Description

The goal of this short project is to study a differentially private algorithm in a centralized publishing context and to challenge it empirically against simple statistical studies that you will have chosen. You will not have to implement any differentially private algorithm, too time-consuming. You will rather rely on the OpenDP[1] library. You will **work in pairs** and have **20 mins for defending** your study (**15 mins talk and 5 mins questions and answers**). When a task below is tagged by the ⋆ symbol, it is optional.

Dataset The dataset that you will use is the adult data set[2] that you already used in the practical work session about the Laplace mechanism.

Algorithm The differentially private algorithm that we propose you to study will be one of the following:

- A differentially private synopsis computation[3].
- A differentially private synthetic data generator among those proposed by OpenDP[4]. In order to limit technical issues, we recommend you to favor marginal synthesizers[5] (*e.g.*, AIM, MST) against neural network synthesizers[6] (*e.g.*, DP-GAN, PATE-GAN).

Tasks Your mission, provided that you agree, will consist in the following tasks:

T1 **Choose an algorithm:** Choose the differentially private algorithm that you want to study according to its specifications. No need to get informed about all the algorithms. You can first pick a couple of algorithms according to a first brief overview, and then elect your algorithm by reading more. Disclaimer: you *will* need to read one or two additional documents (e.g., paper, slides, tutorial on the web) in order to catch the intuitions behind the algorithms.

T2 **Identify a simple statistical question:** Based on your background, experience, or external references in public statistics, identify a simple question that will need to be answered through statistics. For example, does the income of women differ from the income of men, all along their life?

T3 **Evaluate utility (simple question):** Run the chosen algorithm on the adult dataset for answering the chosen statistical question. Is the result with differential privacy *close*[7] to the result without differential privacy? Note that for the synopsis algorithm, this will require to identify the aggregates that deserve to be part of the synopsis. You will perform this comparison for two different values of ϵ: $\epsilon \in \{1, 100\}$. If needed by the chosen algorithm, the value of δ can be set to $\delta = 10^{-5}$. The chosen algorithm might need many other parameters. You can let them to their default values for simplicity. Do not hesitate to repeat your comparison in order to obtain average/min/max results.

T4 ⋆ **Identify a harder statistical question:** After having gained insights on the chosen algorithm, identify a question that you expect to be much more challenging to answer with differentially private guarantees. For example, questions that rely on correlations among various dimensions are usually much harder.

T5 ⋆ **Evaluate utility (harder question):** Similarly to above, compare the result obtained with differential privacy to the result obtained without differential privacy.

T6 **Synthesize your study in a set of slides:** Your slides will need to explain the following points:

- The chosen algorithm (intuitively) or the aggregates of the synopsis if you have chosen the synopsis.
- The statistical questions identified. In particular, what makes it simple or challenging.
- The settings of your experiments (*e.g.*, parameters, distance used).
- Your empirical observations (please plot graphs), analysis, and conclusions.

[1] https://opendp.org/
[2] https://archive.ics.uci.edu/ml/datasets/Adult
[3] https://github.com/opendp/smartnoise-sdk/blob/main/sql/samples/Synopsis.ipynb
[4] https://github.com/opendp/smartnoise-sdk/tree/main/synth
[5] https://docs.smartnoise.org/synth/synthesizers/index.html#marginal-synthesizers
[6] https://docs.smartnoise.org/synth/synthesizers/index.html#neural-network-synthesizers
[7] You will need to choose a reasonable distance between the two results (e.g., euclidean, cosine, earth mover, *etc*).

Expectations

- ☞ **Clarity** of the explanations (chosen algorithm, statistical questions, experiments, observations and analysis).
- ☞ **Relevance** of the statistical questions identified.
- ☞ **Rigor** of the observations and of the analysis.

Good luck!