International Journal of Population Data Science

Journal Website: www.ijpds.org

Secure Privacy Preserving Record Linkage of Large Databases by Modified Bloom Filter Encodings

Schnell, Rainer^{1*} and Borgs, Christian²

¹City University London

²German Record Linkage Center - University of Duisburg-Essen

Objective

In most European settings, record linkage across different institutions has to be based on personal identifiers such as names, birthday or place of birth. To protect the privacy of research subjects, the identifiers have to be encrypted. In practice, these identifiers show error rates up to 20% per identifier, therefore linking on encrypted identifiers usually implies the loss of large subsets of the databases. In many applications, this loss of cases is related to variables of interest for the subject matter of the study. Therefore, this kind of record-linkage will generate biased estimates.

These problems gave rise to techniques of Privacy Preserving Record Linkage (PPRL). Many different PPRL techniques have been suggested within the last 10 years, very few of them are suitable for practical applications with large database containing millions of records as they are typical for administrative or medical databases. One proven technique for PPRL for large scale applications is PPRL based on Bloom filters.

all known attacks on PPRL Bloom filters. These computationally simple algorithms modify the identifiers by different cryptographic diffusion techniques. The presentation will demonstrate these new algorithms and show their performance concerning precision, recall and re-identification risk on large databases.

Method

Using appropriate parameter settings, Bloom filter approaches show linkage results comparable to linkage based on unencrypted identifiers. Furthermore, this approach has been used in real-world settings with data sets containing up to 100 Million records. By the application of suitable blocking strategies, linking can be done in reasonable time.

Result

However, Bloom filters have been subject of cryptographic attacks. Previous research has shown that the straight application of Bloom filters has a nonzero re-identification risk. We will present new results on recently developed techniques to defy

*Corresponding Author:

Email Address: rainer.schnell@uni-due.de (R. Schnell)