International Journal of Population Data Science

Journal Website: www.ijpds.org

# Privacy preserving record linkage meets record linkage using unencrypted data

Izakian, H[1]

[1]PolicyWise for Children & Families

## Introduction

Privacy preserving record linkage (PPRL) resolves privacy concerns because of its capabilities to link encrypted identifiers. It encrypts identifiers using bloom filters and performs record matching based on encrypted data using dice coefficient similarity. Matching data based on hashed identifiers impacts the performance of linkage due to loss of information.

## Objectives and Approach

We propose a technique to optimize the bloom filter parameters and examine if the optimal parameters increase the performance of the linkage in terms of precision, recall, and f-measure. Let us consider a set of string values and calculate the similarity between any two of them using the Jaro-Winkler method. Now let us encrypt the string values using bloom filters and calculate the similarity between any two of them using the dice coefficient technique. Optimal parameters of bloom filters are those that minimize the difference between the calculated similarities using Jaro-Winkler vs. the calculated similarities using the dice coefficient technique.

## Results

Using publically available data, several first name and last name datasets each comprising 1000 unique values were generated. The following values for bloom filter parameters were considered: q in q-grams (q=1,2,3), bit array length (l=50,100,200,500,1000), number of hash functions (k=5,10,20,50). The following five setups of bloom filters were able to minimize the difference between the calculated similarities on encrypted data using the dice coefficient technique, and the calculated similarities on unencrypted data using the Jaro-Winkler method: q=1,l=1000,k=50/q=1,l=500,k=20/ q=2,l=1000,k=50/ q=3,l=500,k=50. These setups were considered to perform data linkage over 10 synthetically-generated datasets. Results show that PPRL was able to achieve similar performance compared to data linkage over unencrypted data.

## Conclusion/Implications

This study showed that optimal parameters of bloom filters minimized loss of information resulting from data encryption. Experimental findings indicated that PPRL using optimal parameters of bloom filters achieves almost the same performance as data linkage on unencrypted data in terms of precision, recall, and f-measure.

http://dx.doi.org/10.23889/ijpds.v3i4.651