The Effect of Homogeneity on the Complexity of Anonymizing Data

Robert Bredereck[1,*], André Nichterlein[1], Rolf Niedermeier[1], Geevarghese Philip[2]

[1]TU Berlin, Germany

{robert.bredereck,andre.nichterlein,rolf.niedermeier}@tu-berlin.de

[2]The Institute of Mathematical Sciences, Chennai, India

gphilip@imsc.res.in

Abstract

Motivated by data privacy applications [3], the NP-hard k-Anonymity problem asks, given an $n \times m$ matrix M over a fixed alphabet and an integer $s > 0$, whether M can be made k-anonymous by suppressing (blanking out) at most s entries. A matrix M is said to be k-anonymous if for each row r in M there are at least $k-1$ other rows in M which are identical to r. Complementing previous work [2, 1], we introduce two new "data-driven" parameterizations for k-Anonymity—the number t_{in} of different input rows and the number t_{out} of different output rows—for modeling aspects of data homogeneity. We show that k-Anonymity is fixed-parameter tractable for the parameter t_{in}, and it is NP-hard even for $t_{\text{out}} = 2$ and alphabet size four. Notably, our fixed-parameter tractability result implies that k-Anonymity can be solved in *linear time* when t_{in} is a constant. Our results also extend to some interesting generalizations of k-Anonymity such as ℓ-diversity and anonymization using the so-called "domain generalization hierarchies".

References

[1] P. Bonizzoni, G. D. Vedova, R. Dondi, and Y. Pirola. Parameterized complexity of k-anonymity: Hardness and tractability. In *Proc. 21st IWOCA*. Springer, to appear 2011.

[2] P. A. Evans, T. Wareham, and R. Chaytor. Fixed-parameter tractability of anonymizing data by suppressing entries. *J. Comb. Optim.*, 18(4):362–375, 2009.

[3] L. Sweeney. Uniqueness of simple demographics in the U.S. population. Technical report, Carnegie Mellon University, School of Computer Science, Laboratory for International Data Privacy, 2000.

*Supported by the DFG, research project PAWS, NI 369/10.