Managing Information Leakage
by Steven Euijong Whang and Hector Garcia-Molina

1 Summary

The paper presents two concepts: entity resolution (ER, record linkage) and data privacy (DP). Authors presented a sketch of a framework for managing information leakage, and studied how the framework can be used to answer a variety of questions related to ER and DP. In the paper they studied the problems of measuring the incremental leakage of critical information. The framework bases on definitions and usage of two functions – match and merge. The former function allows to detect attribute values, which describe the same entity, while the latter function merges such values into one record describing such entity. Calling these functions subsequently incrementally builds a set of data that are disclosed about described entity. Authors used disinformation as a mechanism to minimize information leakage. The paper presents a model of the problem, shows an idea of the framework, explains motivation of authors, and provides plenty of examples, but for any details refers to the technical report.

2 Strong Points

- The presented approach considers correlation different attribute values of the same record.
- The approach considers confidence of record identification.
- The approach uses disinformation as a method to decrease leakage. Disinformation is a type of noise, which is rarely used by other authors due to atomic definition of data records, i.e., data record is not partitioned, but treated as one entity.
- Authors defined their approach as a framework and presented examples of exercising it.
- The paper is easy to read and understand.

3 Weak Points

- The concept presented in the paper is similar to quasi-identifier (QID), but is not compared with it.
- The paper is not self contained, but it refers to the technical report frequently. Methods and theoretical details are moved to the report, but mentioned in this paper.
- An example matching function is defined using the “key set”, which in fact is a definition of the identifier attribute (ID). Such connection with the classical model of partitioning attributes as ID, QID, and Sensitive is not made.
- The leakage is measured based on revealed records, not changes in matching confidence.
- The paper does not present any performance results. The solution presented by authors is in the worst-case scenario NP-hard. Heuristics, and other methods proposed by authors and briefly characterized in the paper have computed complexities, but they are not confirmed in experiments.

4 Questions and Discussion Points

The matching function is binary, i.e., it returns 0 or 1. It would be more flexible to allow any value from the range (0, 1).

Using disinformation as a technique of adding noise, i.e., irrelevant information, which is incorrect, and limits the leakage is an interesting approach, which we could apply to other data privacy research problems. For example, to answer questions like – what is the impact of fake data (DOB set on social media portals) to protect privacy of the data owner? Would it help only against linkage attacks?

Paper Presentations – Preferences.

1. Summarization and Matching of Density-Based Clusters in Streaming Environments by D. Yang, E. A. Rundensteiner, and Matthew O. Ward

2. Sketch-based Querying of Distributed Sliding-Window Data Streams by O. Papapetrou, M. Garofalakis, and A. Deligiannakis