n \times n}$ has one row for every $q \in \{0,1\}^n$

Brute force attack

Input: $\hat{q} = \frac{1}{n} Qs + e$

Output: any $\hat{s} \in \{0,1\}^n$ such that $\left\| \hat{q} - \frac{1}{n} Q\hat{s} \right\|_\infty \leq \alpha$

Theorem [Dinur-Nissim'03]: $\frac{1}{n} Ham(\hat{s}, s) \leq 4\alpha$

Proof: $\left\| \hat{q} - \frac{1}{n} Q\hat{s} \right\| = \left\| \frac{1}{n} Qs + e - \frac{1}{n} Q\hat{s} \right\| \geq \left\| \frac{1}{n} Q(s - \hat{s}) \right\| - \|e\|$

- Suppose $Ham(\hat{s}, s) > 4\alpha n$, then there are $> 2\alpha n$ entries on which $s_i = 1$ but $\hat{s}_i = 0$ (without loss of generality).
- Since Q contains a row that is 1 on exactly these entries, we have $\left\| \frac{1}{n} Q(s - \hat{s}) \right\|_\infty - \|e\|_\infty > 2\alpha - \alpha = \alpha$. Contradiction.

Relationship to Statistical Queries

dataset

$(X, s) \in \{0,1\}^{n \times (d+1)}$

identifiers
(e.g. name, demographics)

secret bits
(e.g. party affiliation)

x_1	s_1
011010	1
…	…
x_n	s_n

As long as $d \geq \log(2n)$, we can obtain any linear function of s.

- Statistical query of the form $\frac{1}{n}\sum_i \phi(x_i)s_i$ is exactly:

$\frac{\phi(x_1)}{n}$	$\frac{\phi(x_2)}{n}$	…	$\frac{\phi(x_n)}{n}$

s_1
s_2
…
s_n

Exponentially Many Queries

Suppose we consider *all* $\{0,1\}$-*valued queries*:
$Q \in \{0,1\}^{2^n \times n}$ has one row for every $q \in \{0,1\}^n$

Brute force attack

Input: $\hat{q} = \frac{1}{n} Qs + e$

Output: any $\hat{s} \in \{0,1\}^n$ such that $\left\| \hat{q} - \frac{1}{n} Q\hat{s} \right\|_\infty \leq \alpha$

Theorem [Dinur-Nissim'03]: $\frac{1}{n} Ham(\hat{s}, s) \leq 4\alpha$

Corollary: if $d \geq \log(2n)$, then there is no differentially private algorithm that answers 2^n arbitrary statistical queries on $x \in \{0,1\}^{n \times d}$ with error $\alpha = o(1)$.

High Accuracy Answers

Suppose we have only a modest number of queries
$Q \in \{0,1\}^{n \times n}$ is a some set of n queries.

Matrix (pseudo-) inversion attack

Input: $\hat{q} = \frac{1}{n} Qs + e$
Let $\tilde{s} = nQ^{inv}\hat{q} = s + nQ^{inv}e$
Output: $\hat{s} = \tilde{s}$ rounded to $\{0,1\}$

Theorem [DN'03, DY'08]: If $n = 2^{\ell}$, and $Q = H_n$ is the Hadamard matrix, then $\frac{1}{n} Ham(\hat{s}, s) \leq 4\alpha^2 n$

Fourier transform of s.

Proof: Useful fact 1: $nQ^{inv} = H_n$
Useful fact 2: All eigenvalues of H_n are $\pm\sqrt{n}$

Therefore, $\|\tilde{s} - s\|_2^2 = \left\|nQ^{inv}e\right\|_2^2 = \|H_n e\|_2^2 \leq n\|e\|_2^2 \leq \alpha^2 n^2$

High Accuracy Answers

Suppose we have only a modest number of queries
$Q \in \{0,1\}^{n \times n}$ is a some set of n queries.

Matrix (pseudo-) inversion attack

Input: $\hat{q} = \frac{1}{n} Qs + e$
Let $\tilde{s} = nQ^{inv}\hat{q} = s + nQ^{inv}e$
Output: $\hat{s} = \tilde{s}$ rounded to $\{0,1\}$

Theorem [DN'03, DY'08]: If $n = 2^{\ell}$, and $Q = H_n$ is the Hadamard matrix, then $\frac{1}{n} Ham(\hat{s}, s) \leq 4\alpha^2 n$

Proof: Now observe that $\|\tilde{s} - s\|_2^2 \geq \frac{1}{4} Ham(\hat{s}, s)$, because if $\hat{s}_i \neq s_i$, then we must have $(\tilde{s}_i - s_i)^2 \geq 1/4$. Rearranging gives

$$\frac{1}{n} Ham(\hat{s}, s) \leq \frac{4}{n} \|\tilde{s} - s\|_2^2 \leq 4\alpha^2 n$$

High Accuracy Answers

Suppose we have only a modest number of queries
$Q \in \{0,1\}^{n \times n}$ is a some set of n queries.

Matrix (pseudo-) inversion attack

Input: $\hat{q} = \frac{1}{n} Qs + e$
Let $\tilde{s} = nQ^{inv}\hat{q} = s + nQ^{inv}e$
Output: $\hat{s} = \tilde{s}$ rounded to $\{0,1\}$

Theorem [DN'03, DY'08]: If $n = 2^{\ell}$, and $Q = H_n$ is the Hadamard matrix, then $\frac{1}{n} Ham(\hat{s}, s) \leq 4\alpha^2 n$

Corollary: if $d \geq \log(2n)$, then there is no differentially private algorithm that answers $2n$ arbitrary statistical queries on a dataset $x \in \{0,1\}^{n \times d}$ with error $\alpha = o\left(\frac{1}{\sqrt{n}}\right)$.

Spectral Bounds

Suppose we have only a modest number of queries
$Q \in \{0,1\}^{n \times n}$ is a some set of n queries.

Matrix (pseudo-) inversion attack

Input: $\hat{q} = \frac{1}{n} Qs + e$
Let $\tilde{s} = nQ^{inv}\hat{q} = s + nQ^{inv}e$
Output: $\hat{s} = \tilde{s}$ rounded to $\{0,1\}$

Theorem [KRSU'10]: For any $k \geq n$ and queries $Q \in \{0,1\}^{k \times n}$,

$$\frac{1}{n} Ham(\hat{s}, s) \leq \frac{4\alpha^2 nk}{\sigma_{min}^2(Q)}$$

Discrepancy Bounds

$Q \in \{0,1\}^{k \times n}$ is an arbitrary set of queries

Brute force attack

Input: $\hat{q} = \frac{1}{n} Qs + e$

Output: any $\hat{s} \in \{0,1\}^n$ such that $\left\| \hat{q} - \frac{1}{n} Q\hat{s} \right\|_\infty \leq \alpha$

Theorem [MN'12]: If $\text{partialdisc}(Q) \geq 2\alpha n$, then $\frac{1}{n} Ham(\hat{s}, s) \leq \frac{1}{10}$

Define the partial discrepancy of a matrix $Q \in \mathbb{R}^{k \times n}$ to be

$$\text{partialdisc}(Q) = \min_{\substack{z \in \{-1,0,1\}^n \\ \|z\|_1 \leq n/10}} \|Qz\|_\infty$$

Theorem [DNT'13]: A related quantity, $\text{hereditarydisc}(Q)$ characterizes the error required to answer Q up to factors of $\text{poly}(d, \log k)$.

High Accuracy Answers

Suppose we consider a *modest number of random queries*
$Q \in \{0,1\}^{k \times n}$ has $n \leq k \leq 2^n$ random rows in $\{0,1\}^n$

Brute force attack

Input: $\hat{q} = \frac{1}{n} Qs + e$

Output: any $\hat{s} \in \{0,1\}^n$ such that $\left\| \hat{q} - \frac{1}{n} Q\hat{s} \right\|_\infty \leq \alpha$

Theorem [Dinur-Nissim'03, Smith]: for every $n \leq k \leq 2^n$, and every $\alpha = o\left(\sqrt{\frac{\ln(k/n)}{n}}\right)$, $\frac{1}{n} Ham(\hat{s}, s) \leq o(1)$

Corollary: if $d \geq \log(2n)$, then there is no differentially private algorithm that answers $k \gg n$ random statistical queries on a dataset $x \in \{0,1\}^{n \times d}$ with error $\alpha = o\left(\frac{\ln(k)}{n}\right)^{1/2}$.

Reconstruction vs. Differential Privacy

Recall: for every d, there is a differentially private algorithm that answers k arbitrary statistical queries on a dataset $x \in \{0,1\}^{n \times d}$ with error $\alpha = \tilde{O}\left(\frac{d \ln(k)}{n}\right)^{1/3}$

Reconstruction vs. Differential Privacy

Later on: for every d, there is a differentially private algorithm that answers k arbitrary statistical queries on a dataset $x \in \{0,1\}^{n \times d}$ with error $\alpha = \tilde{O}\left(\frac{\sqrt{d}\,\ln(k)}{n}\right)^{1/2}$

Corollary: if $d \geq \log(2n)$, there is no differentially private algorithm that answers $k \gg n$ random statistical queries on a dataset $x \in \{0,1\}^{n \times d}$ with error $\alpha = o\left(\frac{\ln(k)}{n}\right)^{1/2}$.

- Reconstruction attacks essentially "characterize" privacy for low-dimensional datasets
- Understanding high-dimensional data requires very different attacks (fingerprinting codes / tracing attacks)

Outline

- Reconstruction Attacks [Dinur-Nissim'03]
  - "Releasing overly accurate answers to too many statistics is blatantly non-private."
    - Establishes limits on the accuracy achieved by any private algorithm, not just differentially private ones
  - Neat connections to linear algebra, discrepancy theory, and error correcting codes