Section 2: Attacks Review

CS 208 Applied Privacy for Data Science, Spring 2025

February 3, 2025

1 Agenda

- Discuss any lingering questions related to PSET 1.
- Discuss reidentification and reconstruction attacks.
- Create and run a reconstruction attack.

2 Overview of Reidentification and Reconstruction Attacks

In class, we discussed that re-identification attacks occur when an adversary links anonymized data with auxiliary datasets to re-associate records with personally identifiable information. Techniques such as k-anonymity aim to prevent re-identification by ensuring that each set of unique quasi-identifiers appears in the dataset at least k times. However, we also learned that attacks against k-anonymity are still possible since, in practice, *any attribute* can be a quasi-identifier, as it's challenging to anticipate what auxiliary information an attacker might possess.

In a reconstruction attack, the dataset itself is unavailable to the attacker, but the attacker receives aggregate statistics over the dataset. For the case of subset sum queries, we learned about the following influential result from Dinur and Nissim [1].

Theorem 2.1 ([1])**.** *If an analyst is allowed to ask $O(n)$ subset queries, and the curator adds noise with a bound $E = o(\sqrt{n})$, then a computationally efficient adversary can reconstruct a $1 - o(1)$ fraction of the dataset with high probability.*

Recall that for a *subset sum* query over a dataset x, the adversary receives answers of the form $q_S(x) = \sum_{i \in S} x_i$ for any $S \subseteq [n]$. Thus, the theorem states that an adversary that can specify $m = n$ subsets S_j and receive answers a_j such that $|a_j - q_{S_j}(x)| \leq E = o(\sqrt{n})$ can reconstruct a $1 - o(1)$ fraction of the bits x_i. The lesson here is that if answers to $q_S(x)$ are too accurate (in particular, within $o(\sqrt{n})$ of the truth), the adversary can reconstruct with high probability. See lecture notes for more details.

Warmup question: We have discussed two types of privacy attacks: *re-identification attacks* and *reconstruction attacks*. Briefly discuss the differences between the two types of attacks. Does one attack seem stronger than the other? Does one attack imply the other?

3 Performing a Reconstruction Attack

In this exercise, we investigate the privacy implications of releasing too many accurate summary statistics. We consider a **diabetes dataset** from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDKD) (publicly available on Kaggle[1]) which contains sensitive medical information about individuals. In particular, the dataset contains an **outcome** column (denoted by y), where

$$y[j] = \begin{cases} 1, & \text{if the individual corresponding to row } j \text{ has diabetes,} \\ 0, & \text{otherwise.} \end{cases}$$

For this exercise, we will consider the first $n = 500$ records from the dataset.

Imagine that the NIDDKD dataset is not publicly available. Instead, a privacy-conscious data curator at the NIDDKD decides to release summary statistics rather than raw data to protect individual privacy. The curator allows an analyst to specify a set of m *subset sum queries* and provides *noisy answers* in return.

3.1 Attack Setup

Each query is defined by a binary indicator vector $q_i \in \{0,1\}^n$, and the corresponding query answer is the sum over the selected records:

$$a_i = \sum_{j=1}^{n} q_i[j] \cdot y[j], \quad \text{for } i = 1, \ldots, m$$

To further protect privacy, the curator adds noise to the query results. Specifically, the released (noisy) query answer is given by:

$$\tilde{a}_i = a_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

The goal of the attacker is to specify a set of m queries such that $m = O(n)$ and use the noisy query answers $\tilde{a} \in \mathbb{R}^m$ to reconstruct the sensitive attribute column y of the dataset.

3.2 Regression-based Reconstruction

We will demonstrate that an attacker can attempt a reconstruction attack by formulating the following least squares problem:

$$\hat{x} = \min_{x \in \mathbb{R}^n} \|Qx - \tilde{a}\|_2^2$$

where Q is defined as the query matrix:

$$Q = \begin{bmatrix} q_1[0] & q_1[1] & \ldots & q_1[n-1] \\ q_2[0] & q_2[1] & \ldots & q_2[n-1] \\ \vdots & \vdots & \ddots & \vdots \\ q_m[0] & q_m[1] & \ldots & q_m[n-1] \end{bmatrix} \in \{0,1\}^{m \times n},$$

[1] `https://www.kaggle.com/datasets/mathchi/diabetes-data-set`

After computing a solution $\hat{x}$, the attacker rounds the values to $\{0, 1\}$.

$$\hat{y}_j = \mathbf{1}_{\{\hat{x}_j \geq 0.5\}}, \quad \text{for } j = 1, \ldots, n$$

In this exercise, we will implement the reconstruction attack and study how reconstruction accuracy varies as a function of:

- The number of queries m
- The noise level σ

Use the starter code available here: Google Colab Notebook to implement your reconstruction attack.

3.3 Some Questions

- What do we expect the success rate of the attacker to be if they just apply random guessing to the `outcome` column?
- For $m = n$, at what level of σ does the reconstruction attack start to fail (achieves roughly the same accuracy as the random guessing baseline)?

4 Theory exercises

Note: the following exercises will not be covered in section. We provide them in these notes as a reference.

Claim 4.1 (A Chernoff-Hoeffding Bound)**.** *For $i = 1, \ldots, k$, let X_i be an independent random variable within $[a, b]$ with mean μ. Then,*

$$\Pr\left[\frac{1}{k}\sum_{i=1}^{k} X_i - \mu \geq t\right] \leq \exp\left(-\frac{2t^2 k}{(b-a)^2}\right).$$

Exercise 4.2. Show that subsampling k out of n rows allows us to estimate m averages each to within $\pm O\left(\frac{1}{\sqrt{k}}\sqrt{\log(m)}\right)$.

Solution. First, we recall the Chernoff bound (above) where we assume the bounds of the random variables (i.e., a, b) are constant.

$$\Pr[\frac{1}{k}\sum_{i=1}^{k}(X_i - \mu) \geq t] \leq e^{-\Omega(2kt^2)}$$

Next, we set $t = \sqrt{\frac{\log(m)}{k}}$. Then, we can re-write the right side of the inequality as

$$\begin{aligned} e^{-2kt^2} &= e^{-2k\left(\sqrt{\frac{\log(m)}{k}}\right)^2} \\ &= e^{-2\log(m)} \quad (\leq e^{-2\ln(m)} \text{ for } m > 1) \\ &\leq \frac{1}{m^2} \end{aligned}$$

Taking the union bound over m queries, we have that that the probability of any sample mean deviating from the true mean by greater than $\sqrt{\frac{\log(m)}{k}}$ is upper bounded by $\frac{1}{m}$. Thus, subsampling k of n rows allows us to estimate m averages each to within $\pm O\left(\frac{1}{\sqrt{k}}\sqrt{\log(m)}\right)$ with high probability.

Exercise 4.3. Show that on average, we can successfully trace $\Omega(\frac{1}{\alpha^2})$ individuals in a dataset at best.

Solution. If we subsample $\Omega(\frac{1}{\alpha^2})$ rows, then the sample mean has a standard deviation of

$$O\left(\frac{1}{\sqrt{1/\alpha^2}}\right) = O\left(\frac{1}{1/\alpha}\right) = O(\alpha)$$

Recall our assumption that the error in answers is bounded by α, $|a_j - \bar{x}_j| < \alpha$, with high probability. Since subsampling $\Omega(\frac{1}{\alpha^2})$ rows maxes out the allowed error to $O(\alpha)$, we are on average only able to trace the individuals included in the subsample.

Exercise 4.4. Show that $\langle y - p, a - p\rangle$ approaches a normal distribution.

Solution. We can express this inner product as follows:

$$\langle y - p, a - p \rangle = \sum_{j=1}^{d} (y_j - p_j) \cdot (a_j - p_j)$$

First, note that p and a are given to the adversary, and our only randomness comes from the sampling of y. From our assumption that the d attributes are independent, we know that each y_j is an independent draw from a Bernoulli distribution with expectation p_j.

Since the d attributes of y are independent samples, we can apply the central limit theorem: if n is sufficiently large, then the distribution of $\langle y - p, a - p \rangle$ is approximately normal, $N(0, \sigma^2)$. (This is because a is a noisy mean of the form $\bar{a}_n = \frac{1}{n} \sum_{i=1}^{n} a_i$.)

The variance σ^2 is equal to $\sum_{j=1}^{d} (a_j - p_j)^2 \cdot p_j \cdot (1 - p_j)$. Note that σ^2 is at most $\frac{d}{4}$, since $p_j = 0.5$ yields the maximum variance of $(p_j)(1 - p_j) = 0.25$ for every j. However, we can get a better hypothesis test by using the actual variance instead of this upper bound $\frac{d}{4}$. This is why using a normal approximation in the homework gives us a $T_{a,p}$ that is better than the T given by the Hoeffding bound.

Exercise 4.5. To make the false positive probability at most δ, show that we can choose $T = O\big(\sqrt{d \log(1/\delta)}\big)$ using Chernoff-Hoeffding bounds.

Solution. We translate the Hoeffding bound for this problem. Note that since $(y_j - p_j) \cdot (a_j - p_j)$ is in $[-1, 1]$, we can replace $(b - a)^2$ with 4.

$$\Pr[\sum_{j=1}^{d} (y_j - p_j) \cdot (a_j - p_j) \geq T] \leq e^{-\Omega\left(\frac{2T^2}{4d}\right)}$$

Then, we simply solve for T in terms of d and δ.

$$\begin{aligned}
e^{-\frac{2T^2}{4d}} &= \delta \\
-\frac{T^2}{2d} &= \ln(\delta) \\
\frac{T^2}{2d} &= -\ln(\delta) = \ln(1/\delta) \\
T^2 &= 2d \ln(1/\delta) \\
T &= \sqrt{2d \ln(1/\delta)} = O\big(\sqrt{d \log(1/\delta)}\big)
\end{aligned}$$

References

[1] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 202–210, 2003.