CSC2412: Algorithms for Private Data Analysis Updated Fall 2020

Basic Model, Attacks on Privacy

Aleksandar Nikolov *Scribe: Eric Bannatyne*

1 Introduction

Private data analysis is concerned with studying mechanisms that enable the analysis of datasets containing sensitive information in a way that provides useful information about the data as a whole, without compromising the privacy of any individuals in the dataset. One could consider many potential definitions of privacy, often concerned with legal, ethical, and otherwise philosophical aspects of privacy. We will focus on definitions of privacy that lend themselves to mathematical analysis, through the framework of *differential privacy*.

For now, we will hold off on providing a precise, mathematical definition of privacy. Instead, we will look at a few scenarios in which we can definitely say that we do *not* have any privacy protection, because we can describe attacks that recover the private data. Even though we have not yet defined exactly what privacy is, it will be clear that algorithms that make these attacks possible do not satisfy reasonable definition of privacy. This will provide some motivation for technical definitions of privacy that we will see later.

1.1 The Setup

In all of our examples, we will consider a database represented by a sequence $X \in \mathcal{X}^n$ of n data points (or database rows), where $\mathcal{X}$ is the universe of possible data points. We will typically have $\mathcal{X} = \{0,1\}^d$ for some d, in which case X is an $n \times d$ table of boolean values. The notation x_i refers to the ith row of X. Typically, we think of a row x_i of a database as a record containing sensitive information about a single person.

Next, we assume that the database X is held by a trusted curator. Rather than directly accessing data in X, we issue *queries* $q_1, q_2, \ldots, q_m$ to the curator, where each query comes from some class of functions. The curator answers queries $q_1, \ldots, q_m$ according to some algorithm or "mechanism" $\mathcal{M}(X, q_1, \ldots, q_m)$ that outputs approximations to $q_1(X), \ldots, q_m(X)$. Some examples of queries that we might want to consider are *counting queries*: Given some predicate $q : \mathcal{X} \to \{0,1\}$,

$$q(X) = \frac{1}{n}\sum_{i=1}^{n} q(x_i)$$

is the fraction of rows in X that satisfy q. (Note that we have slightly overloaded the notation.)

In a broad sense, we wish to design mechanisms $\mathcal{M}$ that enable us to compute useful statistics from our data X, while preventing privacy attacks that might dissuade people from contributing to the database. Thus the central question of private data analysis is: If we can query a database via "aggregate" queries q—that is, they don't depend "too much" on individual rows—can we approximate $q(X)$ without revealing too much about any individual row x_i? One of our goals will be to state this question precisely and mathematically.

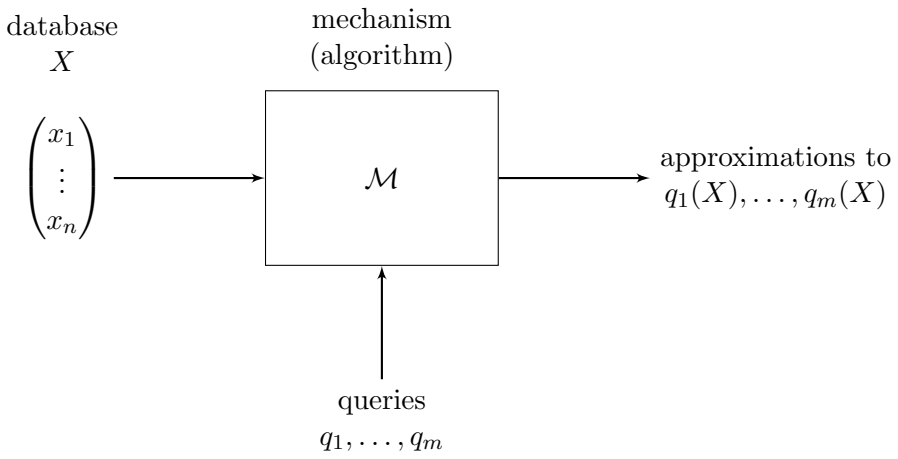

Figure 1: The setting in which we describe attacks on alleged private mechanisms $\mathcal{M}$.

2 Attacks on Simple Mechanisms

2.1 Reconstruction Attack Against Correlation Queries

In this section we consider some negative results, by studying *reconstruction attacks*. The goal of a reconstruction attack is to reconstruct most of the rows of a database from noisy answers to certain queries. This will provide us with examples of (blatant) *non-privacy*, introduced by Dinur and Nissim in [2].

Suppose that each database row has the form $x_i = (y_i, b_i)$, where $y_i \in \mathcal{Y}$ is some identifying information already known to the attacker, and $b_i \in \{0, 1\}$ is a secret bit encoding private information. We assume that the y values are all distinct. We will consider *correlation queries* on a database $X \in \mathcal{X}^n$ over the universe $\mathcal{X} = \mathcal{Y} \times \{0, 1\}$.

Definition 1. *Let $\pi : \mathcal{Y} \to \{0, 1\}$ be a predicate over the the known information. Then the correlation query associated with π is defined by*

$$q_\pi(X) = \frac{1}{n} \sum_{i=1}^{n} \pi(y_i) b_i.$$

In other words, a correlation query asks "What fraction of the database rows satisfy the predicate π and have their sensitive bit set to 1?" Intuitively, such questions give only aggregate information about the sensitive bits. The reconstruction attacks show that, only using the already known information given by the y_i, and these aggregate correlation queries, the attacker can reconstruct most of the private bits.

To quantify the distance between two bit vectors, we use the normalized hamming distance.

Definition 2. *Let $b, b' \in \{0, 1\}^n$. The normalized hamming distance between b and b' is given by $d_H(b, b') = \frac{1}{n}|\{i : b_i \neq b'_i\}|$.*

The following theorem will show that, given noisy answers to every possible correlation against a database with n rows, it is possible to reconstruct most of the original database.

Theorem 3. *Let $\alpha \in [0,1]$, and suppose that, for every predicate $\pi : \mathcal{Y} \to \{0,1\}$, the attacker is given an approximate answer $a_\pi \in \mathbb{R}$ to the correlation query such that $|a_\pi - q_\pi(X)| \leq \alpha$. Then the attacker can compute a b' such that $d_H(b,b') \leq 4\alpha$.*

Proof. Given a_π for every $\pi : \mathcal{Y} \to \{0,1\}$ as in the statement of the theorem, the reconstruction attack will be to simply output any $b' \in \{0,1\}^n$ such that

$$\left| a_\pi - \frac{1}{n}\sum_{i=1}^{n} \pi(y_i) b'_i \right| \leq \alpha \quad \forall \pi : \mathcal{Y} \to \{0,1\}. \tag{1}$$

First, note that the attacker can indeed check (1), because $y_1, \ldots, y_n$ are known to him. Second, note that a b' satisfying (1) must exist, since, by our assumption on the a_π, b satisfies (1) (as do, potentially, many other bit vectors). Finally, note that the conditions above imply that, for any $\pi : \mathcal{Y} \to \{0,1\}$,

$$\left| q_\pi(X) - \frac{1}{n}\sum_{i=1}^{n} \pi(y_i) b'_i \right| \leq |q_\pi(X) - a_\pi| + \left| a_\pi - \frac{1}{n}\sum_{i=1}^{n} \pi(y_i) b'_i \right| \leq 2\alpha. \tag{2}$$

To show that b' is indeed an approximate reconstruction of b, we bound the normalized hamming distance between the two vectors. Let us consider two predicates π_{10} and π_{01} defined to take value 0 on $\mathcal{Y} \setminus \{y_1, \ldots, y_n\}$, and defined to take the following values on $\{y_1, \ldots, y_n\}$:

$$\pi_{10}(y_i) = \begin{cases} 1 & b'_i = 1, b_i = 0 \\ 0 & \text{otherwise} \end{cases},$$

$$\pi_{01}(y_i) = \begin{cases} 1 & b'_i = 0, b_i = 1 \\ 0 & \text{otherwise} \end{cases}.$$

We now have

$$\begin{aligned} d_H(b,b') &= \frac{1}{n}|\{i : b'_i = 1, b_i = 0\}| + \frac{1}{n}|\{i : b'_i = 0, b_i = 1\}| \\ &= \frac{1}{n}\sum_{i=1}^{n} \pi_{10}(y_i) b'_i - q_{\pi_{10}}(X) + q_{\pi_{01}}(X) - \frac{1}{n}\sum_{i=1}^{n} \pi_{01}(y_i) b'_i \\ &\leq 2\alpha + 2\alpha = 4\alpha. \end{aligned}$$

In the final inequality we used (2). □

While the above theorem certainly demonstrates that the private data can be reconstructed from noisy answers to correlation queries, the attack is not exactly computationally feasible, requiring noisy answers to *all* possible correlation queries. We will see that this can be improved, by showing that it is possible to approximately reconstruct the private data using only a number of correlation queries that is at most linear in the size of X. In the proof, we will use the following lemma.

Lemma 4. *Let $Z_1, \ldots, Z_n$ be independent random variables such that for all i, we have $|Z_i| \leq 1$, and let $Z = Z_1 + \cdots + Z_n$. There exists an absolute constant $c > 0$ such that, for any $\theta \geq 0$,*

$$\Pr[|Z| \geq \theta\sqrt{\mathbb{E}[Z^2]}] \geq \frac{c(1-\theta^2)^2}{1 + (1/\mathbb{E}[Z^2])}.$$

Proof. Our strategy is to apply the Paley-Zygmund inequality, which tells us that

$$\Pr[|Z| \geq \theta\sqrt{\mathbb{E}[Z^2]}] = \Pr[|Z|^2 \geq \theta^2\mathbb{E}[Z^2]] \geq (1-\theta^2)^2\frac{\mathbb{E}[Z^2]^2}{\mathbb{E}[Z^4]}. \tag{3}$$

In order to use the inequality, we need to show that $\mathbb{E}[Z^4]$ is not much larger than $\mathbb{E}[Z^2]^2$. This is easier to show directly when the summands have expectation 0, so we will first center our variables Z_i by subtracting their expectations, and show the bound on the fourth moment of the centered variables. Then we will come back to the original variables.

Let us then write $Y_i = Z_i - \mathbb{E}[Z_i]$, and let $Y = Y_1 + \cdots + Y_n$. We make the following observations:

- the Y_i are independent;
- $|Y_i| \leq |Z_i| + |\mathbb{E}[Z_i]| \leq 2$ for all i;
- $\mathbb{E}[Y_i] = 0$ for all i, and, therefore, $\mathbb{E}[Y] - 0$;
- $Z = Y + \mathbb{E}[Z]$;
- $\mathbb{E}[Z^2] = \mathbb{E}[Y^2] + \mathbb{E}[Z]^2$.

By independence, and since each Y_i has expectation 0, we have that, for any $i \neq j$, $\mathbb{E}[Y_iY_j] = \mathbb{E}[Y_i]\mathbb{E}[Y_j] = 0$. Therefore, we have

$$\mathbb{E}[Y^2] = \sum_{i=1}^{n}\mathbb{E}[Y_i^2].$$

Similarly, using the multinomial theorem, and the observation that, for any $i \neq j$, $\mathbb{E}[Y_i^3Y_j] = \mathbb{E}[Y_i^3]\mathbb{E}[Y_j] = 0$, we have the following bound.

$$\begin{aligned}
\mathbb{E}[Y^4] &= \sum_{i=1}^{n}\mathbb{E}[Y_i^4] + 12\sum_{i<j}\mathbb{E}[Y_i^2]\mathbb{E}[Y_j^2] \\
&= 6\mathbb{E}[Y^2]^2 + \sum_{i=1}^{n}\left(\mathbb{E}[Y_i^4] - 6\mathbb{E}[Y_i^2]^2\right) \\
&\leq 6\mathbb{E}[Y^2]^2 + \sum_{i=1}^{n}\left(4\mathbb{E}[Y_i^2] - 6\mathbb{E}[Y_i^2]^2\right) \\
&\leq 6\mathbb{E}[Y^2]^2 + 4\sum_{i=1}^{n}\mathbb{E}[Y_i^2] = 6\mathbb{E}[Y^2]^2 + 4\mathbb{E}[Y^2]. \qquad (\star)
\end{aligned}$$

Here, in the third line, we used $|Y_i| \leq 2$ to derive $Y_i^4 \leq 4Y_i^2$.

Returning to Z, we have

$$\begin{aligned}
\mathbb{E}[Z^4] &= \mathbb{E}[(Y + \mathbb{E}[Z])^4] \\
&\leq 8\mathbb{E}[Y^4] + 8\mathbb{E}[Z]^4 \\
&\leq 48\mathbb{E}[Y^2]^2 + 32\mathbb{E}[Y^2] + 8\mathbb{E}[Z]^4 \\
&\leq 48(\mathbb{E}[Y^2]^2 + \mathbb{E}[Z]^4) + 32\mathbb{E}[Z^2] \\
&\leq 48(\mathbb{E}[Y^2] + \mathbb{E}[Z]^2)^2 + 32\mathbb{E}[Z^2] = 48\mathbb{E}[Z^2]^2 + 32\mathbb{E}[Z^2]. \qquad (\star)
\end{aligned}$$

Above, in the second line we used the inequality $(a+b)^4 \leq 8a^4 + 8b^4$, valid for all real numbers a and b. Then the third line follows from $(\star)$, the fourth line from $\mathbb{E}[Y^2] \leq \mathbb{E}[Z^2]$, and the fifth line follows from the inequality $a^2 + b^2 \leq (|a| + |b|)^2$, also valid for all for all real numbers a and b.

Now we can plug back into (3), and we get

$$\Pr[|Z| \geq \theta\sqrt{\mathbb{E}[Z^2]}] \geq (1-\theta^2)^2 \frac{\mathbb{E}[Z^2]^2}{48\mathbb{E}[Z^2]^2 + 32\mathbb{E}[Z^2]} \geq \frac{(1-\theta^2)^2}{48\left(1 + \frac{1}{\mathbb{E}[Z^2]}\right)}. \qquad \square$$

Theorem 5. *Fix any $\alpha \in (0,1)$. For any $y_1, \ldots, y_n$, there exist $k = O(n)$ predicates $\pi_1, \ldots, \pi_k : \mathcal{Y} \to \{0,1\}$ such that, given $a_1, \ldots, a_k$ satisfying $|a_i - q_{\pi_i}(X)| \leq \frac{\alpha}{\sqrt{n}}$ for all i, the attacker can compute a $b' \in \{0,1\}^n$ such that $d_H(b, b') \leq O(\alpha^2)$.*

Proof. Let k be some integer, whose value will be determined later. Choose correlation queries $\pi_1, \ldots, \pi_k$ independently and uniformly at random from the set all possible predicates. In other words, for each i, and each $y \in \mathcal{Y}$, we independently pick $\pi_i(y)$ to be 0 with probability $\frac{1}{2}$, and 1 with probability $\frac{1}{2}$. The idea will be to show that, when k is sufficiently large, these predicates satisfy the conditions of the theorem with high probability.

The attack itself is similar to the one we already saw. Given $a_1, \ldots, a_k$, the attacker outputs any $b' \in \{0,1\}^n$ such that

$$\left| a_i - \frac{1}{n}\sum_{j=1}^{n} \pi_i(y_j) b'_j \right| \leq \frac{\alpha}{\sqrt{n}} \quad \forall i$$

Once again, the attack can be executed because it only involves information known to the attacker; also, there necessarily exists a bit vector b' satisfying the conditions above, because b is one such bit vector.

As before, the definition of the attack implies that b' satisfies that, for every i,

$$\left| q_{\pi_i}(X) - \frac{1}{n}\sum_{j=1}^{n} \pi_i(y_j) b'_j \right| \leq |q_{\pi_i}(X) - a_i| + \left| a_i - \frac{1}{n}\sum_{j=1}^{n} \pi_i(y_j) b'_j \right| \leq \frac{2\alpha}{\sqrt{n}}.$$

Unwrapping the definition of $q_{\pi_i}(X)$, this is equivalent to

$$\left| \sum_{j=1}^{n} \pi_i(y_j)(b_j - b'_j) \right| \leq 2\alpha\sqrt{n} \quad \forall i. \tag{4}$$

We want to show that the bounds (4) imply that, with high probability, $d_H(b, b') = O(\alpha^2)$.

Towards this goal, let us fix some b' such that $d_H(b, b') \geq C\alpha^2$, where C is a parameter whose value will be determined later. We will show that the probability that b' satisfies (4) is very low. We have that, for any i, the random variable

$$Z = \sum_{j=1}^{n} \pi_i(y_j)(b_j - b'_j)$$

is a sum of the independent random variables $Z_j = \pi_i(y_j)(b_j - b'_j)$. Then

$$\mathbb{E}[Z^2] \geq \mathrm{Var}(Z) = \sum_{j=1}^{n} \mathrm{Var}(Z_j) = \sum_{j=1}^{n} \frac{(b_j - b'_j)^2}{4} \geq \frac{C\alpha^2 n}{4},$$

where we used the fact $d_H(b, b') = \frac{1}{n}\sum_{j=1}^{n}(b_j - b_j)^2$. By Lemma 4, we have

$$\Pr\left[|Z| \geq \frac{\sqrt{C}\alpha\sqrt{n}}{4}\right] \geq \frac{c'}{1 + \frac{4}{C\alpha^2 n}} \geq c'',$$

for some absolute constants $c', c'' > 0$, and any large enough n. Therefore, for $C \geq 64$,

$$\Pr\left[\left|\sum_{j=1}^{n} \pi_i(y_j)(b_j - b'_j)\right| \geq 2\alpha\sqrt{n}\right] \geq c''.$$

So far, we showed that a single query rules out b' as a candidate reconstruction with constant probability. Now, the probability that *no query* rules out b' is bounded as

$$\Pr\left[\forall i \quad \left|\sum_{j=1}^{n} \pi_i(y_j)(b_j - b'_j)\right| \leq 2\alpha\sqrt{n}\right] \leq (1 - c'')^k \leq e^{-c''k}.$$

Letting $k \geq (2n+1)\ln(2)/c''$, the right hand side is bounded by 2^{-2n-1}. Taking a union bound over all 2^{2n} possible pairs b, b' yields

$$\Pr\left[\exists b, b' \, \forall i \quad \left|\sum_{j=1}^{n} \pi_i(y_j)(b_j - b'_j)\right| \leq 2\alpha\sqrt{n}\right] \leq \frac{1}{2}.$$

Since a random choice of the predicates $\pi_1, \ldots, \pi_k$ works with constant probability, then, certainly, there exists such a choice. □

2.2 Tracing Attack

Tracing attacks are another form of attack in which an adversary tries to guess whether or not an individual is in the database, rather than attempting to reconstruct the entire database. That is, given the answers to a set of queries, the adversary wishes to determine a particular row x_i of the database X, perhaps using some auxiliary information. This type of tracing attack was studied by Homer et al. [4] in the context of genetic data. It is also very closely related to fingerprinting codes in cryptography, introduced by Boneh and Shaw [1], with a tight construction due to Tardos [5].

For the purposes of tracing attacks, we will consider *marginal queries*. We assume that the database X is over the universe $\mathcal{X} = \{0,1\}^d$. A marginal query q takes as parameters the database X, along with an index j, and is defined by

$$q(X, j) = \frac{1}{n}\sum_{i=1}^{n} x_{ij},$$

the fraction of rows in X whose jth column is equal to 1. We have the following, attack, due to Dwork et al. [3]. Here, a "nice" distribution is what the authors refer to as a strong distribution, as defined in [3].

Theorem 6. *Let $d \geq (Cn^2 \log \frac{1}{\delta})/\alpha$. Let $p = (p_1, \ldots, p_d)$, where each p_i is drawn independently from some "nice" distribution, and generate X so that $x_{ij} = 1$ with probability p_j, and $x_{ij} = 0$ with probability $1 - p_j$.*

If $\mathcal{M}$ is an algorithm such that, for all X, j, $|\mathcal{M}(X, j) - q(X, j)| \leq \alpha$, then there is an algorithm $\mathcal{A}$ such that

$$\Pr[\exists i \ \ \mathcal{A}(x_i, \mathcal{M}(X, 1), \ldots, \mathcal{M}(X, d), p) = \text{IN}] \geq 1 - \delta.$$

Moreover, if y is drawn from the same distribution as $x_1, \ldots, x_n$ (but independently from them), then

$$\Pr[\mathcal{A}(y, \mathcal{M}(X, 1), \ldots, \mathcal{M}(X, d), p) = \text{IN}] \leq \delta.$$

The attack is in fact very simple. For some $\tau = \Theta(\sqrt{d \log 1/\delta})$, the algorithm $\mathcal{A}(y, q_1, \ldots, q_d, p)$ says IN if and only if

$$\sum_{i=1}^{d} (y_i - p_i) q_i \geq \tau.$$

I.e., the attacker simply checks that y is better correlated with the output q of $\mathcal{M}(X)$ than the distribution p. The analysis is involved: check [3] for the details.

References

[1] D. Boneh and J. Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897–1905, 1998.

[2] 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. ACM, 2003.

[3] C. Dwork, A. D. Smith, T. Steinke, J. Ullman, and S. P. Vadhan. Robust traceability from trace amounts. In V. Guruswami, editor, *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 650–669. IEEE Computer Society, 2015. DOI: 10.1109/FOCS.2015.46. URL: http://dx.doi.org/10.1109/FOCS.2015.46.

[4] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8):e1000167, 2008.

[5] G. Tardos. Optimal probabilistic fingerprint codes. *J. ACM*, 55(2):10:1–10:24, 2008. DOI: 10.1145/1346330.1346335. URL: http://doi.acm.org/10.1145/1346330.1346335.