International Journal of Population Data Science

Journal Website: www.ijpds.org

Evaluation of advanced techniques for multi-party privacy-preserving record linkage on real-world health databases

Ranbaduge, Thilina¹, Vatsalan, Dinusha¹, Randall, Sean², and Christen, Peter^{1*}

¹Research School of Computer Science, The Australian National University

²Centre for Data Linkage, Curtin University

Objective

The linking of multiple (three or more) health databases is challenging because of the increasing sizes of databases, the number of parties among which they are to be linked, and privacy concerns related to the use of personal data such as names, addresses, or dates of birth. This entails a need to develop advanced scalable techniques for linking multiple databases while preserving the privacy of the individuals they contain. In this study we empirically evaluate several state-of-the-art multi-party privacy-preserving record linkage (MP-PPRL) techniques with large real-world health databases from Australia.

Approach

MP-PPRL is conducted such that no sensitive information is revealed about database records that can be used to infer knowledge about individuals or groups of individuals. Current state-of-the-art methods used in this evaluation use Bloom filters to encode personal identifying information. The empirical evaluation comprises of different multi-party private blocking and matching techniques that are evaluated for different numbers of parties. Each database contains more than 700,000 records extracted from ten years of New South Wales (NSW) emergency presentation data. Each technique is evaluated with regard to scalability, quality and privacy. Scalability and quality are measured using the metrics of reduction ratio, pairs completeness, precision, recall, and F-measure. Privacy is measured using disclosure risk metrics that are based on the probability of suspicion, defined as the likelihood that a record in an encoded database matches to one or more record(s) in a publicly available database such as a telephone directory. MP-PPRL techniques that either utilize a trusted linkage unit, and those that do not, are evaluated.

Results

Experimental results showed MP-PPRL methods are practical for linking large-scale real world data. Private blocking techniques achieved significantly higher privacy than standard hashing-based techniques with a maximum disclosure risk of 0.0003 and 1, respectively, at a small cost to linkage quality and efficiency. Similarly, private matching techniques provided a similar acceptable reduction in linkage quality compared to standard non-private matching while providing high privacy protection.

Conclusion

The adoption of privacy-preserving linkage methods has the ability to significantly reduce privacy risks associated with linking large health databases, and enable the data linkage community to offer operational linkage services not previously possible. The evaluation results show that these state-of-the-art MP-PPRL techniques are scalable in terms of database sizes and number of parties, while providing significantly improved privacy with an associated trade-off in linkage quality compared to standard linkage techniques.

*Corresponding Author:

Email Address: peter.christen@anu.edu.au (P. Christen)