MSI 2023/2024

Security and Privacy

Assignment 1: Anonymization of Datasets with Privacy, Utility and Risk Analysis

1. Introduction

Objective: The main objective of this assignment is to perform a detailed analysis of the **anonymization process** of a dataset.

Groups: maximum 2 students
Specify your group and fill the following form:
https://docs.google.com/spreadsheets/d/1z1Fk4NhQnEUlINT4YifpjwUColbD1hxqcBb02zA8cBM/edit?usp=sharing

Final deadline: March 19, 2024. (Submit the project and a report in PDF at Inforestudante)

Assignment defenses: slots will be made available.

2. Description

The assignment is organized in 5 steps, in each step you should perform a detailed analysis of the choices made and include it in the final report:

1. Find a dataset
2. Choosing, sanitizing and characterizing a dataset for the anonymization process and conducting a detailed analysis, selection and configuration of appropriate anonymization/privacy models that you will then apply to the dataset.
3. Configure and apply the appropriate anonymization/privacy models.
4. Perform an analysis and optimization of the utility and privacy levels of the selected anonymization/privacy models, as well an analysis of the risk of re-identification of selected anonymization/privacy models.
5. Write a report documenting, analyzing, and reasoning on the choices made in the previous steps.

On steps 3 & 4, you should explore the parameter space of the privacy models (e.g., suppression limits, coding models, attribute weights, utility measures, etc.) for a in-depth analysis of results.

Step #1 – Selection of Dataset
You should select a dataset of your choice for anonymization. The dataset should be rich enough (e.g., in terms of number of rows and columns) to allow for effective anonymization with different privacy

models (required in step #2).

You should specify the goal with the release of the anonymized dataset. For example, say you have a dataset with information about smartphone apps. You may want to determine what are the categories with greater success (ratings), without being able to identify concrete apps in the anonymized dataset. In the end of this project, you should evaluate how does that goal fare when determined through the anonymized dataset vs the original dataset.

Step #2 –Importing and Characterizing of Dataset

The dataset must be imported into ARX, and this may require sanitization (e.g., fixing charsets, conversion of dates, fixing CSV delimiters, eliminating non-conformant registers, mixing tables, etc.), depending on quality level of the dataset at hand.

Upon importing the dataset into ARX, you should now characterize the dataset, by classifying attributes (identifying , quasi-identifying, insensitive, or sensitive). Analyze in deep the distinction and separation of the different potential quasi-identifiers. You should also characterize/analyze the privacy risks of the dataset in original form, as well as analyze the characteristics of the dataset to make sure it follows a reasonable distribution of data (this is particularly important if you are generating a synthetic dataset).

At this stage you should also look at the distribution of the attributes. Based on this analysis, you should define and configure the coding model to use. In particular, specify the hierarchies to be used for anonymization and the attribute weights.

Based on this analysis, you should also define privacy requirements, i.e., acceptable intervals for parameters of the anonymization process (e.g., suppression limit, coding model, attribute weights, etc.). This may be an iterative process with step #3.

Step #3 – Apply Privacy Models

At this stage you should apply **at least three privacy models**[1] ***(groups should apply at least 4 privacy models),*** **one of which is a new privacy model that is not discusses in the class** to your target dataset **(**i.e., you need to understand and describe the new privacy model in the document**)**. The privacy models can be applied all together or in different combinations. You must justify your choice, by making an informed decision of the privacy models according to the desired privacy requirements and characteristics of the dataset.

Then, conduct a detailed analysis about the performance of each privacy model applied to the selected dataset (e.g., according to parameters such as suppression limit, coding model, attribute weights, utility metrics, etc.). If the results are not satisfying according to the requirements, you should perform several iterations, by either adapting the parameters of the privacy models or considering other privacy models.

Step #4 Analysis of Utility, Privacy and Risk Assessment

At this stage you should conduct a comprehensive analysis of the utility and privacy levels achieved by each of the privacy models. You should choose and analyze the results of appropriate utility and privacy metrics and strive for an acceptable balance between the two. This may require revisiting the privacy models for refinement.

You should also perform a detailed utility and risk analysis by considering appropriate metrics and models for measuring the utility and the re-identification risk.

Step #5 Final Delivery

Write a final report that should include the reasoning behind the choices made in each of the previous

[1] You should do 2 (3 for groups) independent analysis in ARX, it is not enough to just put all the privacy models in one analysis.

steps. It should also include a set of recommendations on the process of anonymizing a dataset. You should see this final report as a professional service of consultancy that could be sent to the security/privacy department of a company that own the dataset and is looking for services for anonymization of gathered data.

3. Evaluation Criteria

- Selection and characterization of dataset [20%]
  - Characterization of dataset
  - Classification of attributes
  - Definition of the goal of the anonymization process
  - Definition of privacy and utility requirements
- Coding Model [20%]
  - Characterization and analysis of dataset
  - Definition and configuration of coding model (e.g., hierarchies)
- Privacy models: utility, privacy, and risk assessment [40%]
  - Detailed analysis of privacy models' results according to varying privacy model's parameters
  - Comprehensive analysis of the utility and privacy levels
  - Detailed risk-analysis of re-identification risk
- Assignment defense [20%]