Total points: 60 (+20 bonus)

Assignment 3

Due: 3:00 PM, 9th May, 2017 (Tue)

Written assignment

1. [13 points] Fill in the following table, which compares four data anonymization techniques (k-anonymity, differential privacy, secure multiparty computation, private information retrieval):

Name	# Data holders	Comp. cost	Distortion	Leakage	Private queries
	(1/1+/2+)	(Low/High)	(Yes/No)	(Yes/No)	(Yes/No)
k-anonymity	1			Yes	
	1+	High			Yes
	2+	High			No
		Low		No	

In the table, “Distortion” indicates whether or not the data is changed at all by the anonymization procedure. “Leakage” indicates whether or not it is possible to deduce new sensitive information that the data holder did not intend to give. “Private queries” indicates whether or not the data holder(s) will know the query/algorithm.

2. [11 points] Company A keeps a backup of its data. Every Sunday, a full backup is done. Every Monday, Tuesday, and Wednesday, a differential backup is performed. Every Thursday, Friday, and Saturday, an incremental backup is performed.

   (a) [4 points] Suppose on Sunday Company A needs to recover its data after the Saturday backup. However, one of the backup copies during the week has been corrupted and cannot be used. Which copy, when corrupted, would cause the most data to be lost? Explain your answer.

   (b) [3 points] Suggest a technique to minimize the chance of data being corrupted. Suggest one downside of using your technique.

   (c) [2 points] How can Company A minimize the worst case data loss, supposing that the company must use three differential backups and three incremental backups throughout the week?

   (d) [2 points] Explain why backups can be encrypted versions of the original data, but backups cannot be hashes of the original data.

3. [16 points] Imagine you are running the website of a company, and the company wants the website to be resilient to DDoS attacks. To do so, you are to write a business continuity plan. Briefly draft a short business continuity plan containing the following four sections: (1) Business Impact Analysis, including recovery objectives; (2) Solutions,

including continuity strategy and security controls; (3) Disaster Recovery Plan; and (4) Ongoing Plan Implementation and Maintenance.

All sections in your draft should specifically concern DDoS attacks against the company's website. You are free to make up numbers, statistics, and other details about the company to augment your draft. (You should probably write at least three sentences per section.)

Programming assignment

k-**anonymity** [20 points]

We learned in class that k-anonymity is used to prevent linkage between *identifiers* and *sensitive attributes*. k-anonymity has some limitations; in particular, it cannot account for background knowledge, especially from complementary data release. In this question, we will examine another issue of k-anonymity known as *minimality*. While it is desirable that k-anonymity mechanisms should minimize how much they *change* the data set, it is sometimes computationally hard to ensure minimal change.

(a) [20 points] The Vidiian Sodality (government) wishes to study a disease called the Phage. They have asked hospitals to release a data set to investigate a possible link between Age and Phage infection, while protecting data privacy. For this question, there will be one identifier: **Age**, which is an integer. There is one sensitive attribute, whether or not the person has contracted a sensitive disease called the **Phage**, indicated by 0 or 1.

To anonymize the data set, you are allowed to change any Age identifier (but not the Phage attribute). You must change the Age to another integer (e.g. change 28 to 25); you **cannot** change any data to a string (e.g. change 28 to 2X). Two people with the same Age are considered to be in the same anonymity set (even if they have a different Phage attribute).

In order to preserve the usefulness of the data set, you are asked to change the identifiers minimally. The definition of minimal change is as follows. Suppose there are n entries, and the list of Age was $a_1, a_2, \ldots, a_n$ Suppose after your anonymization changes, they become $a'_1, a'_2, \ldots, a'_n$ respectively. The **change** between the two data set is defined as:

$$\sum_{i=1}^{n}(|a'_i - a_i|)$$

In other words, the change is computed by summing all the differences between the Ages of the original data set and the anonymized data set.

Write a program, `anonymize`, which implements 5-anonymity on a data set with **minimal** change. The name of the data set file will be the first argument, so that `anonymize` would be called as follows to anonymize `datafile`:

./anonymize datafile

Remember that the data set file can have other names than datafile.

The data set is in CSV format, where each line is an entity's Age, Salary, and Phage, separated by commas (a sample has been provided). You may change any Age and Salary value to another **integer**, but do not change the Phage attribute. `anonymize` should overwrite the original file with the anonymized data set. There may be several different anonymized data sets that achieve minimal change; any is acceptable.

The following fact may be useful for you: suppose the minimal change k-anonymity solution to the set $a_1, a_2, \ldots, a_n$ (sorted ascendingly) is $A_1, A_2, \ldots, A_{m-1}, A_m$. Each A_i contains at least k elements of $a_1, a_2, \ldots, a_n$ and they are all distinct; let us write $A_m = \{a_f, a_{f+1}, \ldots, a_n\}$. Then $A_1, A_2, \ldots, A_{m-1}$ is a minimal change k-anonymity solution to the set $a_1, a_2, \ldots, a_{f-1}$.

Your program should be fast. (See below on testing.)

(b) [20 points (bonus)] It turns out that the study in part (a) did not reveal any notable relationship between Age and Phage. The Vidiian Sodality has asked hospitals to release another data set, with more identifiers: Age, Height, Weight, Width, and Number of Children. (Width is a particularly private identifier in Vidiian society.) All of those are integers; two people belong to the same anonymity set only if all five identifiers match. As before, the last (sixth) column will be the sensitive attribute, Phage, which is 0 or 1, and should not be changed. Similarly, you are asked to achieve **minimal** change.

Write a program, `anonymize_full`, which implements 5-anonymity on a data set with **minimal** change. The name of the data set file will be the first argument, so that `anonymize_full` would be called as follows to anonymize `datafile`:

./anonymize_full datafile

`anonymize_full` should overwrite the original file with the anonymized data set.

Note that the problem of **minimal** change is in general a computationally hard (NP-hard) problem. In other words, achieving minimal change efficiently in general is impossible. You should simply try your best to minimize change as much as possible while still achieving 5-anonymity. (See below for testing.)

Submission instructions

All submissions should be done through the CASS system. For this assignment, there is **no** Milestone deadline. Submit the following programs:

- `a3.pdf`, containing all your written answers.
- Any amount of source code, with a `Makefile` that will create `anonymize`, and `anonymize_full`, for the programming assignment.

You may submit in C++ without a `Makefile`. In that case, the two source code files must be exactly named as `anonymize.cpp`, and `anonymize_full.cpp`. (Do not submit compiled files.)

For Python, you also do not need to submit a `Makefile`. In that case, simply call your source code file `anonymize` and `anonymize_full`. You will probably need to add a shebang to the top of your code (`!/usr/bin/python2` or `!/usr/bin/python3`) so that they can be run correctly.

Keep in mind that plagiarism is a serious academic offense; you may discuss the assignment, but write your assignment alone and do not show anyone your answers and code.

The submission system will be closed exactly 48 hours after the due date of the assignment. Submissions after then will not be accepted unless you have requested an extension before the due date of the assignment. You will receive no marks if there is no submission within 48 hours after the due date.

Testing

For part (a), there will be 10 data sets, ranging from simple, small data sets to large ones with a maximum of 100 elements. I will run your code for at most 10 seconds on each data set. You will score 2 points per test for a correct answer and no points for an incorrect answer.

For part (b), I will run your program alongside my own program on 5 randomly generated data sets. Each data set has 500 entities with reasonable random values for all attributes, and I will run our programs for 10 seconds. After 10 seconds, I will kill our programs, and compare our changes by reading the input file again. Your score will be as follows:

My algorithm beats yours by...	Score
≤ 5	10
6–10	9
11–25	8
26–50	7
51–75	6
76–100	5
101–200	4
200–500	3
500–1000	2
1000+	1
Not 5-anonymized	0

Your overall mark for this question will be the average of all 5 scores. In addition, if you beat my algorithm at least on one data set, you will gain a further bonus of 10 points. (My algorithm is not supremely advanced or optimized, so this is entirely possible; go for it.)

Since I will simply kill our programs at the time limit, your code should overwrite the original file frequently, perhaps whenever it finds a better solution. If it waits too long, I might kill your code before it outputs anything, which would be bad. You may consider it a good idea to voluntarily halt your program when the time is up.

A note on minimality

An interesting fact about minimality on a k-anonymized data set is that it opens the data set to inference attacks. Even if the attacker does not know anything about the data, she can still deduce new information just from the fact that the anonymization procedure was specifically designed to minimize information loss. The issue is even more serious if the anonymization procedure attempted to obey ℓ-diversity as well. This fact highlights a conflict between utility and privacy in k-anonymity.

Prof. Raymond Wong at our CSE department wrote the first academic paper on this phenomenon.