CSC2419:Algorithms and Complexity in Private Data Analysis Winter 2017

Lecture 1 — January 10, 2017

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 examples of naive mechanisms that do *not* provide any privacy guarantees, by describing attacks against these mechanisms. Even though we have not yet defined exactly what privacy is, it will be clear that these examples of mechanisms 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 vector $X \in \mathcal{U}^n$, where n is the number of rows in the database, and $\mathcal{U}$ is the universe of possible rows. We will typically have $\mathcal{U} = \{0,1\}^d$ 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 accessing 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 of X. 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 $q : \mathcal{U} \to \{0,1\}$,

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

is the fraction of rows in X that satisfy q.

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.

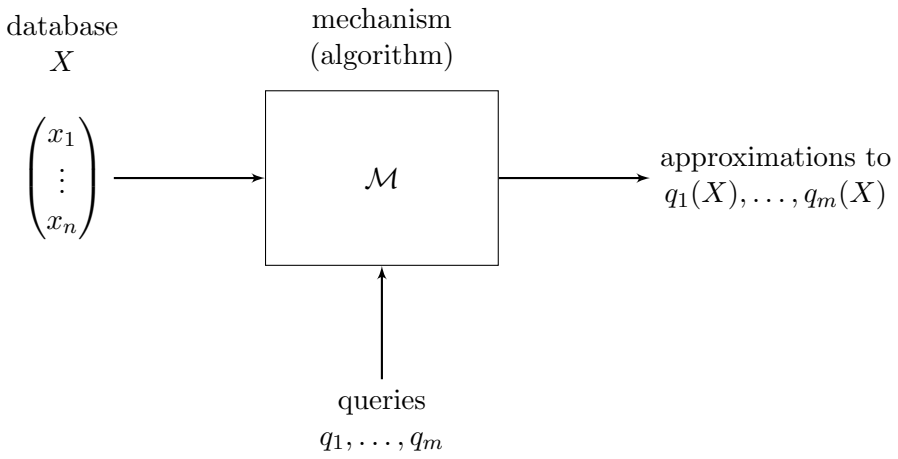

Figure 1: The setting in which we describe attacks on alleged private mechanisms $\mathcal{M}$.

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 [1]. In particular, we will consider *correlation queries* on a database $X \in \mathcal{U}^n$ over the universe $\mathcal{U} = \{0, 1\}$, that is, the universe contains a private bit x_i for each person.

Definition 1. *Let $q = (q_1, \ldots, q_n) \in \{0, 1\}^n$ be some bit vector. The correlation query associated with q is defined by*

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

To quantify the distance between two databases over $\{0, 1\}$ (or, more generally, bit vectors), we use the normalized hamming distance.

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

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 $\alpha > 0$, and suppose that, for every $q \in \{0, 1\}^n$, we have some $y_q \in \mathbb{R}$ such that $|y_q - q(X)| \leq \alpha$. Then we can compute a database $X' \in \{0, 1\}^n$ such that $d_H(X, X') \leq 4\alpha$.*

Proof. Given y_q for every $q \in \{0, 1\}^n$ as in the statement of the theorem, the reconstruction attack will be to simply output any database $X' = (x'_1, \ldots, x'_n) \in \{0, 1\}^n$ such that $|y_q - q(X')| \leq \alpha$ for all $q \in \{0, 1\}^n$. We know it must be possible to find such an X' since, in particular, the original database X itself satisfies this condition, as well as potentially many other candidate databases.

To show that X' is indeed an approximate reconstruction of X, we can bound the normalized hamming distance between the two databases. To do so, consider two queries $q_1 = X$, and $q_0 = 1 - X$, where q_0 is obtained by flipping all of the bits in X. We then have

$$\begin{aligned} d_H(X, X') &= \frac{1}{n}|\{i : x_i = 1, x'_i = 0\}| + \frac{1}{n}|\{i : x_i = 0, x'_i = 1\}| \\ &= q_1(X) - q_1(X') + q_0(X') - q_0(X) \\ &\leq (y_{q_1} + \alpha) - (y_{q_1} - \alpha) + (y_{q_0} + \alpha) - (y_{q_0} - \alpha) \\ &= 4\alpha. \end{aligned}$$

□

While the above theorem certainly demonstrates that a database 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 a database 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 d such that $\Pr[|Z| \geq \sigma/10] \geq d/\left(1 + O\left(1/\sigma^2\right)\right)$, where $\sigma^2 = \mathrm{Var}(Z)$.*

Proof. We have two cases to consider.

Case 1. $|E[Z]| \geq \frac{\sigma}{10}$. Since $E[|Z|] \geq |E[Z]|$, we have $\Pr[|Z| \geq \frac{\sigma}{10}] \geq \Pr[|Z| \geq \frac{1}{10}E[|Z|]$. Applying the Paley-Zygmund inequality gives us

$$\Pr\left[|Z| \geq \frac{\sigma}{10}\right] \geq \left(1 - \frac{1}{10}\right)^2 \frac{E[|Z|]^2}{E[Z^2]} \geq \frac{81}{100} \cdot \frac{E[Z]^2}{100E[Z]^2 + E[Z]^2} > 10^{-4}.$$

Case 2. $|E[Z]| < \frac{\sigma}{10}$. For each $i \in [n]$ let $Y_i = Z_i - E[Z_i]$, and let $Y = Y_1 + \cdots + Y_n$, so that, for all i, we have $E[Y_i] = 0$ and $Var(Y_i) = Var(Z_i) = E[Y_i^2]$. Moreover, for each i, we have $|Y_i| \leq |Z_i| + |E[Z_i]| \leq 2$.

Next, we can upper bound the fourth moment of Y as follows, by applying the multinomial theorem and independence.

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

Returning to Z, we have

$$\Pr\left[|Z| \geq \frac{\sigma}{10}\right] = \Pr\left[|Z| - |E[Z]| \geq \frac{\sigma}{10} - |E[Z]|\right] \geq \Pr\left[|Y| \geq \frac{\sigma}{5}\right] = \Pr\left[Y^2 \geq \frac{\sigma^2}{25}\right]$$
$$= \Pr\left[Y^2 \geq \frac{1}{25}E[Y^2]\right].$$

Applying the Paley-Zygmund inequality to Y^2 and substituting $(\star)$ yields

$$\begin{aligned}\Pr\left[|Z| \geq \frac{\sigma}{10}\right] &\geq \left(\frac{24}{25}\right)^2 \frac{E[Y^2]^2}{E[Y^4]} \\ &\geq \left(\frac{24}{25}\right)^2 \frac{E[Y^2]^2}{6E[Y^2]^2 + 4E[Y^2]} \\ &= \left(\frac{24}{25}\right)^2 \frac{\sigma^4}{6\sigma^4 + 4\sigma^2} \\ &= \left(\frac{24}{25}\right)^2 \frac{1}{1 + O\left(\frac{1}{\sigma^2}\right)}.\end{aligned}$$

□

Theorem 5. *There exist $m = O(n)$ queries $q_1, \ldots, q_m$ such that, given $y_1, \ldots, y_m$ such that $|y_i - q_i(X)| \leq \frac{\alpha}{\sqrt{n}}$ for all i, we can compute a database $X' \in \{0,1\}^n$ such that $d_H(X, X') = O(\alpha^2)$.*

Proof. Let m be some integer, whose value will be determined later. Choose correlation queries $q_1, \ldots, q_m$ independently and uniformly at random from $\{0,1\}^n$. The idea will be to show that, when m is sufficiently large, these queries satisfy the conditions of the theorem with high probability. Given these queries, along with their answers $y_1, \ldots, y_n$, the attack will be to output a database $X' \in \{0,1\}^n$ such that $|y_i - q_i(X')| \leq \frac{\alpha}{\sqrt{n}}$ for all i.

Let q_{ij} denote the jth bit of q_i. Now, fix any $X' = (x'_1, \ldots, x'_n)$ such that $d_H(X, X') \geq c\alpha^2$, where c is a parameter whose value will be determined later. Then

$$q_i(X) - q_i(X') = \frac{1}{n}\sum_{j=1}^{n} q_{ij}(x_j - x'_j)$$

is a sum of independent random variables. Now, since $d_H(X, X') \geq c\alpha^2$, it follows that $\sum_{i=1}^{n}(x_j - x'_j)^2 \geq c\alpha^2 n$. Therefore, we can compute the variance,

$$\begin{aligned}\mathrm{Var}\left(q_i(X) - q_i(X')\right) &= \frac{1}{4n^2}\sum_{j=1}^{n}(x_j - x'_j)^2 \\ &\geq \frac{c\alpha^2}{4n}.\end{aligned}$$

By Lemma 4, we have

$$\Pr\left[|q_i(X) - q_i(X')| \geq \frac{\alpha\sqrt{c}}{10\sqrt{n}}\right] \geq \frac{d}{1 + \frac{C}{\sigma^2}},$$

for some constant C, where $\sigma^2 = \mathrm{Var}(q_i(X) - q_i(X'))$ and d is the absolute constant from Lemma 4. Therefore, the probability that the condition in the expression does not hold for any i is

$$\Pr\left[\forall i \;\; |q_i(X) - q_i(X')| \leq \frac{\alpha\sqrt{c}}{10\sqrt{n}}\right] \leq \left(1 - \frac{d}{1 + \frac{C}{\sigma^2}}\right)^m.$$

Now, let $c_2 = \frac{d}{1+C/\sigma^2}$, and let

$$m \geq \frac{2n+1}{\log\left(\frac{1}{1-c_2}\right)} \quad \text{and} \quad c \geq 800.$$

Substituting these bounds for m and c, it follows that

$$\Pr\left[\forall i \;\; |q_i(X) - q_i(X')| \leq \frac{2\alpha}{\sqrt{n}}\right] \leq 2^{-2n-1}.$$

Taking a union bound over all 2^{2n} possible pairs X, X' yields

$$\Pr\left[\exists X, X' \forall i \;\; |q_i(X) - q_i(X')| \leq \frac{2\alpha}{\sqrt{n}}\right] \leq \frac{1}{2}.$$

Since, by the triangle inequality, $|y_i - q_i(X)| \leq \frac{\alpha}{\sqrt{n}}$ and $|y_i - q_i(X')| \leq \frac{\alpha}{\sqrt{n}}$ implies that $|q_i(X) - q_i(X')| \leq \frac{2\alpha}{\sqrt{n}}$, we have

$$\Pr\left[\exists X, X' \;\; d_H(X, X') \geq c\alpha^2\right] \leq \frac{1}{2}.$$

Thus, if the queries $q_1, \ldots, q_m$ are chosen correctly, which happens with high probability, then the database X' cannot be too far from X. □

2.2 Tracing Attack

Tracing attacks, first introduced by Tardos in [3], 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.

For the purposes of tracing attacks, we will consider *marginal queries*. We assume that the database X is over the universe $\mathcal{U} = \{0,1\}^d$. 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. [2]. Here, a "nice" distribution is what the authors refer to as a strong distribution, as defined in [2].

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

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

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*. ACM, 2003, pp. 202–210.

[2] C. Dwork, A. D. Smith, T. Steinke, J. Ullman, and S. P. Vadhan. Robust traceability from trace amounts. In *IEEE 56th annual symposium on foundations of computer science, FOCS 2015, berkeley, ca, usa, 17-20 october, 2015*. V. Guruswami, editor. IEEE Computer Society, 2015, pp. 650–669. DOI: 10.1109/FOCS.2015.46. URL: http://dx.doi.org/10.1109/FOCS.2015.46.

[3] 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.