Blocking-Aware Private Record Linkage

Reviewed by Liyue Fan

This paper is an early study in private record linkage. Each record is composed of tokens, and record pairs are compared based on TFIDF weights of tokens contained in them. The security is protected by one-way hashing and secure communication channel. They also propose several blocking methods compatible with the TFIDF weight vector representation of the records.

Comments:
1. Matching model used for document retrieval highlights token-based similarity.
2. Proposed blocking methods are based on shared tokens but may introduce more than one block assignments for each record.
3. Security is protected by hashing and the use of a third party. The analysis of divulged information as well as the extent of disclosure is notable.

Discussions:
1. This paper assumes that each record contains a set of tokens (bag-of-words) which need to be tokenized first. This step adds to computation cost and also omits possible typos/errors within each token. The model is usually used to match documents, rather than short records. In the evaluation, the authors conducted studies with citation data. What would be a real scenario that concerns privacy?
2. Computation time turns out to be a concern. It is reported that using the most efficient blocking method would result in 100s computation time for matching two databases with 5000 records each. Without blocking at all, the computation time is 450s for pairwise TFIDF vector comparison on the same datasets. Is it possible to improve the scalability?
3. Most private information retrieval literature study the privacy of user/query issued rather than that of the underlying database. Can we calculate TFIDF vectors with differential privacy?

Preferences:
1. Efficient Privacy-Aware Record Integration, EDBT
2. Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage, Information Fusion