IMPERIAL COLLEGE LONDON

TIMED REMOTE ASSESSMENTS 2020-2021

MEng Honours Degree in Electronic and Information Engineering Part IV
MEng Honours Degree in Mathematics and Computer Science Part IV
MEng Honours Degrees in Computing Part IV
MSc Advanced Computing
MSc in Computing (Specialism)
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant assessments for the Associateship of the City and Guilds of London Institute

PAPER COMP70018=COMP97012=COMP97013

PRIVACY ENGINEERING

Monday 14 December 2020, 14:00
Duration: 140 minutes
Includes 20 minutes for access and submission

Answer ALL TWO questions
Open book assessment

By completing and submitting work for this assessment, candidates confirm that the submitted work is entirely their own and they have not (i) used the services of any agency or person(s) providing specimen, model or ghostwritten work in the preparation of the work they have submitted for this assessment, (ii) given assistance in accessing this paper or in providing specimen, model or ghostwritten answers to other candidates submitting work for this assessment.

Paper contains 2 questions

1 Attempting to anonymize data

Your friend just had an idea for a fancy new pseudonymization mechanism. The mechanism M_N uses a secret parameter, an integer N taken uniformly at random in the range $[N_1, N_2]$. For a given *N*, the mechanism applies a cryptographically secure hashing function *hash* repeatedly *N* times to the input *x*:

$$M_N(x) = (hash)^N (x) = hash(hash(...hash(x))) \text{ [repeated n times]}$$

Your friend intends to use this mechanism to transform the original IDs and compute the pseudonymized IDs of a dataset with 100,000 users (and one row for each user). You know the original ID x_a of a record in the dataset. Assume that the hashing function *hash* has no collisions, that you can compute hash(x) in 1 millisecond, and that the time taken by all other operations is negligible.

a i) If $N \in [1, 10^3]$, how long does it take for you to find the record with original ID x_a in the dataset in the worst case?

ii) What is the worst case and the average case (both in years) if $N \in [10^{10}, 10^{100}]$? (Note: you can use any tool incl. Wolfram Alpha to convert the result from ms to years)

iii) How long in years does it take to hash a dataset of 100,000 entries with this mechanism if $N = 10^{56}$?

iv) Consider the standard mechanism from the course which concatenates a salt to records, $M_C(x) = \text{hash}(x||\text{salt})$, where the salt is a random string of numeral characters (0,...,9) of length 100. How long would this take to break in the worst case? How long would it take to pseudonymize the dataset? Which mechanism is better?

b Imperial has a dataset about students with three columns: date of birth, postcode and department. The College decides to give you access to this dataset via a QBS. You know that the QBS allows *only* counting queries that test conjunctions of equalities (not others such as ≠, >, or <), i.e. exclusively queries of the type:

count(date of birth == X AND postcode == Y AND department == Z)

You have a friend Bob who was born on 11-10-1995 and lives in postcode SW79TY (and you know he's the only one born on that day in that postcode).

Bob claims that he's registered in the Department of Computing, but you've never seen him in class.
i) How could you find out if he's lying by checking if he is in the dataset? What's the name of this attack?

Now suppose that Imperial introduced query set size restriction (QSR) with threshold T to avoid this attack. (Recall: this means that the QBS returns TooLow if the true count is smaller than T, and returns the true count otherwise.) However, the threshold T used for QSR is not consistent, but rather sampled at random for each query (even if the same query is repeated, a new T is sampled). Specifically, T is sampled from a normal distribution with mean 4 and standard deviation 2, i.e. T ~ $N(4,2^2)$.

ii) Suppose that you can send an unlimited amount of queries. How can you infer if Bob is in the dataset?

c You have access to the Department of Computing's students' final grades for one course. Every grade is an integer between 1 and 20. This is an example of how a dataset could look like:

Pseudonymized student ID	**Grade**
C6JXiPa1	12
eqtuSdbj	14
o7L3ZQk4	17
JfgSR923	9

We define D as the set of all possible such tables.

You want to release the means of grades in the dataset in a way that is both useful and privacy-preserving. To do this, we propose to first compute the sum and the number of grades separately in a differentially private (DP) way, then use their ratio as an estimate of the mean.

We call C: D → N the function that takes a dataset D and returns its number of users, i.e. C(D) = |D|.

We call S: D → N the function that takes a dataset D and returns the sum of all grades S(D) = $\sum_{i=1}^{|D|} grade_i$.

i) Compute the sensitivity of the function S. Justify your answer step by step (e.g., define the neighboring datasets).

ii) What is the scale *b* of the Laplace noise N_1 that you need to add in order for the result of the sum query to be ε-DP with $\varepsilon = 1$? What is the variance of the noise?

We call M_1 the mechanism $M_1(D) = S(D) + N_1$, where $N_1 \sim$ Lap*(b)*, and M_2 the mechanism $M_2(D) = C(D) + N_2$ where $N_2 \sim$ Lap($1/\varepsilon$).

We finally define the mechanism for the mean grade as:

$$M(D) = M_1(D) / M_2(D)$$

To avoid releasing weird results to the analyst (such as a negative average), we slightly modify the mechanism M to output only values greater or equal than 1. Specifically, we define this modified mechanism M' as M'(D) = max(1, M(D)).

iv) Is M' still differentially private? Justify your answer

However, as we discussed during the course we want to build solutions that preserve privacy but also allow the data to be used for good, e.g. for academic research.

v) Knowing that, for a Laplace random variable with scale parameter *s*, the 95% confidence interval is approximately *[-s*3, s*3]* and assuming that the noise on the denominator is negligible ($M_2(D) = |D|$), how large is the confidence interval of the noise added to the mean, as a function of the dataset size |D| and the privacy budget ε ?

Reminder: the 95% confidence interval is the range of values (centered in the mean value) within which the noise will be sampled with probability 95%

vi) Using this result, how many students need to be in D for the error on the result to be < 1 point with 95% probability, with ε =0.1? How useful would your mechanism be for the Privacy Engineering class (80 students) ?

The department looked at your mechanism and is very interested. They want to use DP to release the collection of mean grades for *every* course in the department. You can assume that a student can take at most 6 courses, and that there are 100 available courses in the department. Here’s an example of how the dataset can look like:

Pseudonymized student ID	**Grade for course 1**	**Grade for course 2**	**Grade for course 3**	**...**	**Grade for course 100**
8dj20s2h	18	N/A	19	...	13
safhu71n	N/A	N/A	N/A	...	N/A
2398nf0a	N/A	17	15		13
...	...	...	...	...	...

vii) How can you modify the mechanism M to produce a mechanism M’’ that releases all the 100 mean grades and achieves ε-DP? Is the total number of available classes important?

Hint: use the sensitivity of a multivariate function.

The three parts carry, respectively, 27.5%, 20%, and 52.5% of the marks.

2a Show that the sum of the Lagrange basis polynomials $\delta_i(x)$ is 1 for any N distinct points $x_1, x_2, \dots, x_N$. *Hint: interpolate points on the function* $f(x) = 1$.

b Consider the following secure Multi-party computation.

Three parties wish to compute the function

$$f(A, B, C) = (A + B) * C \quad (mod\ 11)$$

using the BGW protocol for polynomials of degree 1 and arithmetic ***modulo*** **11**.

The private values of Parties 1, 2 and 3 are A=2, B=3 and C=4 respectively.

For this configuration answer questions b(i) to b(v) below.

Use a table with 3 columns, one for each of the party when doing the calculations b(i) to b(iv) below.

You will lose marks for not showing your working.

(i) Calculate the shares that each Party computes for its private values.
Use ***Sxy*** to denote the share that Party ***x*** computes for Party ***y***.
Use the coefficients 1, 2 and 3 for the polynomials of Parties 1, 2 and 3 respectively.

(ii) Calculate the output shares that each party computes for its ADD gate.
Use ***Ax*** to denote the output share that Party ***x*** computes.

(iii) Calculate the recombination vector $(\delta_1(0), \delta_2(0), \delta_3(0))$ for all polynomials of degree up-to-at-most 2.

(iv) Using the recombination vector in part b(iii) calculate the output share that each party produces for its MUL gate.
Use ***Mx*** to denote the output share for Party ***x***.
Use the same coefficients as in part b(i) for the degree reduction step.

(v) Explain whether the final BGW broadcast and recombination step is really required to produce the correct result for a circuit ending in a MUL gate.

Question 2 is continued on the next page.

c Consider the following **1-from-*N*** oblivious transfer protocol in an honest-but-curious setting.

All bit-strings in this protocol are of the same length.

Alice's messages are the bit-strings $M_1, \ldots, M_N$.

Bob's message selection *value* is $\boldsymbol{B}$. $1 \leq B \leq N$.

The protocol proceeds as follows:

1. Alice generates the all-zeros bit-string K_0

2. Alice and Bob then do Steps 2a and Step 2b for $H = 1, \ldots, N$

   2a. Alice generates a random bit-string K_H.

   2b. Alice and Bob run the following **1-from-2** oblivious transfer protocol:

   Bob uses 0 for the selection *bit* if $H = B$ otherwise Bob uses 1.

   Alice uses the following two messages:

   $C_0 = K_0 \oplus K_1 \oplus \ldots \oplus K_{H-1} \oplus M_H$ and
   $C_1 = K_H$

For this protocol:

(i) What values does Bob learn from Alice?

(ii) Show how Bob can recover M_B?

(iii) Explain how the protocol satisfies the properties of an oblivious transfer.

The three parts carry, respectively, 10%, 60%, and 30% of the marks.