CIS 800/002 The Algorithmic Foundations of Data Privacy September 27, 2011

Lecture 5

Lecturer: Aaron Roth *Scribe: Aaron Roth*

The Net Mechanism

In this lecture we consider the *query release* problem:

> Given a collection of queries $Q_1, \ldots, Q_k : \mathbf{N}^{|\mathcal{X}|} \to \mathbf{R}$, come up with answers $v_1, \ldots, v_k$ to minimize the maximum error: $\max_i |Q_i(D) - v_i|$.

Definition 1 (Accuracy) *A query release mechanism M is (α, β)-*accurate *with respect to queries in class C if for every database $D \in \mathbf{N}^{|\mathcal{X}|}$, with probability at least $1 - \beta$, the output of the mechanism represents answers $\{v_i\}$ such that:*

$$\max_{Q_i \in C} |Q_i(D) - v_i| \leq \alpha$$

Recall that last class we proved:

Theorem 2 *Let $\epsilon, \delta \geq 0$. The class of ϵ'-differentially private mechanisms satisfies (ϵ, δ)-differential privacy under k-fold adaptive composition for:*

$$\epsilon = \sqrt{2k \ln(1/\delta')}\epsilon' + k\epsilon'(e^{\epsilon'} - 1)$$

Typically, we will design an algorithm in which each of k intermediate steps is ϵ'-differentially private, and we will want to solve for the largest value of ϵ' that causes the entire algorithm to be ϵ, δ-differentially private. For $\epsilon' \leq 1$, $\epsilon'(\exp(\epsilon') - 1) \leq 2\epsilon'^2$, and so we can solve for:

$$\epsilon = \sqrt{2k \ln(1/\delta')}\epsilon' + 2k\epsilon'^2$$

An easy way to do this is to simply solve:

$$\frac{\epsilon}{2} \geq \sqrt{2k \ln(1/\delta')}\epsilon' \qquad \frac{\epsilon}{2} \geq 2k\epsilon'^2$$

to find:

$$\epsilon' \leq \frac{\epsilon}{\sqrt{8k \ln 1/\delta'}} \qquad \epsilon' \leq \sqrt{\frac{\epsilon}{4k}}$$

For $\epsilon \leq 1$, the first expression is always the binding constraint, so for ϵ-differential privacy, we can simply set:

$$\epsilon' = \frac{\epsilon}{\sqrt{8k \ln 1/\delta'}}$$

Compare this to if we wanted pure ϵ-differential privacy, in which case we would have had to set $\epsilon' = \epsilon/k$.

Lets consider what this means for answering k sensitivity 1 queries with the Laplace mechanism. For ϵ-differential privacy, we can answer each query by adding noise $Y_i \sim \text{Lap}(k/\epsilon)$. Recall that $\text{E}[|Y_i|] = k/\epsilon$, and that:

$$\Pr[|Y_i| \geq t \cdot \frac{k}{\epsilon}] = \exp(-t)$$

Therefore, if we have queries $Q_1, \ldots, Q_k : \mathbf{N}^{|\mathcal{X}|} \to \mathbf{R}$ and release the answers $v_i = Q_i(D) + Y_i$, we have that except with probability at most β:

$$\max_i |Q_i(D) - v_i| \leq \frac{k}{\epsilon} \cdot \log\left(\frac{k}{\beta}\right)$$

On the other hand, for (ϵ, δ)-differential privacy, we can answer each query by adding noise

$$Y_i \sim \text{Lap}\left(\frac{\sqrt{8k \ln 1/\delta}}{\epsilon}\right)$$

And so have that except with probability at most β:

$$\max_i |Q_i(D) - v_i| \leq \frac{\sqrt{8k \ln 1/\delta}}{\epsilon} \cdot \log\left(\frac{k}{\beta}\right)$$

More generally, we have shown:

Theorem 3 *The Laplace mechanism is (α, β)-accurate for any of k sensitivity Δ queries for*

$$\alpha = \Delta \cdot \frac{k}{\epsilon} \cdot \log\left(\frac{k}{\beta}\right)$$

when preserving ϵ-differential privacy and:

$$\alpha = \Delta \cdot \frac{\sqrt{8k \ln 1/\delta}}{\epsilon} \cdot \log\left(\frac{k}{\beta}\right)$$

when preserving (ϵ, δ)-differential privacy.

Our goal will be to design a query release algorithm that has its usefulness parameter scale only logarithmically with k. Such a mechanism will be able to answer exponentially many more queries than the Laplace mechanism, given the same accuracy guarantees.

One way to concisely represent the answers to many queries is by outputting another database. A mechanism that does this is said to release *synthetic data*. For such a mechanism $M : \mathbf{N}^{|\mathcal{X}|} \to \mathbf{N}^{|\mathcal{X}|}$, if $D' = M(D)$, we say $v_i = Q_i(D')$.

We will derive a useful mechanism from the existence of small α-nets:

Definition 4 (α-net) *An α-net of databases with respect to a class of queries C is a set $N \subset \mathbf{N}^{|\mathcal{X}|}$ such that for all $D \in \mathbf{N}^{|\mathcal{X}|}$, there exists an element of the α-net $D' \in N$ such that:*

$$\max_{Q \in C} |Q(D) - Q(D')| \leq \alpha$$

We write $N_\alpha(C)$ to denote an α-net of minimum cardinality among the set of all α-nets for C.

Note that we can index an element in a set S by only $\log |S|$ bits. Therefore, intuitively, if a class of queries has a small α-net, then it means that the answers to every query in that class (up to error α) represent only a small amount of information. We want to take advantage of this.

Algorithm 1 The Net Mechanism

NetMechanism(D, C, ϵ, α)

Let $\mathcal{R} \leftarrow N_\alpha(C)$

Let $q : \mathbf{N}^{|\mathcal{X}|} \times \mathcal{R} \to \mathbf{R}$ be defined to be:

$$q(D, D') = -\max_{Q \in C} |Q(D) - Q(D')|$$

Sample And Output $D' \in \mathcal{R}$ with the exponential mechanism Exponential$(D, q, \mathcal{R})$

We first observe that the Net mechanism preserves ϵ-differential privacy.

Proposition 5 *The Net mechanism is ϵ-differentially private.*

Proof The Net mechanism is simply an instantiation of the exponential mechanism. We showed in Lecture 3 that the exponential mechanism preserves differential privacy. ■

What about accuracy?

Proposition 6 *For any class of queries C the Net Mechanism is $(2\alpha, \beta)$-useful for any α such that:*

$$\alpha \geq \frac{2\Delta}{\epsilon} \log \frac{N_\alpha(C)}{\beta}$$

Where $\Delta = \max_{Q \in C} GS(Q)$.

Proof Recall our utility theorem for the exponential mechanism:

$$\Pr\left[q(D, \text{Exponential}(D, q, \mathcal{R})) \leq \text{OPT}_q(D) - \frac{2\Delta}{\epsilon}\left(\log\left(|\mathcal{R}|\right) + t\right)\right] \leq e^{-t}$$

By the definition of an α-net, $\text{OPT}_q(D) \geq -\alpha$, and we have $\mathcal{R} = N_\alpha(C)$. Therefore, setting $t = \log(1/\beta)$, we have for $D' = \text{NetMechanism}(D)$:

$$\Pr[q(D, D') \leq -2\alpha] \leq \Pr[q(D, D') \leq -\alpha - \frac{2\Delta}{\epsilon} \log \frac{N_\alpha(C)}{\beta}] \leq \beta$$

which proves the claim. ■

So we have reduced the problem of proving upper bounds for the accuracy of differentially private release mechanisms for a class of queries C to the problem of finding small nets for the class of queries C.

We will now consider the natural class of *counting queries* and show that they do indeed admit small nets.

Definition 7 *A* counting query Q_φ*, defined in terms of a predicate $\varphi : X \to \{0, 1\}$ is defined to be*

$$Q_\varphi(D) = \frac{\sum_{x \in D} \varphi(x)}{|D|}.$$

It evaluates to the fraction of elements in the database that satisfy the predicate φ.

For the sake of this discussion, lets assume that $n = |D|$ is public knowledge (If not, it is easy to estimate with the Laplace mechanism). Note that a counting query is just a normalized subset sum query. If we write $D \in \mathbf{N}^{|\mathcal{X}|}$ as a historam, we can write $Q_\phi \in \mathbf{N}^{|\mathcal{X}|}$, which is just the truth table of ϕ: $Q_\phi[i] = \phi(x_i)$. Then $Q_\varphi(D) = \frac{1}{n}\langle Q_\varphi, D\rangle$.

Observation 8 *For any predicate $\varphi : X \to \{0, 1\}$, the corresponding counting query $Q_\varphi : X^n \to [0, 1]$ has global sensitivity $GS_{Q_\varphi} \leq 1/n$*

For counting queries, we therefore have the following corollary:

Corollary 9 *For any class of counting queries C the Net Mechanism is $(2\alpha, \beta)$-useful for any α such that:*

$$\alpha \geq \frac{2}{\epsilon n} \log \frac{N_\alpha(C)}{\beta}$$

It remains to bound the size of the smallest α-net for a class of counting queries.

We recall the additive Chernoff bound, which is a special case of Azuma's inequality which we saw earlier:

Theorem 10 (Additive Chernoff Bound) *Let $X_1, \ldots, X_m$ be independent random variables bounded such that $0 \leq X_i \leq 1$ for all i. Let $S = \frac{1}{m}\sum_{i=1}^{m} X_i$ denote their mean, and let $\mu = \mathrm{E}[S]$ denote their expected mean. Then:*

$$\Pr[S > \mu + \epsilon] \leq e^{-2m\epsilon^2}$$

$$\Pr[S < \mu - \epsilon] \leq e^{-2m\epsilon^2}$$

Theorem 11 *For any finite class of counting queries C:*

$$|N_\alpha(C)| \leq |X|^{\frac{\log|C|}{\alpha^2}}$$

In order to prove this theorem, we will show that for any collection of counting queries C and for any database D, there is a "small" database D' of size $|D'| = \frac{\log|C|}{\alpha^2}$ that approximately encodes the answers to every query in C, up to error α. Crucially, this bound will be independent of $|D|$.

Lemma 12 *For any $D \in X^*$ and for any finite collection of counting queries C, there exists a database D' of size*

$$|D'| = \frac{\log|C|}{\alpha^2}$$

such that:

$$\max_{Q \in C} |Q(D) - Q(D')| \leq \alpha$$

Proof Let $m = \frac{\log|C|}{\alpha^2}$ We will construct a database D' by taking m uniformly random samples from the elements of D. Specifically, for $i \in \{1, \ldots, m\}$ let X_i be a random variable taking value x_j with probability $|\{x \in D : x = x_j\}|/|D|$, and let D' be the database containing elements $X_1, \ldots, X_m$. Now fix any $Q_\varphi \in C$ and consider the quantity $Q_\varphi(D')$. We have: $Q_\varphi(D') = \frac{1}{m}\sum_{i=1}^{m} \varphi(X_i)$. We note that each term of the sum $\varphi(X_i)$ is a bounded random variable taking values $0 \leq \varphi(X_i) \leq 1$, and that the expectation of $Q_\varphi(D')$ is:

$$\mathrm{E}[Q(D')] = \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}[\varphi(X_i)] = \mathrm{E}[\varphi(X_i)] = \sum_{x_j \in X} \frac{|\{x \in D : x = x_j\}|}{|D|}\phi(x_j) = Q_\varphi(D)$$

Therefore, we can apply the Chernoff bound which gives:

$$\Pr\left[|Q_\varphi(D') - Q_\varphi(D)| > \alpha\right] \leq 2e^{-2m\alpha^2}$$

Taking a union bound over all of the counting queries $Q_\varphi \in C$, we get:

$$\Pr\left[\max_{Q_\varphi \in C} |Q_\varphi(D') - Q_\varphi(D)| > \alpha\right] \leq 2|C|e^{-2m\alpha^2}$$

Plugging in m makes the right hand side smaller than 1 (so long as $|C| > 2$), proving that there exists a database of size m satisfying the stated bound, which completes the proof of the lemma. ■

Now we can complete the proof of Theorem 11.

Proof [of Theorem 11] By Lemma 12, we have that for any D there exists a database D' with $|D'| = \frac{\log|C|}{\alpha^2}$ such that $\max_{Q_\varphi \in C} |Q_\varphi(D) - Q_\varphi(D')| \leq \alpha$. Therefore, if we take $N = \{D' \in X^* : |D'| = \frac{\log|C|}{\alpha^2}\}$ to be the set of *every* database of size $\frac{\log|C|}{\alpha^2}$, we have an α-net for C. Since

$$|N| = |X|^{\frac{\log|C|}{\alpha^2}}$$

and by definition $|N_\alpha(C)| \leq |N|$, we have proven the theorem. ■

Bibliographic Information The Net Mechanism is from "A Learning Theory Approach to Non-Interactive Database Privacy" by Blum, Ligett, and Roth, 2008.