Security and Privacy
2020/2021
Nuno Antunes, nmsa@dei.uc.pt

Assignment 2

January 05th @ 23:59

Objectives:

- Understand how to apply in practice syntactic and semantic models for privacy preserving data publishing.
- Compare advantages and disadvantages of syntactic and semantic models.

Preparation: Selection of dataset and problem

1. Select a dataset to use, identifying the privacy concerns included in the dataset.
2. Convert the dataset to a tabular format (if it is not already) and perform the necessary sanitization (e.g., fixing charsets, conversion of dates, fixing delimiters, eliminating non-conformant registers, etc).
3. Define **3** different analysis that are relevant on the dataset. Implement those analysis in terms of a query or a procedure and obtain the resulting values.

⇒ the dataset resulting of this phase serves as input for both of the following phases, which can be developed in parallel.

Exercise 1: Use syntactic models to improve privacy

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. In particular, specify the hierarchies to be used for anonymization and the attribute weights.
5. Apply two different privacy models/configurations on the dataset.
6. Study, compare and discuss utility and privacy of the two resulting datasets.
7. Execute the analysis defined before and compare the results with the baseline.

Exercise 2: Use differential privacy to improve privacy

1. Study the sensitivity of the analysis implemented during the preparation phase.
2. Implement a differential privacy mechanism to add to the analyses.
3. Execute the analysis with the implemented differential privacy.
4. Analyze the results and compare with the results obtained in the other phases.
5. Discuss advantages and disadvantages.

Deliverable:

- A small report containing the results obtained and the analyses/discussions;
- The complete sources developed and necessary for the results obtained;