PRIVACY ENHANCING TECHNOLOGIES – ASSIGNMENT # 1

"Anonymization of Datasets with Privacy, Utility and Risk Analysis"

The goal of this assignment is to perform a detailed analysis of the **anonymization process** of a dataset.

The assignment is organized in 4 steps, in each step you should perform a detailed analysis of the choices made and include it in the final report:

1. You will start by choosing, sanitizing and characterizing a dataset for the anonymization process;
2. In the second step, you will conduct a detailed analysis, selection and configuration of appropriate anonymization/privacy models that you will then apply to the dataset;
3. You will 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.
4. As final step, you will write a report documenting, analyzing and reasoning on the choices made on each of the previous steps.

On steps 2 and 3, you should consider exploring the parameter space of the privacy models (e.g. suppression limits, coding models, attribute weights, utility measures, etc) for an encompassing analysis of results.

Step #1 – Selection, Importing and Goal 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).

Some suggested websites:
http://www.kaggle.com
https://archive.ics.uci.edu/ml/index.php
https://www.analyticsvidhya.com/blog/2016/11/25-websites-to-find-datasets-for-data-science-projects/

You should specify the *goal for 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 the project, you should evaluate how does that goal fare when determined through the anonymized dataset vs the original dataset. I.e. you should produce statistics about your goal for the raw dataset and compare those results with the anonymized dataset.

The dataset must be imported into ARX, and this may require a bit of 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.

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 #2.

Step #2 – Characterization of the Dataset and Coding Models

Upon importing the dataset into ARX, you should now characterize the dataset, by classifying attributes (insensitive, sensitive, quasi-identifying, or identifying). 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 or using 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.

Step #3 – Privacy Models: Utility, Privacy and Risk Assessment

At this stage you should apply **at least two privacy models** to your target dataset. 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, 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.

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.

You should evaluate the feasibility of extracting information from the anonymized dataset (vs the original dataset) for the goal you have defined initially, and try other anonymization approaches if needed. Note, however, that the dataset has to be anonymized properly (i.e. low levels to re-identification risk), not only to achieve the goal that you have defined, but for public release for others to eventually pursue other goals.

FINAL REPORT

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

Submit the PDF report at the course moodle.

Final deadline: **26th October 2022**

Assignment defenses: week of Nov 10 (TBC).

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
    - Including the utility with respect to your goal, but also other utility metrics
  - 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
    - Again, including the utility with respect to your goal, but also other utility metrics
  - Detailed risk-analysis of re-identification risk
- Assignment defense [20%]