CIS 800/002 The Algorithmic Foundations of Data Privacy October 20, 2011

Lecture 11

Lecturer: Aaron Roth *Scribe: Aaron Roth*

The Sparse Vector Technique

We're going to take a short break from the problem of private query release to develop another fundamental technique in differential privacy. Don't worry – we'll soon use this to improve our query release algorithms.

Suppose that a data analyst wants to know the answers to k adaptively chosen, low-sensitivity queries on a private database. At the moment, the only way we know how to handle adaptively chosen queries is by using the Laplace mechanism, and paying a cost in our privacy parameter proportional to k (or $\sqrt{k}$ for (ϵ, δ)-privacy). But what if the data analyst has reason to believe that only a very small number of his queries (say c of them) will take value above a certain threshold T? Moreover, what if he only cares about the values of those queries that actually evaluate above the threshold? If he knew which queries those were, he could ask only the c relevant queries, and pay a privacy cost proportional only to c. The problem is he doesn't...

In this lecture, we'll show an algorithm for answering any sequence of k adaptively chosen low sensitivity queries, while paying a privacy cost proportional only to those queries that are above a given threshold T.

Algorithm 1 Input is a private database D, an adaptively chosen stream of sensitivity $1/n$ queries $Q_1, \ldots$, a threshold T, and a cutoff point c. Output is a stream if answers $a_1, \ldots$

```
Sparse($D, \{Q_i\}, T, c$)
  Let $\hat{T} = T + \text{Lap}\left(\frac{2}{\epsilon n}\right)$
  Let $\sigma = \frac{2c}{\epsilon \cdot n}$
  Let count = 0
  for Each query $i$ do
    Let $\nu_i = \text{Lap}(\sigma)$
    if $Q_i(D) + \nu_i \geq \hat{T}$ then
      Output $a_i = Q_i(D) + \nu_i$.
      Let count = count +1.
    else
      Output $a_i = \perp$.
    end if
    if count $\geq c$ then
      Abort.
    end if
  end for
```

Definition 1 (Accuracy) *We will say that Sparse is (α, β)-accurate for a sequence of k queries $Q_1, \ldots, Q_k$, if except with probability at most β, the algorithm does not abort before Q_k, and for all $a_i \in \mathbf{R}$:*

$$|a_i - Q_i(D)| \leq \alpha$$

and if for all $a_i = \perp$:

$$Q_i(D) \leq T + \alpha$$

.

Theorem 2 *For any sequence of k queries $Q_1, \ldots, Q_k$ such that $L(T) \equiv |\{i : Q_i(D) \geq T - \alpha\}| \leq c$, Sparse$(T, C)$ is (α, β) accurate for:*

$$\alpha = 2\sigma \left(\log k + \log \frac{2}{\beta} \right) = \frac{4c(\log k + \log(2/\beta))}{\epsilon n}.$$

Proof Observe that the theorem will be proved if we can show that except with probability at most β:

$$\max_{i \in [k]} |\nu_i| + |T - \hat{T}| \leq \alpha$$

If this is the case, then for any $a_i \in \mathbf{R}$, we have: $|a_i - Q_i(D)| = |\nu_i| \leq \alpha$, for any $a_i = \bot$ we have:

$$Q_i(D) + \nu_i \leq \hat{T} \leq T + |T - \hat{T}|$$

i.e. $Q_i(D) \leq T + |T - \hat{T}| + |\nu_i| \leq T + \alpha$. We will also have that for any $i \notin L$: $Q_i(D) < T - \alpha < T - |\nu_i| - |T - \hat{T}|$, and so: $Q_i(D) + \nu_i \leq \hat{T}$, meaning $a_i = \bot$. Therefore the algorithm does not halt before k queries are answered.

We now complete the proof.

Recall that if $Y \sim \text{Lap}(b)$, then: $\Pr[|Y| \geq t \cdot b] = \exp(-t)$. Therefore we have:

$$\Pr[|T - \hat{T}| \geq \frac{\alpha}{2}] = \exp\left(-\frac{\epsilon n \alpha}{4}\right)$$

Setting this quantity to be at most $\beta/2$, we find that we require $\alpha \geq \frac{4 \log(2/\beta)}{\epsilon n}$

Similarly, by a union bound, we have:

$$\Pr[\max_{i \in [k]} |\nu_i| \geq \alpha/2] \leq k \cdot \exp\left(-\frac{\epsilon n \alpha}{4c}\right)$$

Setting this quantity to be at most $\beta/2$, we find that we require $\alpha \geq \frac{4c(\log(2/\beta) + \log k)}{\epsilon n}$ These two claims combine to prove the theorem.

■

Theorem 3 *The sparse vector algorithm is ϵ-differentially private.*

Proof The output of the algorithm is $a \in (\mathbf{R} \cup \{\bot\})^t$. Write $a^{<i}$ to denote the prefix $a_1, \ldots, a_{i-1}$. We want to show for all neighboring pairs of databases D, D' and for all output vectors $\hat{a}$:

$$\log \left(\frac{\Pr_D[a = \hat{a}]}{\Pr_{D'}[a = \hat{a}]} \right) = \sum_{i=1}^{t} \log \left(\frac{\Pr_D[a_i = \hat{a}_i | \hat{a}^{<i}]}{\Pr_{D'}[a_i = \hat{a}_i | \hat{a}^{<i}]} \right) \leq \epsilon$$

Recall that our outputs are either numeric or $\bot$. Let $N = \{i : \hat{a}_i \in \mathbf{R}\}$ denote the set of indices of the numeric answers, and let $N^C = \{i : \hat{a}_i = \bot\}$ denote the indices of the non-numeric answers. Of course:

$$\log \left(\frac{\Pr_D[a = \hat{a}]}{\Pr_{D'}[a = \hat{a}]} \right) = \sum_{i \in N} \log \left(\frac{\Pr_D[a_i = \hat{a}_i | \hat{a}^{<i}]}{\Pr_{D'}[a_i = \hat{a}_i | \hat{a}^{<i}]} \right) + \sum_{i \in N^C} \log \left(\frac{\Pr_D[a_i = \bot | \hat{a}^{<i}]}{\Pr_{D'}[a_i = \bot | \hat{a}^{<i}]} \right)$$

We bound the two sums separately. Consider the first sum. By design, $|N| \leq c$, since the algorithm Aborts after c numeric queries. Therefore, by the properties of the Laplace mechanism:

$$\sum_{i \in N} \log \left(\frac{\Pr_D[a_i = \hat{a}_i | \hat{a}^{<i}]}{\Pr_{D'}[a_i = \hat{a}_i | \hat{a}^{<i}]} \right) = \sum_{i \in N} \log \left(\frac{\Pr[\nu_i = \hat{a}_i - Q_i(D)]}{\Pr[\nu_i = \hat{a}_i - Q_i(D')]} \right) \leq \sum_{i \in N} \frac{\epsilon}{2c} \leq \frac{\epsilon}{2}$$

Now consider the second sum, and simultaneously bound all terms using the independent randomness we used to select $\hat{T}$. Define $\mathcal{A}_Z(D)$ to be the set of all of the values of the noise variables $\nu_1, \ldots, \nu_t$ which lead to $a_i = \perp$ for all $i \in N^C$ when the mechanism is run on D, conditioning on $\hat{T} = Z$ and $a_i = \hat{a}_i$ for all $i \in N$. Because the sensitivity of each query is $1/n$, we have:

$$\mathcal{A}_{Z-1/n}(D') \subseteq \mathcal{A}_Z(D) \subseteq \mathcal{A}_{Z+1/n}(D')$$

This is because switching from D to D' can raise the value of each query by at most $1/n$, which will cause each below-threshold query to remain below-threshold if we also raise the threshold by $1/n$. Also, because we selected $\hat{T}$ by perturbing T from the Laplace distribution, we have:

$$\Pr[\hat{T} = Z] \leq \exp(\epsilon/2) \cdot \Pr[\hat{T} = Z + \frac{1}{n}]$$

Therefore, we can calculate:

$$\begin{aligned}
\prod_{i \in N^C} \Pr_D[a_i = \perp | \hat{a}^{<i}] &= \int_{-\infty}^{\infty} \Pr[\hat{T} = Z] \cdot \Pr[(\nu_1, \ldots, \nu_t) \in \mathcal{A}_Z(D)] dZ \\
&\leq \exp\left(\frac{\epsilon}{2}\right) \int_{-\infty}^{\infty} \Pr[\hat{T} = Z + \frac{1}{n}] \cdot \Pr[(\nu_1, \ldots, \nu_t) \in \mathcal{A}_Z(D)] dZ \\
&\leq \exp\left(\frac{\epsilon}{2}\right) \int_{-\infty}^{\infty} \Pr[\hat{T} = Z + \frac{1}{n}] \cdot \Pr[(\nu_1, \ldots, \nu_t) \in \mathcal{A}_{Z+1/n}(D')] dZ \\
&= \exp\left(\frac{\epsilon}{2}\right) \int_{-\infty}^{\infty} \Pr[\hat{T} = Z] \cdot \Pr[(\nu_1, \ldots, \nu_t) \in \mathcal{A}_Z(D')] dZ \\
&= \exp\left(\frac{\epsilon}{2}\right) \prod_{i \in N^C} \Pr_{D'}[a_i = \perp | \hat{a}^{<i}]
\end{aligned}$$

Therefore we have:

$$\sum_{i \in N^C} \log\left(\frac{\Pr_D[a_i = \perp | \hat{a}^{<i}]}{\Pr_{D'}[a_i = \perp | \hat{a}^{<i}]}\right) = \log\left(\frac{\prod_{i \in N^C} \Pr_D[a_i = \perp | \hat{a}^{<i}]}{\prod_{i \in N^C} \Pr_{D'}[a_i = \perp | \hat{a}^{<i}]}\right) \leq \frac{\epsilon}{2}$$

which completes the proof. ■

What did we show in the end? That if we are given a sequence of queries together with a guarantee that only at most c of them have answers above $T - \alpha$, we can estimate the answers to every query that takes value at least $T + \alpha$ to error α, while reporting that all of the others are below this threshold. This accuracy is equal, up to a factor of $\log k$, to the accuracy we would get, given the same privacy guarantee, if we knew the identities of these large above-threshold queries ahead of time, and answered them with the Laplace mechanism. That is, the sparse vector technique allowed us to fish out the identities of these large queries almost "for free", paying only logarithmically for the irrelevant queries. This is the same guarantee that we could have gotten by trying to find the large queries with the exponential mechanism and then answering them with the Laplace mechanism. This algorithm, however, is trivial to run, and crucially, allows us to choose our queries adaptively.

If we wanted a guarantee of (ϵ, δ)-differential privacy, we could make a similar claim by using the composition theorems[1]. Specifically, we could have modified the algorithm by selected $\sigma = \frac{\sqrt{32c \ln 1/\delta}}{\epsilon n}$. By making a similar argument, but employing our composition theorems, we could have shown:

Theorem 4 *The modified sparse vector algorithm is (ϵ, δ)-differentially private.*

[1] Actually the composition theorems cannot be applied in a black box manner. This is because in our analysis considers c invocations of the Laplace mechanism under non-trivial conditioning. Nevertheless, a careful analysis shows that similar bounds do indeed still hold under this conditioning.

and:

Theorem 5 *For any sequence of k queries $Q_1, \ldots, Q_k$ such that $L(T) \equiv |\{i : Q_i(D) \geq T - \alpha\}| \leq c$, the modified Sparse$(T, C)$ is (α, β) accurate for:*

$$\alpha = 2\sigma \left(\log k + \log \frac{2}{\beta} \right) = \frac{\sqrt{128c \ln 1/\delta}(\log k + \log 2/\beta)}{\epsilon n}.$$

Bibliographic Information The sparse vector algorithm and analysis given here is from Hardt and Rothblum, "A Multiplicative Weights Mechanism for Privacy Preserving Data Analysis", 2010. Previous variants on this technique with inferior bounds and privacy guarantees had been used by Dwork, Naor, Reingold, Rothblum, and Vadhan in "On the Complexity of Differentially Private Data Release", 2009, and by Roth and Roughgarden in "Interactive Privacy via the Median Mechanism", 2010.