Institute of Computer Science, University of Bern
Privacy and Data Security, HS 2021

Prof. Christian Cachin
Orestis Alpos, Luca Zanolini

Exercise 6

6.1 k-Anonymity and ℓ-diversity (2pt)

Prove that a dataset that has been sanitized either according to distinct ℓ-diversity or according to probabilistic ℓ-diversity also satisfies ℓ-anonymity.

6.2 Implementing k-anonymity with the Mondrian algorithm (8pt)

In this problem, we explore an algorithm to compute a k-anonymous version of a given dataset. In particular, we explore the *Mondrian* algorithm [LDR06] as implemented in this basic prototype, available as open-source program written in Python:

`https://github.com/qiyuangong/Mondrian`

Download `Mondrian`, study the documentation and the source, and modify it to process the file `ex05-fake-registrations.csv`. You may want to extend the program a bit to make it more flexible to use.

For assessing the utility retained in the sanitized dataset, `Mondrian` computes a measure called *normalized certainty penalty (NCP)* as a real value in $[0, 1]$ [GKKM07, Sec. 2.1].

As in Examples 1–3 given in the course, remove the identifiers (*name*, *first name*, and *email*) from the dataset. Then explore different parameter settings and choices for the quasi-identifiers (QI) and for the sensitive attribute (S). Make sure you understand categorical attributes and numerical attributes.

a) Using only *PLZ* and *points* as QI (and represented as numerical attributes), and with *system* as S, compute at least a 3-, 5-, and 10-anonymization of the dataset and report its NCP.

   What is the NCP of a 1-anonymization and that of a 74-anonymization?

b) Permute the dataset randomly (e.g., calling `shuf`) and observe the outcome. Extend the algorithm (using randomization) to compute improved 3-, 5-, and 10-anonymizations, that is, achieving better NCP than under a).

References

[GKKM07] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, *Fast data anonymization with low information loss*, Proceedings of the 33rd International Conference on Very Large Data Bases, University of Vienna, Austria, September 23-27, 2007 (C. Koch, J. Gehrke, M. N. Garofalakis, D. Srivastava, K. Aberer, A. Deshpande, D. Florescu, C. Y. Chan, V. Ganti, C. Kanne, W. Klas, and E. J. Neuhold, eds.), ACM, 2007, `http://www.vldb.org/conf/2007/papers/research/p758-ghinita.pdf`, pp. 758–769.

[LDR06] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, *Mondrian multidimensional k-anonymity*, Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA (L. Liu, A. Reuter, K. Whang, and J. Zhang, eds.), IEEE Computer Society, 2006, https://doi.org/10.1109/ICDE.2006.101.