In this lecture we will show some statistical limitations of mechanisms for counting queries: If such a mechanism is too accurate then a large portion of the database can be reconstructed, and so the mechanism cannot be almost always differentially private.

Then we show that if there are too few rows and each row is large then even some simple counting queries cannot always be answered both accurately and privately. For certain settings of parameters, almost always differential privacy is achievable, while differential privacy is not.

1 Counting queries require noise

Let us recall the privacy guarantee of the Blum-Ligett-Roth mechanism from Lecture 2: Given an n-row database over domain D, this mechanism provides an ε-differentially private *additive* approximation of αn to all queries within a given set Q with probability $1-\exp(-\Omega(\varepsilon\alpha n))$, provided $\alpha \geq K(\log|D|\log|Q|/n)^{1/3}$ for a sufficiently large constant K. If $|D|$, $|Q|$, and $1/\varepsilon$ are much smaller than n, our analysis of this mechanism guarantees an additive approximation on the order of $n^{2/3}$ to the true answers (which take value between 0 and n). Can we do better?

The following theorem shows that any mechanism for counting queries that is too accurate cannot be differentially private.

Theorem 1. *For every sufficiently large n and domain D of size $4n$ any mechanism that answers all sets of $O(n)$ counting queries on input $x \in D^n$ with additive error at most $0.02\sqrt{n}$ (with probability 1) is not $(1, 0.1)$-differentially private.*

The database x will consist of n rows of the type $(i,-1)$ or $(i,1)$, where the index i takes values between 1 and $2n$, all distinct. We will show that any mechanism with such good additive error is able to reconstruct at least $8/9$ of the rows of x. Intuitively, this ability to reconstruct such a large portion of x should be incompatible with differential privacy.

Instead of counting queries, it will be easier to work with difference queries. To each subset $S \subseteq [2n]$, we associate the difference query q'_S given by

$$q'_S(x) = \sum_{i\in S} x_i - \sum_{i\notin S} x_i.$$

where

$$x_i = \begin{cases} 1, & \text{if the row } (i,1) \text{ is present in } x \\ -1, & \text{if the row } (i,-1) \text{ is present in } x \\ 0, & \text{otherwise.} \end{cases}$$

The answer to a difference query can be determined from the answers to four counting queries:

$$q'_S(x) = |\{i \in S\colon x_i = 1\}| - |\{i \in S\colon x_i = -1\}| - |\{i \in \overline{S}\colon x_i = 1\}| + |\{i \in \overline{S}\colon x_i = -1\}|.$$

Moreover, if all four counting queries are answered within additive error 0.02, then the difference query can be answered with additive error 0.08.

Lemma 2. *Let $\mathcal{S}$ be a family of s uniformly and independently chosen random subsets of $[n]$. Then the probability that there exists a pair of databases x and $x^\star$ (of the desired form) that differ in at least $n/9$ rows such that $|q'_S(x^\star) - q'_S(x)| \le 0.16\sqrt{n}$ for all $S \in \mathcal{S}$ is at most $16^n \cdot (1 - 10^{-4})^s$.*

To prove this lemma we will need the following probabilistic claim.

Claim 3. *Let $a_1, \ldots, a_m$ be a sequence of numbers each one of which equals -2, -1, $+1$, or $+2$ and S be a uniformly random subset of $[m]$. Then the expression $X = \sum_{i \in S} a_i - \sum_{i \notin S} a_i$ has magnitude at least $0.5\sqrt{m}$ with probability at least 10^{-4}.*

Proof of Lemma 2. Fix a pair of databases x, x^* that differ in at least $n/10$ rows. Then

$$q'_S(x^\star) - q'_S(x) = \sum_{i \in S}(x^\star_i - x_i) - \sum_{i \notin S}(x^\star_i - x_i).$$

The difference $x^\star_i - x_i$ takes one of the values -2, -1, $+1$, or $+2$ when and only when $x^\star$ and x differ in the row indexed by i, so we can apply Claim 3 to conclude that $|q'_S(x^\star) - q'_S(x)|$ has magnitude at least $0.5\sqrt{n/9} > 0.16\sqrt{n}$ with probability at least $1/16$ over the choice of a random set S. Since $\mathcal{S}$ consists of an independent collection of s such sets, the probability that $|q'_S(x^\star) - q'_S(x)| \le 0.16\sqrt{n}$ for all sets $S \in \mathcal{S}$ is at most $(15/16)^s$. Since the number of pairs of databases $(x, x^\star)$ of the desired form is at most $(2^{2n})^2 = 16^n$, the lemma follows by taking a union bound over all such databases. □

We now prove Theorem 1. Choose a sufficiently large constant K so when we set $s = Kn$, then $16^n \cdot (1 - 10^{-4})^{-s} < 1$. Then there must exist a set $\mathcal{S}$ of size s such that for every pair of databases $x, x^\star$ that differs in at least $n/9$ rows, $|q'_S(x^\star) - q'_S(x)| > 0.16\sqrt{n}$ for at least one query q'_S, $S \in \mathcal{S}$. Let Q' be this set of difference queries (and Q the corresponding set of counting queries).

We now describe the reconstruction algorithm A: Given a sequence of answers (a_S) for $S \in \mathcal{S}$, A outputs some $x^\star \in D^n$ such that $|q'_S(x^\star) - a_S| \le 0.08\sqrt{n}$ for every $S \in \mathcal{S}$ if one exists and $\perp$ if such an $x^\star$ does not exist.

We claim that if $M(x)$ answers all queries in Q' with accuracy $0.08\sqrt{n}$ then the output $x^\star$ of $A(M(x))$ differs from x in at most $n/9$ rows. First, $A(M(x))$ does not output $\perp$ because the database $x^\star = x$ satisfies the desired criterion. Now let $x^\star$ be a possible output of $A(M(x))$. By the triangle inequality, for every query $q'_S \in Q'$,

$$|q'_S(x^\star) - q'_S(x)| \le |q'_S(x^\star) - a_S| + |q'_S(x^\star) - a_S| \le 0.08\sqrt{n} + 0.08\sqrt{n} = 0.16\sqrt{n}$$

so x and $x^\star$ differ in at most $n/9$ rows as desired.

Finally, we show that M cannot be $(1, 0.1)$-differentially private. Let $T_{j,x}$ be the set of all y such that $A(y)$ contains the j-th row of x. We just showed that for every database x,

$$\mathrm{E}_j\big[\Pr_M[M(x) \in T_{j,x}]\big] \ge 8/9$$

so averaging over a uniformly random database $X \in D^n$ of the desired form,

$$\mathrm{E}_j\big[\Pr_{M,X}[M(X) \in T_{j,X}]\big] \geq 8/9$$

By the definition of differential privacy, if $X^{(j)}$ is the database obtained by replacing the j-th row of X by a random row of the form $(i,-1)$ or $(i,+1)$ where index i is not present in X we get

$$\Pr_{M,X}[M(X) \in T_{j,X}] \leq e \Pr_{M,X}[M(X^{(j)}) \in T_{j,X}] + 0.1.$$

Conditioned on $X^{(j)}$, the j-th row of X is a random row whose index is not present in $X^{(j)}$. The output $A(M(X^{(j)}))$ contains n rows, $8n/9$ of which must be present in $X^{(j)}$. So the probability that $A(M(X^{(j)}))$ contains a random row whose index is not present in $X^{(j)}$ can be at most $1/9$. We can conclude that for every j,

$$\Pr_{M,X}[M(X) \in T_{j,X}] \leq e \cdot 1/9 + 0.1.$$

averaging over j, it follows that $8/9 \leq e/9 + 0.1$, a contradiction.

Proof of Claim 3. Let $X_i = a_i$ if $i \in S$ and $X_i = -a_i$ if $i \notin S$. Then $X_1, \ldots, X_m$ are independent, *symmetric* unbiased random variables, taking values $\{-2,-1,+1,+2\}$, and $X = X_1 + \cdots + X_m$. One elegant way to prove a *lower bound* on the magnitude of X is via the Paley-Zygmund inequality, which says that for every nonnegative random variable Z and $0 \leq \theta \leq 1$,

$$\Pr\big[Z \geq \theta \,\mathrm{E}[Z]\big] \geq (1-\theta)^2 \frac{\mathrm{E}[Z]^2}{\mathrm{E}[Z^2]}.$$

We plug in $Z = X^2$ and $\theta = 1/4$. Using independence and symmetry, one can calculate

$$\mathrm{E}[Z] = \mathrm{E}[X^2] = \sum_{i=1}^{n} \mathrm{E}[X_i^2] \geq m$$

and

$$\mathrm{E}[Z^2] = \mathrm{E}[X^4] = \sum_{i=1}^{m} \mathrm{E}[X_i^4] + \sum_{i \neq j} \mathrm{E}[X_i^2]\,\mathrm{E}[X_j^2] \leq 16m + 48m(m-1) \leq 48m^2$$

from where

$$\Pr\big[|X| \geq \tfrac{1}{2}\sqrt{m}\big] = \Pr\big[X^2 \geq \tfrac{1}{4}\,\mathrm{E}[X^2]\big] \geq \big(\tfrac{3}{4}\big)^2 \cdot \frac{\mathrm{E}[X^2]^2}{\mathrm{E}[X^4]} = \frac{9}{16} \cdot \frac{m^4}{(48m^2)^2} > 10^{-4}. \qquad \square$$

Efficiency One possible objection to the algorithm A is that it may be computationally inefficient. There are two reasons for this: First, we did not describe how to find the relevant difference queries but merely showed their existence; second, finding a database $x^\star$ such that $q'_S(x^\star)$ is close to a_S for all S in $\mathcal{S}$ might be computationally expensive. This algorithm can be made efficient if we slightly increase the number of queries from $O(n)$ to $O(n(\log n)^2)$. You will work this out in Homework 2.

2 Databases with few large rows

We now give an example of a database over domain $D = \{0,1\}^k$ for which it is impossible to answer k counting queries in an ε-differentially private way, even with linear additive error, unless $n = \Omega(k/\varepsilon)$.

The k queries $q_1, \ldots, q_k$ are the 1-way marginals of x: The i-th 1-way marginal q_i counts the number of rows r in x such that the i-th bit r_i of r equals 1.

Theorem 4. *If there exists an ε-private mechanism that answers all 1-way marginals $q_i(x)$ for every $x \in D^n$ within additive error less than $n/2$ with probability at least $1/2$ then $n = \Omega(k/\varepsilon)$.*

Contrast this lower bound with the performance of the Blum-Ligett-Roth mechanism, which gives a privacy guarantee even for a much larger set of queries but only if the domain is smaller.

Also, this theorem shows that the product mechanism is essentially optimal for answering marginal queries in an ε-differentially private way. Recall that this mechanism can handle more queries if we only require almost always differential privacy.

Proof. Let the database x_r consist of n identical rows $r \in \{0,1\}^k$. In this case, the true answer to $q_i(x_r)$ is n if the i-th bit of r equals 1 and 0 if not. Let $B_r \subseteq \mathbb{R}^k$ be the "open subcube"

$$B_r = (nr_1 - n/2, nr_1 + n/2) \times \cdots \times (nr_k - n/2, nr_k + n/2).$$

Since the sets B_r, $r \in \{0,1\}^k$, are pairwise disjoint, for every $s \in \{0,1\}^k$,

$$\sum_{r \in \{0,1\}^k} \Pr[M(x_s) \in B_r] = \Pr[M(x_s) \in B_r \text{ for some } r \in \{0,1\}^k] \leq 1$$

and so there must exist some B_r such that

$$\Pr[M(x_s) \in B_r] \leq 2^{-k}.$$

If $M(x_r)$ provides approximations to $(q_1(x_r), \ldots, q_k(x_r))$ within the desired additive accuracy of $n/2$ then $M(x_r)$ must belong to the set B_r with probability at least $1/2$:

$$\Pr[M(x_r) \in B_r] \geq 1/2.$$

The database x_r can be obtained from x_s after modifying all n of its rows, so by differential privacy

$$\Pr[M(x_r) \in B_r] \leq e^{\varepsilon n} \Pr[M(x_s) \in B_r].$$

Putting all three inequalities together, we conclude that $1/2 \leq e^{\varepsilon n} \cdot 2^{-k}$, so $n \geq ((\ln 2)k - 1)/\varepsilon$. □

References

These notes are based on Chapter 8 of the survey *The Algorithmic Foundations of Differential Privacy* by Cynthia Dwork and Aaron Roth.

For the presentation of Theorem 1 I was guided by the paper *Revealing Information while Preserving Privacy* of Irit Dinur and Kobbi Nissim and some help from Aaron Roth, who explained how the ability to reconstruct a large portion of a database implies lack of differential privacy.

You can find a proof of the Paley-Zygmund inequality in the Wikipedia article on it.