International Journal of Population Data Science

Journal Website: www.ijpds.org

# Scalable Secure Privacy-Preserving Record Linkage (PPRL) Methods Using Cloud-based Infrastructure

Ong, T[1], Lazrig, I[2], Ray, I[2], Ray, I[1], and Kahn, M[1]

[1]University of Colorado Anschutz Medical Campus

[2]Colorado State University

## Introduction

Bloom Filters (BFs) are a scalable solution for probabilistic privacy-preserving record linkage but BFs can be compromised. Yao's garbled circuits (GCs) can perform secure multi-party computation to compute the similarity of two BFs without a trusted third party. The major drawback of using BFs and GCs together is poor efficiency.

## Objectives and Approach

We evaluated the feasibility of BFs+GCs using high capacity compute engines and implementing a novel parallel processing framework in Google Cloud Compute Engines (GCCE). In the Yao's two-party secure computation protocol, one party serves as the generator and the other party serves as the evaluator. To link data in parallel, records from both parties are divided into chunks. Linkage between every two chunks in the same block is processed by a thread. The number of threads for linkage depends on available computing resources. We tested the parallelized process in various scenarios with variations in hardware and software configurations.

## Results

Two synthetic datasets with 10K records were linked using BFs+GCs on 12 different software and hardware configurations which varied by: number of CPU cores (4 to 32), memory size (15GB – 28.8GB), number of threads (6-41), and chunk size (50-200 records). The minimum configuration (4 cores; 15GB memory) took 8,062.4s to complete whereas the maximum configuration (32 cores; 28.8GB memory) took 1,454.1s. Increasing the number of threads or changing the chunk size without providing more CPU cores and memory did not improve the efficiency. Efficiency is improved on average by 39.81% when the number of cores and memory on the both sides are doubled. The CPU utilization is maximized (near 100% on both sides) when the computing power of the generator is double the evaluator.

## Conclusion/Implications

The PPRL runtime of BFs+GCs was greatly improved using parallel processing in a cloud-based infrastructure. A cluster of GCCEs could be leveraged to reduce the runtime of data linkage operations even further. Scalable cloud-based infrastructures can overcome the trade-off between security and efficiency, allowing computationally complex methods to be implemented.

http://dx.doi.org/10.23889/ijpds.v3i4.638