Submissions

⚠ **Group** : work by **pairs**.

⚠ **Deliverable:** your **pdf report** (**5 pages** at most) and the accompanying code (**a zipfile**).

⚠ **Submission:** by email to tristan.allard@irisa.fr.

⚠ **Deadline:** the 10th of Dec 2023 (11:59pm).

Expectations

This project requires a fair amount of **experimental work**. Please check that:

✓ Your report **describes your approach** in a **self-contained** manner (i.e., there must be no need to read other documents in order to understand your report).

✓ Your report explains precisely the **experimental environment** (e.g., programing language, hardware, external libraries, dataset) and the **experimental methodology** (e.g., number of repetitions, parameters).

✓ Your report **displays, describes, and thoroughly analyzes** the graphs you plot (e.g., the error according to the privacy level ϵ).

Indications

☞ We suggest you to work with Python, but you are free to choose your favorite language.

☞ This project uses the Adult dataset (only `adult.data`).

☞ You can either implement from scratch the privacy mechanism that you will use, or import it from external libraries (e.g., the `reprosyn` package [1]).

Your Mission

In order to obtain the prestigious position of privacy engineer at Koukle labs ©, you need to show your skills in privacy-preserving data publishing. The head of the lab wants to know the performances of lets you **until the 10th of Dec 2023** to **choose, implement (or rely on an external library), and evaluate experimentally the utility of a centralized publishing mechanism** enabling statisticians to compute **offline arbitrary** count queries. He asked you to use the well-known Adult dataset. The queries targetted are **count queries** on the `age` attribute, over **arbitrary ranges** – for example, "the number of individuals between 20 and 29 years old".

Your mission might be split in the following tasks:

- **Design** of your approach: privacy model (e.g., k-anonymity, ϵ-differential privacy), privacy mechanism (e.g., Mondrian, histogram, hierarchy of histograms [2], synthetic data generation [1, 3]), utility measures (e.g., the average relative error used in the practical work sessions of the class).
- **Implementation** of your approach: from scratch or based on external libraries.
- **Design** of your experiments: imagine the graphs that will allow you to fulfill your mission (keep them simple, draw them before running the experiments: what do **you** expect?), values of the various parameters (e.g., the ranges queried, the values for ϵ, values for the specific paramters of your privacy mechanism (default ones?), number of repetitions).
- **Running** your experiments and plotting the graphs.
- **Write** your report: describe your privacy mechanism and the privacy model that it satisfies (explain your choices), your experimental environment, your experimental methodology, include the most relevant graphs and analyze them carefuly, conclude.

[1] https://github.com/alan-turing-institute/reprosyn

References

[1] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the nist contest: A scalable and general approach to differentially private synthetic data, 2021.

[2] Wahbeh H. Qardaji, Weining Yang, and Ninghui Li. Understanding hierarchical methods for differentially private histograms. *Proceedings of VLDB*, 6(14):1954–1965, 2013.

[3] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (SIGMOD '14)*, 2014.