**This page is intentionally left blank.**
**Please do not open this question set before you are allowed to do so.**

**4810-1184 Algorithms for Information Security and Privacy (Autumn 2018)**

**Midterm Problem 1**

Suppose that we have a table $T$ with the weight information of all persons in a city. We want to use the Laplacian mechanism to publish the average weight information $AVG(T)$. However, for now, let suppose that $f(T) = AVG(T)/2$.

Question 1.1: Calculate the value of $GS(f)$.

Question 1.2: What is the distribution that the noise we add to $f(T)$ should be drawn from?

Question 1.3: Discuss why your noise in Question 1.2 should be smaller than when $f(T) = AVG(T)$.

Question 1.4: Discuss why, even the noise is smaller, scaling the value of $f$ will not help users to have a clearer information.

From next question, we will use the exponential mechanism to privately publish the average weight $AVG(T)$. Suppose that a publication of the mechanism $p$ must be a member of the set $\{40,50,60,70,80,90\}$.

Question 1.5: Design a utility function for each $p$.

Question 1.6: Calculate the value of $\Delta Utility$.

Question 1.7: Discuss why, when using your utility function, the exponential mechanism will work well.

Question 1.8: Discuss why publishing $AVG(T)/2$ instead of $AVG(T)$ will also not work in the exponential mechanism with your utility function. You can assume that, when trying to publish $AVG(T)/2$, the possible publication can be $\{20,25,30,35,40,45\}$.

**4810-1184 Algorithms for Information Security and Privacy (Autumn 2018)**

**Midterm Problem 2**

Suppose that our database $T$ has 2 information, chosen party and ages, of all users. Also, for simplicity, let assume that there is only 2 parties in our country, Melon-pan and Gyu-don. We know in advance that there will be two queries to our database, which are:

1) How many persons choose "Melon-pan" and has age below 30?
2) What is the percentage of persons who is below 40 and choose "Melon-pan" party?

Question 2.1: Discuss why reducing the table $T$ to $T'$ with the following information:

1) chosen party
2) if the person ages below 30 (boolean variable)
3) if the person ages below 40 (boolean variable)

will help reducing the computation time of the smallDB algorithm.

Question 2.2: Design a function $f_1$ for Query 1. Your function must follow the assumption on the function for smallDB algorithm discussed in the class.

Question 2.3: Design a function $f_2$ for Query 2. Your function must follow the assumption on the function for smallDB algorithm discussed in the class.

From next question, suppose that $\alpha = 0.3$.

Question 2.4: Calculate the number of records in the published small database, when $\log 2 = 0.7$.

Question 2.5: Calculate the number of possible small tables.

From next question, assume that, the histogram of the table $T'$ is as follows:

- Number of records with "Melon-pan" and "age below 30" is 500000
- Number of records with "Melon-pan" and "age over 30" is 0
- Number of records with "Gyu-don" and "age below 40" is 0
- Number of records with "Gyu-don" and "age over 40" is 500000

Question 2.6: Calculate the number of possible small table $U$ with $\max_i[f_i(T') - f_i(U)] = 0$.

Question 2.7: Calculate the number of possible small table $U$ with $\max_i[f_i(T') - f_i(U)] = 1/8$.

Question 2.8: Based on your answer in Questions 2.6 and 2.7, try to a fast algorithm for smallDB algorithm. By the algorithm, can you reduce the value of $\alpha$ to a smaller value?

**4810-1184 Algorithms for Information Security and Privacy (Autumn 2018)**

**Midterm Problem 3**

Suppose that we want to publish the percentage of persons choosing Melon-pan party, i.e.

$$f(T) = \frac{\#Melon - pan}{||T||}.$$

We know that $f(T)$ must be between 0 and 1.

Question 3.1: Discuss why Laplacian mechanism can give a number smaller than 0 or larger than 1.

Question 3.2: Discuss why there would not be a privacy problem although the number is not between 0 and 1.

From next question, let consider the case that it is very unlikely to have $f(T) + noise$ smaller than 0 or larger than 1. The probability that we have such the case it negligible.

Suppose that there is an attacker who want to know $f(T)$. The attacker keeps asking from the value $f(T) + noise$ for $n$ times. Assume that he has got $X_1, \dots, X_n$ for the queries.

Question 3.3: Discuss why the expected value of $X_i$ is $f(T)$.

Let $S$ be an average of $X_i$. The attacker will predict that $f(T)$ is $S$.

Question 3.4: Discuss why the expected value of $S$ is $f(T)$.

Question 3.5: Discuss why the probability that $S$ is different from $f(T)$ by more than $\alpha$ is no more than $2 \cdot \exp(-2n\alpha^2)$.

Question 3.6: Calculate the probability that the attacker's prediction has error smaller than $\sqrt{\frac{\ln 10}{200}} \approx 0.11$ when $n = 100$.

Question 3.7: Discuss why your answer in Question 3.6 would imply a privacy problem for Laplacian mechanism. Discuss why we have already included this problem in the composition theorem.