CSCI 5520: Foundations of Data Privacy | Homework 1
The Chinese University of Hong Kong, Spring 2015 | due Tuesday 3 March

Please turn in your solution in class on Tuesday 3 March. You are encouraged to collaborate on the homework and ask for assistance, but you are required to write your own solutions, list your collaborators, acknowledge any sources of help, and provide external references if you have used any.

Question 1

In this question you will show that the database reconstruction algorithm from Lecture 6 can be made efficient.
We will say that a vector $y \in [-2,2]^m$ is *β-heavy* if at least $m/10$ of its coordinates have absolute value at least β. Let

$$q'_S(y) = \sum_{i \in S} y_i - \sum_{i \notin S} y_i$$

where S is a subset of $[m]$ and y is a vector in $\mathbb{R}^m$.

(a) Show that if $y \in [-2,2]^m$ is 1/4-heavy and S is a random subset of $[m]$, then there exists a sufficiently small constant γ (independent of m) such that

$$\Pr[q'_S(y) \geq \gamma\sqrt{m}] \geq \gamma.$$

(b) Let G be a finite subset of $[-1,1]^m$ and $\mathcal{S}$ be a collection of s random independent subsets of $[m]$. Show that the probability there exist $x \in \{-1,0,1\}^m$ and $x' \in G$ such $x - x'$ is 1/4-heavy but $q'_S(x-x') < \gamma\sqrt{m}$ for all $S \in \mathcal{S}$ is at most $3^m|G|(1-\gamma)^s$.

(c) Show that if $s \geq Km\log m$ for a sufficiently large constant K, then with probability at least 1/2 over the choice of $\mathcal{S}$, for every $x \in \{-1,0,1\}^m$ and every $x' \in [-1,1]^m$ such that $x - x'$ is 1/3-heavy, there exists a set $S \in \mathcal{S}$ such that $q'_S(x-x') \geq \gamma\sqrt{m}/2$. (**Hint:** Take G to be a sufficiently dense grid in $[-2,2]^m$.)

(d) Suppose that M is a mechanism that on input[1] $x \in \{-1,0,1\}^m$ and query q'_S outputs an approximation to $q'_S(x)$ with additive error $\gamma\sqrt{m}/6$. Show that with constant probability, the following algorithm outputs a vector $\hat{x}$ that agrees with x on $9m/10$ of its coordinates:

(i) Choose a collection $\mathcal{S}$ of s independent uniform random subsets of $[m]$.

(ii) Query M to obtain approximations a_S to $q'_S(x)$ for all $S \in \mathcal{S}$.

(iii) Find $x' \in [-1,1]^m$ such that $q'_S(x') \leq a_S + \gamma\sqrt{m}/6$, if it exists.
(This is a linear program; it can be solved efficiently.)

(iv) For every coordinate i, set

$$\hat{x}_i = \begin{cases} 1, & \text{if } x'_i \geq 1/2, \\ -1, & \text{if } x'_i \leq 1/2, \\ 0, & \text{otherwise} \end{cases}$$

and output $\hat{x}$.

[1] In the actual database, we include the row $(i,1)$ if $x_i = 1$, $(i,-1)$ if $x_i = -1$, and do not include a row that starts with i otherwise.

Question 2

In this question you will that if a synthetic database mechanism is differentially private then its output is unlikely to contain rows from the original database. Let $M\colon D^n \to D^d$ be a synthetic database mechanism.

(a) Let $x \in D^n$ be a database whose rows are independent uniform samples from D and x' be a database obtained by resampling the ith row of x uniformly from D and independently of the other rows. Show that

$$\Pr_{M,x,x'}[M(x') \text{ contains the } i\text{-th row of } x] \le d/|D|.$$

(b) Use part (a) to show that if M is (ε, δ)-differentially private, then

$$\Pr_{M,x,x'}[M(x) \text{ contains at least one row of } x] \le e^{\varepsilon} dn/|D| + \delta n.$$

(c) Now let $\mathcal{D}$ be an arbitrary distribution over D and assume the rows of x and x' are sampled as in part (a), but from $\mathcal{D}$ instead of the uniform distribution over D. Show that

$$\Pr_{M,x,x'}[M(x) \text{ contains at least one row of } x] \le e^{\varepsilon} pdn + \delta n.$$

where $p = \max_r\{\Pr_{R\sim\mathcal{D}}[R = r]\}$. (You do not need to redo the proofs from parts (a) and (b), just explain the differences.)

(d) **(Extra credit)** Now suppose x is chosen from the following distribution: The i-th row of x equals $(i, 0)$ with probability $1/2$ and $(i, 1)$ with probability $1/2$, independently from the other rows. If the output of $M(x)$ contains 99% of the rows of x with probability at least 99%, can M be $(0.1, n^{-100})$-differentially private for sufficiently large n?

Question 3

Let P be a subset of $\{0, 1\}^n$. A *testing algorithm* for property P is a randomized algorithm M such that $\Pr[M(x) \text{ accepts}] \ge 2/3$ for every $x \in P$ and $\Pr[M(x') \text{ accepts}] \le 1/3$ for every $x' \in \{0, 1\}^n$ that differs from all $x \in P$ in at least εn coordinates.

(a) Show that every P has a $O(1/\varepsilon n)$-differentially private testing algorithm.

(b) A testing algorithm is *one-sided* if $\Pr[M(x) \text{ accepts}] = 1$ for every $x \in P$. Which P have a $(100, 0.1)$-differentially private one-sided testing algorithm?