Lecture : Privacy attacks [Dwork et al., "Exposed! A Survey of Attacks on Private Data," 2017]

In this lecture, we shall discuss some well-studied attacks on "aggregate statistics" releases, which compromise the privacy of individuals in the dataset, in some semantic sense.

These attacks, which are broadly called "privacy attacks," are launched by adversaries with one of three possible goals: (a) re-identification (→ or "de-anonymization"), (b) reconstruction of sensitive information, and (c) tracing, or determining an individual's presence in a dataset.

Without loss of generality, a dataset $\mathcal{D} \in \mathbb{D}_n$ (collection of datasets with n users) is a collection of "records" or rows, each of which corresponds to a single user, and contains identity information, "quasi-identifiers" (such as age/gender), and sensitive information.

Example: A dataset $\mathcal{D} \in \mathbb{D}_n$ could consist of records with "Aadhaar number" (or SSN) [identity], age, gender, pincode of neighbourhood (quasi-identifiers), sickle-cell anemia status $\in \{0,1\}$ (0 = No, 1 = Yes). Usually, the identity information is ← Sensitive! → either dropped entirely, or obfuscated via cryptographic methods.

The privacy attacks we shall see work with "aggregate statics" of the dataset (such as counts, mean, etc.) and are yet able to break privacy in a well-accepted sense.

Reconstruction attacks

Consider a dataset $\mathcal{D} \in \mathbb{D}_n$, such that $\mathcal{D} = \{(x_i, s_i) : i \in [n]\}$.

($x_i \in \mathbb{R}^d$: publicly available info. about user i; $s_i \in \mathbb{R}$: sensitive info. about user i.)

To make the construction of privacy-preserving algorithms (or mechanisms) hard, we shall assume that the x_i, $i \in [n]$, are all publicly known. The broad goal of a reconstruction attack by an adversary is the following:

[Informal] **Goal (Reconstruction):** Produce a vector $\underline{c} \in \mathbb{R}^n$ that agrees with $\underline{s}$ in all but $O(n)$ locations.

[Such an attack renders the sequence of queries released "blatantly non-private"]

1. **Reconstruction from linear statistics:** An important class of released aggregate statistics (or "queries") is the class of linear statistics, where each query output q can be written as $q = B\underline{s}$, for some matrix B.

Example: When $\underline{x} \in \{0,1\}^d$ and $s \in \{0,1\}$, the "marginal" of $s=1$ onto a fixed $\underline{x}$ (or, the fraction of rows of $\mathcal{D}$ with public info. $\underline{x}$ and

s being 1) can be thought of as a linear query $q = B\underline{s}$, where

$$B = \frac{1}{n}\underline{X}^T.$$

We focus on the class of fractional linear queries:

Def 1: A fractional linear query $q = q_{\underline{b}}(\underline{s})$ has the form $q = \frac{1}{n}\underline{b}^T\underline{s}$, for some $\underline{b} \in [0,1]^n$. A collection of fractional linear queries is hence identified by the rows of a matrix B, each of whose entries is in $[0,1]$.

We are interested in reconstructing $\underline{s}$ from potentially noisy responses to queries, but under the constraint that the noise is bounded.

Def 2 (B-reconstruction problem): Given a matrix B and a vector $\hat{q} = \frac{1}{n}B\underline{s} + \underline{e}$ where $\|\underline{e}\|_\infty \leq \alpha$ and $\underline{s} \in \{0,1\}^n$, find $\hat{\underline{s}}$ s.t. $d_H(\underline{s}, \hat{\underline{s}}) \leq \frac{n}{10}$.

d_H: Hamming distance

In the next section, we shall discuss the setting where exponentially many (in n) queries are raised.

(A) Exponential # Queries

Consider the setting where B corresponds to "all possible sums of subsets of coordinates in $\underline{s}$", i.e., $B \in M_{2^n \times n}$, with $B_{ij} \in \{0,1\}$,

$\forall i, j$, and the rows of B are all the vectors in $\{0,1\}^n$.

Theorem 1 (Dinur and Nissim (2003)): For the above choice of B, given α-accurate queries, there is a reconstruction attack that solves the B-reconstruction problem with reconstruction error $= \frac{d_H(\underline{s}, \hat{\underline{s}})}{n}$ at most 4α, for all $\alpha > 0$.

Proof: We simply return $\hat{\underline{s}}$ to be any vector s.t. $\frac{1}{n} B\hat{\underline{s}}$ differs from all the query outputs $\hat{q}_{\underline{a}}$ by at most α. Clearly one such $\hat{\underline{s}}$ exists [which one?].

Now, let $\underline{b}_0 = \underline{s}$ and $\underline{b}_1 = \bar{\underline{s}}$ (complement of $\underline{s}$). Then, we have that $\left| \frac{\langle \underline{b}_0, \hat{\underline{s}} \rangle}{n} - \hat{q}_{\underline{b}_0} \right| \leq \alpha$ and $\left| \frac{\langle \underline{b}_0, \underline{s} \rangle}{n} - \hat{q}_{\underline{b}_0} \right| \leq \alpha$.

($\langle \underline{b}_0, \hat{\underline{s}} \rangle = w_H(\underline{s} \cdot \hat{\underline{s}})$; $\langle \underline{b}_0, \underline{s} \rangle = w_H(\underline{s})$)

This implies that $\underline{s}$ & $\hat{\underline{s}}$ disagree on at most $2\alpha n$ positions where $\underline{s}$ is $\underline{1}$. By the same argument using $\underline{b}_1$, we see that $\underline{s}$ & $\hat{\underline{s}}$ disagree on at most $2\alpha n$ positions when $\underline{s}$ is 0. ▨

(B) Polynomial # Queries:

We provide a (mostly complete) proof sketch of the following result:

Theorem 2 (Dwork & Yekhanin (2008)): There exists a matrix $B \in \{0,1\}^{2n \times n}$ and an attack A that solves the B-reconstruction problem with reconstruction error at most $16\alpha^2 n$, when the query responses are α-accurate, when $n = 2^l$, for some integer $l \geq 1$.

Proof: Suppose that $n = 2^l$, for some integer $l \geq 1$. We focus on $B \in \{\pm 1\}^{2n \times n}$, since one can always simulate a query with ± 1 coefficients using 2 queries with 0/1-coefficients. [Why?]

Let B be the $n \times n$ Hadamard matrix H_l (and hence the corresp. 0/1-matrix $\bar{B}$ is a $2n \times n$ matrix) defined recursively as:

$$H_0 = (1) \quad \text{and} \quad H_{i+1} = \begin{pmatrix} H_i & H_i \\ H_i & -H_i \end{pmatrix}.$$

HW: Verify that $H_l^2 = n I_n$, for all $n \geq 1$.

The above fact implies that $(H_l)^{-1} = \frac{1}{n} H_l$ and that the eigenvalues of H_l are all in $\{-\sqrt{n}, +\sqrt{n}\}$.

Step 1: Now, given $\hat{q} = \frac{1}{n} H_l \underline{s} + \underline{e}$, the attack obtains

$$\underline{a} = H_l \hat{q} = \underline{s} + H_l \underline{e}.$$

Now, note that $\|\underline{e}\|_2^2 \leq \alpha^2 n$ and $\|H_l \underline{e}\|_2^2 \leq \alpha^2 n^2$. [Why?]

Thus, we have that $\|\underline{a} - \underline{s}\|_2^2 \leq \alpha^2 n^2$.

Step 2: The attacker then rounds $\underline{a}$ to $\{0,1\}^n$, entry-wise, giving $\hat{\underline{s}}$.

The following claim then holds:

Claim: We have $\frac{d_H(\underline{s}, \hat{\underline{s}})}{n} \leq \frac{4\|\underline{a} - \underline{s}\|_2^2}{n} \leq 4\alpha^2 n$.

Proving the claim concludes the proof of Thm 2, upto constant factors (associated with moving from queries with ± 1 coefficients to $\{0,1\}$).

Proof of Claim: Note that $\underline{s}$ & $\hat{\underline{s}}$ differ in only those positions $j \in [n]$ where $|a_j - s_j| \geq 1/2$ (and hence $(a_j - s_j)^2 \geq 1/4$). By Markov's inequality, we obtain that the fraction of these positions is at most $\frac{4\|\underline{a} - \underline{s}\|_2^2}{n}$. ▨

▨

2. Tracing Attacks

In this section, we briefly discuss a model for tracing an individual based on hypothesis testing, given query outputs. This model was famously used in Homer et al. (2008) on genome-wide association data.

→ We have a population with a product distribution P over $\{+1,-1\}^d$. Let $\underline{p} = \mathbb{E}_P[X]$.

→ A dataset $\mathcal{D} = \{X_1, \dots, X_n\}$ is drawn i.i.d. from P.

→ The sample mean

$$q_f = \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}_{i \sim \mathrm{Unif}([n])}[X_i]$$

is released.

→ The attacker is provided with a collection $R = \{R_1, \dots, R_m\} \overset{iid}{\sim} P$ of "reference samples", in addition to q_f.

The attack is designed based on hypothesis testing; given sample y.

$$H_0: \; y \notin \mathcal{D} \text{ and } y \overset{iid}{\sim} P$$

$$H_1: \; y \in \mathcal{D}$$

Under H_0, y has mean p, while under H_1, y has mean q. The attacker obtains an approximation $\hat{p}$ of p, using R, and then carries out a log-likelihood test to decide which of the hypotheses (H_0 or H_1) is true.

<u>Theorem</u> (Sankararaman et al. (2009)): There is an attack $A(y, q_f, \Sigma)$ such that for any non-trivial product distribution P on sufficiently large dimension d $(= O(n \log(1/\delta)))$, we have

(a) $\Pr[A(y, q_f, \Sigma) = \text{IN} \mid H_1] \geq 1 - \delta$, and

(b) $\Pr[A(y, q_f, \Sigma) = \text{OUT} \mid H_0] \geq 1 - \delta$.

Informally,

The framework of differential privacy (DP) allows for the design of privacy-preserving mechanisms that can answer a number of queries that is roughly equal to those we obtained as upper bounds, via reconstruction attacks, for the same level of accuracy.

A discussion on a simple mechanism to protect against reconstruction attacks:

Earlier, we had seen that if the error α in reconstruction is required to be $\ll 1/\sqrt{n}$, then, reconstruction is possible. However, can we come up with a scheme for preventing reconstruction, when the reconstruction error is larger, for the same number of queries?

Answer: Subsampling

Consider a dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$. Let us take a random subsample of this dataset, i.e., we construct

$$\overline{\mathcal{D}} = \{y_1, \ldots, y_m\},$$

where $y_i \sim \text{Unif}(\mathcal{D})$, where $\frac{m}{n} = c \in (0,1)$.

Now, observe that for statistics of the form $q = \sum_{i=1}^{n} \phi(x_i)$, an

unbiased estimate is $\overline{q} = \frac{1}{C} \cdot \sum_{i=1}^{m} \phi(Y_i)$. [Show this.]

The following result can then be shown via standard concentration arguments:

HW: Show that for any set of statistics $\phi_1, \dots, \phi_K$, we have that with probability at least $1-\delta$,

$$\forall i \in [K], \quad \left| \frac{1}{C} \cdot \sum_{j=1}^{m} \phi_i(Y_j) - \sum_{i=1}^{n} \phi_i(x_i) \right| \leq O\left(\sqrt{n \log k} \cdot \log\left(\tfrac{1}{\delta}\right)\right)$$

Thus, picking $K = \Theta(n)$ [same # queries as in Dwork-Yekhanin (2008)], we see that we can answer queries with error $\frac{\sqrt{\log n}}{\sqrt{n}}$. Thus, it appears that the $\approx \frac{1}{\sqrt{n}}$ threshold on query errors is tight...