Homework 1 – Due Tuesday, February 9, 2010 before the lecture

Collaboration and Honesty Policy: Collaboration on homework problems is permitted. If you choose to collaborate on some problems, you are allowed to discuss each problem with at most 2 other students currently enrolled in the class. Before working with others on a problem, you should think about it yourself for at least 45 minutes. Finding answers to problems on the Web or from other outside sources (these include anyone not enrolled in the class) is strictly forbidden.

- You must write up each problem solution by yourself without assistance, even if you collaborate with others to solve the problem.
- You must identify your collaborators. If you did not work with anyone, you should write ”Collaborators: none.”
- Your solution to each problem should be handed in on a separate sheet of paper. Different problems might be graded by different people.

Problems

1. (Sampling) You are given a list of $n$ real numbers $a_1, \ldots, a_n$ in the interval $[0, 1]$ with the average $\bar{a}$. Let $s_1, \ldots, s_k$ be random samples taken from the list uniformly with replacement, and $\hat{a}$ be the average of the samples. Let $error = |\hat{a} - \bar{a}|$.

   (a) Based on the Chernoff-Hoeffding bound seen in class, what should $k$ be to guarantee that $error \leq \alpha$ with probability at least $1 - \delta$?

   (b) Suppose that half of $a_i$’s are 0s, and the rest are 1s. Show that there exists constant $c_1 > 0$, such that $error \leq c_1/\sqrt{k}$ with probability at least $2/3$. [Hint: There are several ways to approach the problem. One way is to show that $\hat{a}$ takes on any given value with probability at most $O(1/\sqrt{k})$, and then look at how many values $\hat{a}$ can take in the interval $\frac{1}{2} \pm c_1/\sqrt{k}$. Another approach is to apply the Central Limit Theorem from statistics.]

2. (Simple Consequences of the Differential Privacy Definition)

   (a) (Randomization) Show that a non-trivial differentially private algorithm has to be randomized. More specifically, that if a deterministic algorithm $A$ does not output the same answer on all inputs, it is not differentially private.

   (b) (Group Privacy) Your friend and his family are participating in a study where the results will be released via a differentially private algorithm. He is concerned that differential privacy only gives a guarantee for databases that differ in one person, and is wondering whether all but one of family members should withdraw from the study because of privacy concerns. Suppose $A$ is $\varepsilon$-differentially private. What guarantee can you give for two databases that differ in at most $k$ entries?

3. (Composition and Postprocessing)
(a) **(Nonadaptive)** Prove that if a randomized algorithm \( A \) runs \( k \) algorithms \( A_1, \ldots, A_k \), where each \( A_i \) is \( \varepsilon_i \)-differentially private, and outputs \( A(x) = g(A_1(x), \ldots, A_k(x)) \) for some randomized algorithm \( g \) then \( A \) is \( \left( \sum_{i=1}^{k} \varepsilon_i \right) \)-differentially private. (You may use statements proved in class without reproving them.)

(b) **(Adaptive)** Now suppose that algorithm \( A \) chooses algorithms \( A_1, \ldots, A_k \) and \( g \) adaptively. That is, it first decides which algorithm \( A_1 \) to run. Then, when it gets \( A_1(x) \), it chooses \( A_2 \), and so on. Prove the statement in part (a) for an adaptive \( A \).

4. **(Differentially Private Elections)** A function **majority** on 0/1 inputs is defined as follows: \( f_{\text{maj}}(x_1, \ldots, x_n) \) is 1 when \( \geq n/2 \) arguments are 1, and 0 otherwise. Give an \( \varepsilon \)-differentially private algorithm with the following property: if the input contains \( \geq n/2 + k \) occurrences of bit \( b \) then your algorithm should output \( b \) with probability at least \( 1 - e^{\varepsilon \cdot k/4} \). (Hint: Use the global sensitivity framework.)

5. **(Median-finding using sum queries)** Given a set \( X \) of \( n \) real numbers \( x_1, \ldots, x_n \) in \([0, 1] \), the rank of a value \( y \) is the number of indices \( i \) such that \( x_i \leq y \). We say \( y \) is a median of \( X \) if it has rank \( \lceil n/2 \rceil \). Give a differentially private algorithm which takes \( X \) as input and approximates the median in the following sense: after asking \( t \) questions with global sensitivity 1, with probability at least 2/3, the algorithm should output an interval of width \( 2^{-t} \) that contains a value with rank \( n/2 \pm t \log(t) \).