Barriers to the Implementation of k-Anonymity and Related Microdata Anonymization Techniques in a Real-World Application

Implementing anonymity in applications, which involve sensitive microdata, is a desirable goal. [3] In particular, publishing sensitive data is a critical task and requires measures to ensure privacy, even if joining attacks [1, 4] are performed. Treating the data with *k-anonymity* algorithms is one possible solution to this problem. [1,4] Given a dataset of k individuals, k-anonymity holds, if each record is identical to at least (k-1) records over so called quasi-identifiers in a release. [3] This way, even if the tables are joined with external data, it will not be possible, to re-identify the datasets, i.e. linking them back to one individual. [1]

However, the implementation of k-anonymity in real-world applications is not an easy task. Sweeney provides two systems, which have different shortcomings in their implementation of k-anonymity, namely the *Datafly* system and the μ -*Argus* system. [3]

One fundamental step of k-anonymization is to generalize the data, but only as much as necessary. [1] The Datafly system produces sufficient generalizations, but those generalizations are not guaranteed to be the desired k-minimal distortions. [3] Sweeney's analysis shows, that considering all requirements for a correct k-anonymity is not trivial.

In contrast, the μ -Argus system can fail completely to produce k-anonymity, because its algorithms don't consider all possible combinations of quasi-identifiers in order to save computing time. [3] This is a hint to the major problem of the implementation of optimal k-anonymity: Meyerson and Williams proved that the process of k-anonymization is NP-hard. [2] Accordingly, there is a trade-off in real-world applications, which have to process datasets with many quasi-identifiers, between computing time and the quality of the produced k-anonymity. Therefore approximations and heuristics are often used. [2]

References

1. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: Efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. pp. 49–60. SIGMOD '05, ACM, New York, NY, USA (2005), http://doi.acm.org/10.1145/1066157.1066164
2. Meyerson, A., Williams, R.: On the complexity of optimal k-anonymity. In: Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. pp. 223–228. PODS '04, ACM, New York, NY, USA (2004), http://doi.acm.org/10.1145/1055558.1055591
3. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10(5), 571–588 (Oct 2002), http://dx.doi.org/10.1142/S021848850200165X

4. Sweeney, L.: K-anonymity: A model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10(5), 557–570 (Oct 2002), http://dx.doi.org/10.1142/S0218488502001648