1 2 9 0 dei mecd

Security and Privacy

2022/2023

Nuno Antunes, nmsa@dei.uc.pt
Nuno Cardoso, ncardoso@uc.pt

Assignment 2: Privacy-Preserving Data Sharing

December 19th @ 23:59

Objectives

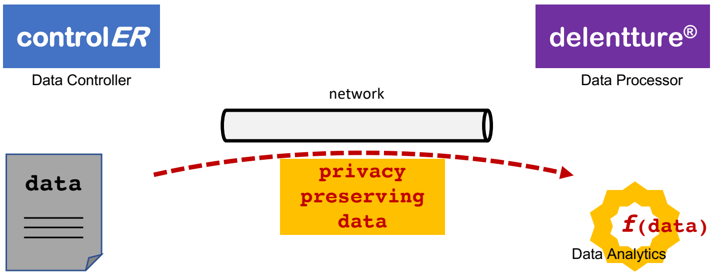
- Study how privacy preserving techniques can be used in data processing scenarios.
- Apply these techniques in the solution of concrete problems.
- Compare advantages and disadvantages of the implemented solutions.

Preparation

Before starting the exercise, you should carefully read this assignment and understand what tasks are to be conducted and the goals to be achieved.
The assignment is prepared to be executed by **Groups of 2 elements**, in a total of 48 hours.

Problem statement

The company **control*ER*** provides a service of consumer loans. Recently, the company noticed a significant raise in loan contract infringements, and it is worried about the current strategy in-place for assessing loan risk. **control*ER*** does not have a data analysis/science team yet but obtaining insights on strategies to restrict loan approval is an urgent matter. As such, **control*ER*** decided to contract the services of **delentture®**, a consulting company with highly specialized data science teams. The figure depicts the scenario.



In the previous exercise you already addressed the problem of transferring highly sensitive customer data to **delentture®**. But now, it was decided that we do not want that the data shared is able to harm privacy of the data subjects, and therefore, new strategies are required. For this, you will design, implement, and evaluate 3 different strategies to address the problem according to what you learned in the classes, each one in one of the following exercises.

Resources

`ucstudent.uc.pt/`**`2023_SP_Dataset_v2.zip`** includes two `.csv` files:

- `columns_description.csv` – describes the meaning of each of the columns of the dataset.
- `infringement_dataset.csv` – the content of the dataset to be used.

(**changes v2**): several columns were removed because they were less relevant

Exercise 1: Anonymization with Privacy Models

You will implement a solution proposed by one **control*ER*** engineers, which consists in anonymize the data before sharing using the studied syntactic privacy models. Using the techniques learned in the course, the **control*ER*** modifies the dataset before sharing it, in such what that it respects the desired levels of privacy risk and maintaining an acceptable level of utility.

Note: you should use the ARX tool. The dataset must be imported, and this may require sanitization (e.g., fixing charsets, conversion of dates, eliminating non-conformant registers, etc).

For this, you will have to perform the following tasks:

1. Characterize the dataset by classifying attributes.
2. Analyze the distinction and separation of the different potential quasi-identifiers.
3. Measure privacy risks of the dataset in original form.
4. Define and configure the coding model to use. Specify the hierarchies to be used for anonymization and the attribute weights.
5. Define privacy requirements, i.e., acceptable intervals for parameters of the anonymization process (e.g., suppression limit, coding model, attribute weights, etc.).
6. Apply **1** privacy models/configurations on the dataset.
    a. Analyze utility and privacy risk of the resulting dataset.
    b. If requirements are not met, tune and adjust according.
7. Study, compare and discuss utility and privacy of the resulting dataset.
8. Execute both analyses ***f()*** prepared in the Assignment 1 and compare the results with the ones originally obtained.
    a. Discuss advantages and disadvantages.
    b. Take note of the results and observations, for your report and future comparisons.

Exercise 2: Differential Privacy

You will implement a second solution proposed by the **control*ER*** engineers: instead of anonymizing the data, you implement a differential privacy solution that provides **delentture®** with the result of the necessary queries, while protecting the privacy of the database records. You should **NOT** use the ARX tool. The engineers know that this solution provides added guarantees of privacy, with less assumptions, but the utility of the results is yet to be understood.

For this, you will have to perform the following tasks:

1. Plan how you will execute both analyses ***f()***.
2. Study the sensitivity of both analyses ***f()***.
3. Implement a differential privacy mechanism to add to the analyses.
4. Execute the analysis with the implemented differential privacy.
5. Analyze the results and compare with the results obtained in the other phases.
    a. Discuss advantages and disadvantages.
    b. Take note of the results and observations, for your report and future comparisons.

Exercise 3: Synthetic Data

You will implement a third solution proposed by the **control*ER*** engineers. In this case, you will generate a synthetic dataset to be shared with **delentture®** for analysis. Using synthetic data raises smaller consent concerns, but also here the data utility needs to be carefully studied.

For this, you will have to perform the following tasks:

1. Plan/decide on the most appropriate tools and techniques for generating the dataset.
2. Generate a syntactic dataset, with similar shape (rows/columns) as the real dataset.
3. Report the steps to generate the synthetic dataset from the real data, including any steps/techniques used to improve privacy.
4. Use evaluation metrics to assess the quality of the synthetic data. Discuss the results.
5. Execute both analyses ***f()*** in the generated dataset and compare the results with the remaining ones.
    a. Discuss advantages and disadvantages.
    b. Include in your report the results and observations.

Deliverable

- A report containing the evaluation results, their analyses and discussion.
- The complete sources developed and necessary for the results obtained.

Questions

If you have questions about the scope of the work or any other aspect, please talk with the Professors of the course. You can do so face-to-face, by email, or using Skype.