CSCI 5520: Foundations of Data Privacy **Lecture 2**
The Chinese University of Hong Kong, Spring 2015 13 January 2015

In the last lecture we saw the Laplace mechanism for privately answering counting queries. We also showed that multiple queries can be answered privately using a product mechanism. However, the privacy parameter deteriorates linearly with the number of queries.

We will now consider some additional types of queries that can be handled using variants of the Laplace mechanism.

We will then show a different mechanism that, for certain settings of parameters, can handle an arbitrarily large number of counting queries without suffering the privacy deterioration of the product mechanism.

1 Beyond counting queries

Histograms

Suppose that we have a database containing CSCI 5520 students' names, some additional information, and their grades. We want to know how many students got an A, how many got a B, and so on. This is an example of a histogram query: Its answer consists of a vector of five numbers $(n_A, n_B, n_C, n_D, n_F)$ giving the number of students with each grade.

In general, given a function h from the domain D to a finite set of *buckets* B, the answer to the *histogram query* $q_h(x)$ consists of the sequence of numbers $(n_b)_{b \in B}$, where n_b is the number of rows i such that $h(x_i) = b$.

Notice that each n_b is a counting query. If we answer each n_b independently by the Laplace mechanism with privacy parameter ε, our analysis of the product mechanism from last time allows us to conclude that this mechanism M is $|B|\varepsilon$-private for the histogram query q_h. In fact we can prove a stronger claim.

Claim 1. *M is 2ε-differentially private.*

So the privacy of the mechanism does not depend at all on the number of buckets in the histogram.

In order to prove Claim 1, it helps to think of a histogram query q_h as a *vector-valued* query taking values in the vector space $\mathbb{R}^B$. Recall that the ℓ_1 distance between two vectors in $v, v' \in \mathbb{R}^B$ is the number $\sum_{b \in B} |v_b - v'_b|$. We observe that if x and x' are databases that differ in at most one row, then the ℓ_1 distance between $q_h(x)$ and $q_h(x')$ is at most two: Changing one row in a database can at most shift one item from one bucket into another.

A query $q \colon D^n \to \mathbb{R}^d$ is *c-Lipschitz* if for every pair of databases x, x' that differ in at most one row, the ℓ_1 distance between $q(x)$ and $q(x')$ is at most c. Claim 1 now follows from the following general theorem:

Theorem 2. *Let $q\colon D^n \to \mathbb{R}^d$ be a c-Lipschitz integer valued query and M be the product of ε-differentially private Laplace mechanisms over the entries of q. Then M is $c\varepsilon$-differentially private.*

Proof. Let x, x' be databases that differ in at most one row and let $c_i = q_i(x) - q_i(x')$. Then

$$\begin{aligned}
\Pr[M(x) = (y_1, \dots, y_d)] &= \Pr[q_1(x) + N_1 = y_1] \cdots \Pr[q_d(x) + N_d = y_d] \\
&= \Pr[N_1 = y_1 - q_1(x)] \cdots \Pr[N_d = y_d - q_d(x)] \\
&\le (e^{\varepsilon|c_1|} \Pr[N_1 = y_1 - q_1'(x)]) \cdots (e^{\varepsilon|c_d|} \Pr[N_1 = y_1 - q_1'(x)]) \\
&= e^{\varepsilon(|c_1| + \cdots + |c_d|)} \Pr[N_1 = y_1 - q_1(x')] \cdots \Pr[N_d = y_d - q_d(x')] \\
&= e^{\varepsilon(|c_1| + \cdots + |c_d|)} \Pr[M(x') = (y_1, \dots, y_d)].
\end{aligned}$$

Because q is c-Lipschitz, $|c_1| + \cdots + |c_d| \le c$, and so M is $c\varepsilon$-differentially private. □

Most popular item

Suppose we have a database that describes whether a child likes a particular type of fruit:

likes	apple	orange	banana
Alice	yes	yes	no
Bob	no	yes	yes
Charlie	no	yes	no
Dave	no	yes	yes
Erica	no	no	yes

The query we want to answer is "What is the most popular fruit?" The correct answer is "orange". Giving out this answer directly is clearly not differentially private: If we change Charlie's row of preferences, the answer to the query may change to "banana". Here is a private mechanism for this task:

The Noisy Max mechanism. Given a database $x \in D^n$,

1. Calculate the number $n_i(x)$ of yes-queries for every column i.
2. Sample independent $\mathrm{Lap}(1/\varepsilon)$ random variables N_i, one for every column.
3. Output the column label i for which the value $n_i(x) + N_i$ is the largest. (In case of ties, choose the first label among those that maximize the value $n_i(x) + N_i$.)

In our example, $n_{\text{apple}} = 1$, $n_{\text{orange}} = 4$, $n_{\text{banana}} = 3$. If the sampled noise values are, say, $N_{\text{apple}} = 1$, $N_{\text{orange}} = -1$, and $N_{\text{banana}} = -1$, the mechanism would output "orange". If, in contrast, $N_{\text{apple}} = 0$, $N_{\text{orange}} = -1$, and $N_{\text{banana}} = 1$, then the mechanism would output "banana".

Theorem 3. *The Noisy Max mechanism is 2ε-differentially private.*

Before we prove this theorem, here is an easy but useful lemma about the Laplace distribution.

Lemma 4. *If N is a* $\mathrm{Lap}(1/\varepsilon)$ *random variable, then for every pair of integers n and k,*

$$\Pr[N > n] \le e^{k\varepsilon} \Pr[N > n + k] \quad \textit{and} \quad \Pr[N < n + k] \le e^{k\varepsilon} \Pr[N < n].$$

Proof Sketch for Theorem 3. For simplicity of notation, we will ignore the possibility of ties among the values $n_i(x) + N_i$. To turn this proof sketch into a real proof, you need to account for that possibility. It is not too difficult to do so.

Let x and x' be two neighboring databases. Then $M(x)$ outputs i if and only the quantity $n_j(x)+N_j$ is maximized at $j = i$, namely

$$\begin{aligned}
\Pr[M(x) = i] &= \Pr[n_i(x) + N_i > n_j(x) + N_j \text{ for all } j \ne i] \\
&= \Pr[N_i > n_j(x) - n_i(x) + N_j \text{ for all } j \ne i] \\
&= \Pr[N_i > \max_{j\ne i}\{n_j(x) - n_i(x) + N_j\}] \\
&= \mathrm{E}\big[\Pr_{N_i}[N_i > \max_{j\ne i}\{n_j(x) - n_i(x) + N_j\} \mid N_{-i}]\big].
\end{aligned}$$

where N_{-i} denotes all random variables N_j except for N_i. By an analogous calculation

$$\Pr[M(x') = i] = \mathrm{E}\big[\Pr_{N_i}[N_i > \max_{j\ne i}\{n_j(x') - n_i(x') + N_j\} \mid N_{-i}]\big].$$

To complete the proof, we will show that for every fixing of N_{-i}, the ratio of the two probabilities is at most $e^{2\varepsilon}$. When we replace x by x', every term $n_j(x) - n_i(x) + N_j$ can change by at most two, so the maximum of all these terms can also change by at most two. It follows that

$$\max_{j\ne i}\{n_j(x) - n_i(x) + N_j\} \ge \max_{j\ne i}\{n_j(x') - n_i(x') + N_j\} - 2$$

and so for every fixing of N_{-i},

$$\begin{aligned}
&\Pr_{N_i}[N_i > \max_{j\ne i}\{n_j(x) - n_i(x) + N_j\} \mid N_{-i}] \\
&\qquad\qquad \le \Pr_{N_i}[N_i > \max_{j\ne i}\{n_j(x') - n_i(x') + N_j\} - 2 \mid N_{-i}].
\end{aligned}$$

By Lemma 4, the last expression is at most $e^{2\varepsilon} \Pr_{N_i}[N_i > \max_{j\ne i}\{n_j(x') - n_i(x') + N_j\} \mid N_{-i}]$, as desired. □

2 The exponential mechanism

The exponential mechanism is a general mechanism for answering queries whose answers can take values in an arbitrary range (not necessarily numerical). Let us illustrate it with an example.

Suppose we have an unlimited supply of pumpkins that we are trying to sell to Alice, Bob, and Charlie. We allow each one of them to bid \$1 or \$2 for a pumpkin, and want to set the price of a pumpkin so as to maximize our revenue.

For example, say we have the following database of bids:

name	bid
Alice	\$1
Bob	\$1
Charlie	\$2

If set the pumpkin price at \$1, this is within everyone's budget so collect \$3 in revenue; if we set the price at \$2, only Charlie can afford to buy and so our revenue is \$2 only. We can describe this information as a *utility function*:

price	utility
\$1	\$3
\$2	\$2

In general, the utility function depends both on the bids and on the price. If our objective is to maximise revenue, we would set the pumpkin price at \$1.

However, publishing the price might violate the privacy of the bidders. For example, suppose we know Alice is poor (she can afford to spend only \$1 on a pumpkin) and Charlie is rich (he can easily bid \$2), but we know nothing about Bob. Then the revenue maximising pumpkin price reveals Bob's budget.

The exponential mechanism is a general mechanism for setting prices in a manner that guarantees differential privacy of the bidders. In fact, this mechanism works in a much more general setting, which we now describe.

In general, we are interested in designing a mechanism for answering a query q that takes as input a database $x \in D^n$ and produces an output an element in some range of values R. To each database x and range value r, there is an associated numerical *utility score* $u(x, r)$.

In our example, q is the query "What is the price of a pumpkin?" and R is the range of possible prices $\{\$1, \$2\}$. The utility is the profit in dollars; for the above database of bids x, $u(x, \$1) = 3$ and $u(x, \$2) = 2$.

We saw that choosing the outcome with the highest utility score is detrimental for privacy. The exponential mechanism will sample the outcome from a probability distribution. The probability of an an outcome will depend on its utility: The higher the utility, the more likely the outcome.

Let Δu be the maximum of $|u(x, r) - u(x', r)|$ over all pairs of neighbouring databases $x, x' \in D^n$ and all possible values $r \in R$.

The exponential mechanism. Given a database x and utility scores $u(x, r)$ for every $r \in R$, output r with probability proportional to $\exp(\varepsilon u(x, r)/2\Delta u)$.

We will shortly prove that

Theorem 5. *The exponential mechanism is ε-differentially private.*

Before we do that, let's see what the mechanism does for our pumpkins. First, we need to calculate Δu. Since utility represents profit, this amounts to figuring out the following: For any fixed collection of bids (database) and pumpkin price (element in the range), by how much can our profit be affected by a single person changing their bid? It is not difficult to see that the answer is \$2, so $\Delta u = 2$. Thus the exponential mechanism will output a pumpkin price of \$1 with probability proportional to $\exp(3\varepsilon/4)$ and \$2 with probability proportional to $\exp(\varepsilon/2)$. For $\varepsilon = 0.2$, a bid of \$1 happens with probability about 51% and a bid of \$2 with probability about 49%.

This is almost like flipping a random coin to determine the price. To get a better feel about the effect of the exponential mechanism, let's take a larger example with a bigger gap in utilities. Suppose you have 100 bidders for pumpkins, out of which 90 bid \$1 and 10 bid \$2, so the utility is 100 for a \$1 price and 20 for a \$2 price. If we want $\varepsilon = 0.2$, the exponential mechanism sets the pumpkin price to \$1 with probability 98.2% and to \$2 with probability 1.8%.

Proof of Theorem 5. Let x and x' be neighbouring databases and $t \in R$ a possible outcome. Then

$$\begin{aligned}
\Pr[M(x) = t] &= \frac{\exp(\varepsilon u(x,t)/2\Delta u)}{\sum_{r\in R}\exp(\varepsilon u(x,r)/2\Delta u)} \\
&\leq \frac{\exp(\varepsilon(u(x',t) + \Delta u)/2\Delta u)}{\sum_{r\in R}\exp(\varepsilon(u(x',r) - \Delta u)/2\Delta u)} \\
&= \frac{\exp(\varepsilon/2)\cdot\exp(\varepsilon u(x',t)/2\Delta u)}{\sum_{r\in R}\exp(-\varepsilon/2)\cdot\exp(\varepsilon u(x',r)/2\Delta u)} \\
&= \exp(\varepsilon)\cdot\frac{\exp(\varepsilon u(x',t)/2\Delta u)}{\sum_{r\in R}\exp(\varepsilon u(x',r)/2\Delta u)} \\
&= \exp(\varepsilon)\cdot\Pr[M(x') = t].
\end{aligned}$$

□

While the exponential mechanism will not, in general, produce an output with maximum utility, the following theorem gives a bound on the probability that the utility the mechanism outputs is substantially smaller than the optimal one.

Theorem 6. *For a given x, let $u^*(x)$ be the maximum of $u(x,r)$ over all $r \in R$. Then for every $t > 0$, the probability that the exponential mechanism produces an output of utility smaller than $u^*(x) - t$ is less than $|R|\exp(-\varepsilon t/2\Delta u)$.*

Proof. Let E denote the exponential mechanism. Fix a database x. Then

$$\begin{aligned}
\Pr[E(x) < u^*(x) - t] &= \sum_{r\in R\colon\, u(x,r)<u^*(x)-t} \Pr[E(x) = r] \\
&= \frac{\sum_{r\in R\colon\, u(x,r)<s}\exp(\varepsilon u(x,r)/2\Delta u)}{\sum_r \exp(\varepsilon u(x,r)/2\Delta u)}.
\end{aligned}$$

We can upper bound the numerator strictly by

$$\sum_{r\in R\colon\, u(x,r)<u^*(x)-t} \exp(\varepsilon u(x,r)/2\Delta u) \leq \sum_{r\in R\colon\, u(x,r)<s} \exp(\varepsilon(u^*(x)-t)/2\Delta u) \leq |R|\exp(\varepsilon(u^*(x)-t)/2\Delta u)$$

as there are at most $|R|$ entries in the summation, and lower bound the denominator by

$$\sum_r \exp(\varepsilon u(x,r)/2\Delta u) \geq \exp(\varepsilon u^*(x)/2\Delta u)$$

by choosing the term in the summation that maximizes $u(x,r)$. Combining the two, we get

$$\Pr[E(x) < u^*(x) - t] < |R| \exp(-\varepsilon t/2\Delta u). \qquad \square$$

3 A mechanism for many counting queries

Consider what happens when you apply the product mechanism to d instances of the same counting query q. Every time the answer is $q(x)$ plus some noise, and the different instances of the noise are statistically independent. By taking enough of them and averaging them one can obtain a very good estimate of the value $q(x)$, violating privacy. To make this argument a bit more quantitative, if the noise is of standard deviation σ, then averaging the value of $O(\sigma^2)$ queries reduces the standard deviation to a small constant. Since counting queries can only take integral values, after this many queries the exact value $q(x)$ is determined with high probability. (In the Laplace mechanism, σ is the inverse of the privacy parameter ε, so $O(1/\varepsilon^2)$ queries suffice for this experiment to succeed.)

To avoid this type of privacy violation, it would be sensible to answer the same query identically, and not independently, every time. More generally, to ensure privacy, related queries better result in statistically correlated answers.

We consider the following general scenario. We want to design a differentially private mechanism that provides accurate answers to a whole class of counting queries Q, which can be potentially very large. Here is one possible design for such a mechanism: Given a (private) database $x \in D^n$, the mechanism will use randomness to create a *synthetic* (fake) database $y \in D^d$ and release all the contents of y to the public. On the one hand, y will have a similar utility as x; when answering queries from Q, it will not matter much whether x or y is used as the reference database. On the other hand, releasing y will cause no privacy harm to the participants in x.

Blum, Ligett, and Roth came up with an ingenious way of choosing the database y. To explain their solution, it helps to replace counting queries by the equivalent notion of *averaging queries*: Given a counting query q over a database with n rows, the corresponding averaging query $\overline{q}$ outputs the answer $\overline{q}(x) = q(x)/n$. So the answer of an averaging query is always a number between 0 and 1. (We also use $\overline{Q}$ for the class of averaging queries corresponding to Q.) We will also think of the number of rows d of the synthetic database as a number much smaller than the number of rows n of the real database.

Let us start with the following question: Given $x \in D^n$, among all databases $y \in D^d$ rows, which is the one that gives the most accurate approximation to all the queries in $\overline{Q}$? One formalization of this question goes by defining a *utility score* $u(x,y)$ that is higher the more similar x and y look to the queries in $\overline{Q}$:

$$u(x,y) = -\max\nolimits_{\overline{q} \in \overline{Q}} |\overline{q}(x) - \overline{q}(y)|.$$

If x and y give exactly the same answer to all the queries in $\overline{Q}$, then $u(x,y) = 0$; otherwise $u(x,y)$ takes some negative value. The closer this value is to zero, the more similar x and y look to $\overline{Q}$.

We choose the synthetic database y by applying the exponential mechanism on input x and utilities u. By Theorem 5, the mechanism is ε-differentially private. It remains to show that it is useful – that is, the answers to queries in $\overline{Q}$ are typically similar; in other words, the mechanism typically outputs a y that achieves high utility.

We will show high utility using Theorem 6. To apply this theorem, we need to first estimate the value

$$u^*(x) = \max_y u(x, y).$$

Claim 7. *For every $x \in D^n$, there exists a $y \in D^d$ such that $u(x, y) \geq -\alpha$ whenever $2|Q| < e^{2d\alpha^2}$.*

We will prove this claim later; let us first see what kind of utility it gives us. Combining Theorem 6 and Claim 7, we get that the mechanism produces an output of utility less than -2α with probability at most

$$|R| \exp(-\varepsilon\alpha/2\Delta u) \leq |D|^d \exp(-\varepsilon\alpha n/4)$$

where we use the fact that there are $|D|^d$ entries in the range of y and the fact that $\Delta u \leq 2/n$: Changing a row of x can affect the value $\overline{q}(x)$, and therefore also $u(x, y)$, by at most $2/n$. If we choose the smallest possible value of d that satisfies Claim 7, we get an upper bound of

$$|D|^{O(\log|Q|/\alpha^2)} \cdot \exp(-\varepsilon\alpha n/4) = \exp(-\Omega(\varepsilon\alpha n))$$

on this probability, as long as $\alpha \geq K(\log|Q| \log|D|/n)^{1/3}$ for a sufficiently large constant K. So if n is sufficiently large, we can get simultaneous accuracy for very large sets of counting queries with very high probability over the choice of y using this mechanism.

Proof of Claim 7 We prove this claim using the probabilistic method. We will need the following form of the Chernoff bound:

Theorem 8 (Chernoff bound)**.** *Let $X_1, \ldots, X_m$ be independent random variables taking values in the range $[0, 1]$ and $\overline{X} = (X_1 + \cdots + X_m)/m$. Then for every $\alpha > 0$*

$$\Pr\big[\overline{X} > \mathrm{E}[\overline{X}] + \alpha\big] \leq e^{-2m\alpha^2} \quad \textit{and} \quad \Pr\big[\overline{X} < \mathrm{E}[\overline{X}] - \alpha\big] \leq e^{-2m\alpha^2}.$$

Proof of Claim 7. Fix x. We will show that if Y consists of a random sample of the rows of x, where each row of Y is chosen uniformly and independently among all rows of x (with repetition), then $u(x, Y) \geq -\alpha$ with probability strictly greater than zero over the choice of Y. Therefore such a Y must exist.

Fix an averaging query $\overline{q}$. Let Y_i be an indicator random variable for the event that the i-th row of y satisfies the predicate for the query $\overline{q}$. Then $Y_1, \ldots, Y_d$ are independent $\{0, 1\}$ random variables and by the Chernoff bound,

$$\Pr[|\overline{Y} - \mathrm{E}[\overline{Y}]| > \alpha] \leq 2e^{-2d/\alpha^2}.$$

where $\overline{Y} = (Y_1 + \cdots + Y_d)/d = \overline{q}(Y)$. Since each Y_i is a uniformly random row of x, it satisfies the predicate of $\overline{q}$ with probability $\overline{q}(x)$, and so $\mathrm{E}[\overline{Y}] = \overline{q}(x)$. So the last inequality says

$$\Pr[|\overline{q}(Y) - \overline{q}(x)| > \alpha] \leq 2e^{-2d/\alpha^2}.$$

Taking a union bound over all $\overline{q} \in \overline{Q}$, we get that

$$\Pr[|\overline{q}(Y) - \overline{q}(x)| > \alpha \text{ for some } \overline{q} \in \overline{Q}] \leq |\overline{Q}| \cdot 2e^{-2d\alpha^2}.$$

By our condition on α, the quantity $2|\overline{Q}|e^{-2d\alpha^2}$ is strictly less than one. In this case, the probability of the complement event $|\overline{q}(Y) - \overline{q}(x)| \leq \alpha$ for all $\overline{q}$ in $\overline{Q}$ is strictly greater than zero, so such a Y must exist. □

References

These notes are based on Chapters 3 and 4 of the survey *The Algorithmic Foundations of Differential Privacy* by Cynthia Dwork and Aaron Roth. For more on the probabilistic method, see the book *The Probabilistic Method* by Alon and Spencer.