International Journal of Population Data Science

Journal Website: www.ijpds.org

Partial Agreements in Probabilistic Linkages

Brown, A¹, Randall, S¹, Ferrante, A¹, and Boyd, J¹

¹Curtin University

Introduction

Record linkage units around the world use probabilistic linkage techniques for routine linkage of large datasets. It is widely known how probabilities are converted to agreement and disagreement weights for each field, yet there has been little exploration of the methodology to optimally convert field similarity scores into partial weights.

linkage where the ability for clerical review is limited. Optimised functions for converting similarities to partial weights can significantly improve the quality of linkage and should not be overlooked.

Objectives and Approach

String similarity comparators such as Jaro-Winkler are commonly used in traditional linkage, other comparators such as the Sorenson Dice coefficient, Jaccard similarity and Hamming distance are used in alternative privacy-preserving record linkage techniques. Determining partial weights to apply at each level of similarity is a non-trivial task. However, both types of linkages would greatly benefit from similarity to weight functions for each field that maximises the accuracy of the linkage.

We evaluated several methods for computing partial agreement weights and applied these to synthetic datasets with varying levels of corruption. We then evaluated the methods on real administrative datasets.

Results

Exact comparisons can miss matches where typographical errors or misspellings produce small changes in value. Similarity comparisons can reduce the number of missed matches, but may also increase the number of incorrect matches.

Various results of the partial agreement methods on Jaro-Winkler, Sorenson Dice coefficient, Jaccard similarity and Hamming distance comparators will be presented. A generic function to convert similarity values to weights, created from synthetic data, can be used on most datasets with a greatly improved result in linkage quality. However, maximising the linkage quality requires the creation of similarity-to-weight functions that are optimised for each dataset.

Conclusion/Implications

Accuracy in record linkage is vital for the correct analysis of linked data. It is even more critical in privacy-preserving record