COMP 430/530: Data Privacy and Security - Fall 2020
Homework Assignment # 1

Question 1. [21 pts] Recall the 7 principles of Privacy by Design (PbD).

Imagine that you are employed as the Chief Privacy Officer of a major hospital. The hospital has made the decision to build an online system where patients' hospital visits, health records, test results, prescribed medication, etc. will be stored and accessed digitally. The system will be open to the hospital personnel (doctors, physicians, nurses) within the hospital intranet, as well as patients who would like to access their information remotely via the Internet.

As the Chief Privacy Officer, you are given the task to write a report regarding how the hospital's system will conform to the 7 principles of PbD. For each of the 7 principles, describe a related design decision you have made when designing the hospital's digital system and how that decision satisfied the corresponding PbD principle.

Example:

Principle #1 - state the principle - give a 2-4 sentence summary of the design decision and how the design decision satisfies the principle.

Principle #2 - state the principle - give a 2-4 sentence summary, etc.

Grading: Each principle and its description together are worth 3 pts, total 7*3=21 pts.

Question 2. [12 pts] Imagine now that the digital hospital system in Question 1 has been built and it is being actively used. There can be several privacy threats to this system. In particular, consider the following actors:

- An **honest-but-curious** doctor.
- A **malicious** nurse.
- A patient's **honest-but-curious** mother.
- A **man-in-the-middle (MITM)** adversary.

For each of these actors, describe one action they can take and how that action may cause a privacy leakage. It is important that the action you describe fits the given adversary type (consider the definitions of keywords: *honest-but-curious*, *MITM*, *malicious*).

Grading: Each actor is worth 3 pts, total 4*3=12 pts.

Question 3. [21 pts] Download the Adult Census Income dataset that is supplied to you on Blackboard. This is a simplified version of the Adult dataset from the UCI ML Repository: `http://archive.ics.uci.edu/ml/datasets/Adult`.

The Adult dataset has been a commonly used benchmark dataset in the privacy literature. It is a tabular dataset, where each row corresponds to one individual. Each column (attribute) contains information such as the individual's age, gender, marital status, education level, etc. The final column (*income*) is a binary column that states whether the individual's yearly income is $\leq$ 50K or $>$ 50K dollars.

Download the dataset **from Blackboard** (adult.csv) and familiarize yourself with it – it is important you download from Blackboard since the version on Blackboard is slightly simplified compared to the original dataset. Then, answer the following:

(a) [1 pt] How many attributes does the dataset have, including *income*?

(b) [2 pts] Clean all rows that contain one or more missing values. You will use your cleaned version in the rest of this question. How many rows does the clean dataset have?

(c) [5 pts] Create a histogram for understanding the education levels of individuals with high income (> 50K). Put the different education levels on the x axis. The y axis should be counts, i.e., the number of individuals with that education and income > 50K. **Note:** Make sure all axes and information in the histogram are properly labeled and readable. You will be graded on the correctness of your histogram **and** its visual presentation.

(d) [6 pts] You are given the task of analyzing whether somebody's race is related to their income. You decide that you can perform such analysis using conditional probabilities. Let X be a random variable denoting race: $X \in \{\text{White, Black, Asian-Pac-Islander}\}$. (You can ignore the remaining races.) For each race, you want to compute:

$$\Pr[\text{income} > 50\text{K} \mid \text{race} = \text{X}] \text{ and } \Pr[\text{income} \leq 50\text{K} \mid \text{race} = \text{X}]$$

Thus, there are 3*2=6 conditional probabilities in total. For each conditional probability, write the range-count queries you need to compute that probability. You may write the range-count queries in SQL syntax or verbally (one sentence for each query).

(e) [4 pts] Execute the queries from part (d) on the Adult dataset and calculate the conditional probabilities.

(f) [3 pts] Interpret the results you obtained in part (e) and give your conclusions to the analysis task from part (d).

Grading: Submit your written answers to parts (a)-(f). If you wrote any code to perform the tasks, e.g., in Python or R, submit the code files via Blackboard.

Question 4. [10 pts] Ali is the CEO of an imaginary company, SecureCorp. In order to protect the privileged information that only he should access, Ali had his engineering team build a retina scanner for biometric authentication. The retina scanner is built to only authenticate Ali and reject everyone else's login attempts.

The retina scanner's logs show that there have been a total of 1,000 login attempts to this day. 865 attempts were successful. Ali says that he made 900 login attempts himself and was able to successfully log in 850 times.

Read about the terminology TPR, TNR, precision, accuracy and F1 score from `https://en.wikipedia.org/wiki/Sensitivity_and_specificity`. Based on the terminology and the information given above, answer the following:

(a) [2 pts] What is the *True Positive Rate (TPR)* of SecureCorp's retina scanner?
(b) [2 pts] What is the *True Negative Rate (TNR)* of SecureCorp's retina scanner?
(c) [2 pts] What is the *precision* of SecureCorp's retina scanner?
(d) [2 pts] What is the *accuracy* of SecureCorp's retina scanner?
(e) [2 pts] What is the *F1 score* of SecureCorp's retina scanner?

Question 5. [24 pts] MegaCorp, an imaginary company, stores customer usernames and passwords on their "secure" servers. They use SHA-512 for hashing passwords but no salts. You just hacked into MegaCorp's servers and stole a part of their username-password list. You stored the stolen list in a file called megacorp.txt (megacorp.txt is available on Blackboard).

Since you are an elite hacker, you are also aware of the Rockyou breach and you suspect some Rockyou users are also customers of Megacorp. You download the Rockyou password dataset – a small version, which is sufficient for this question, is available on Blackboard (rockyou.txt).

(a) [6 pts] Create a dictionary attack using Rockyou. Given rockyou.txt, your code should generate a csv file that contains the dictionary attack table. Submit your source code and the output of your code (the attack table).

(b) [3 pts] What are the passwords of MegaCorp users Alice, Bob, Charlie? Use the attack table from part (a) to infer their passwords.

Now consider that MegaCorp have updated their security and use salts when storing passwords. Yet, they have not updated remaining aspects of their security; thus, you were able to hack into their servers again and this time you stole the salted username-password file: salty-megacorp.txt.

(c) [3 pts] Does your dictionary attack from parts (a) and (b) work? Why or why not?

(d) [4 pts] The file salty-megacorp.txt contains users Dave, Elaine, Faith. Devise an attack to find their passwords. Verbally describe your attack strategy and discuss whether this attack requires less computation or more computation than what you did in parts (a) and (b).

(e) [8 pts] Implement your new attack. Submit its source code as well as the true passwords of Dave, Elaine and Faith.

Hint: If you are using Python, here is a good resource for SHA-512 hashing: `https://docs.python.org/3/library/hashlib.html`.

Hint 2: You cannot directly hash characters, you should first convert them to bytes.

Question 6. [12 pts] Assume an SQL database with 2 tables:

STUDENTS(studentid, firstname, lastname, DateOfBirth)

COURSES(courseid, coursename, capacity)

Write the appropriate access control statement for each of the following.

(a) [3 pts] Frank just joined the Registrar's Office as a new employee. He should be given rights to add new students to the STUDENTS table. But since Frank is a new employee, he should be able to only *add* new students, not delete or update existing students.

(b) [3 pts] Jill is in charge of course planning. She should be given all rights on the COURSES table.

(c) [3 pts] Frank has left his job. His rights should be removed.

(d) [3 pts] Matthew is working under Jill. His task is to manage course capacities. He should be given rights to update course capacities, but not other columns.

Grading: Minor errors in syntax will be forgiven, but logic errors (e.g., giving redundant rights or not giving the sufficient rights) will be penalized.