Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage

Reviewed by Liyue Fan

1 Summary

This paper provides an overview of recently proposed privacy-preserving string comparators (PPSCs), including Exact Matching with SHA-1 [9], Bloom Filters [12], Trigrams [7], Embedding [11], secure Edit Similarity [1], and Phonetic Filter [8]. While two parties hold private databases, a third party is needed to perform the matching. Once individual field has been compared, Winkler modification for continuous Fellegi-Sunter [10] is used to produce a similarity score for each record pair. For each method, $N$ top-ranked record pairs out of $N \times N$ candidates are emitted as matches, where $N$ is the size of one party database.

The above methods are evaluated with respect to correctness, computation time, and security. The correctness is measured by TPR (True Positive Rate). Computation time is measured from the program running time. Security is measured by mutual information entropy since the third party only sees the encoded data without any knowledge of the protocol.

The dataset used for evaluation is North Carolina voter registration (NCVR) files, which is NOT available anymore at the provided link\(^1\). 1000 records are selected for each database, and a “data corrupter” based on [2] is used to generated the database to be matched. 100 iterations of such experiments have been conducted. According to their results, Embedding has lowest TPR but provides high security and Edit Similarity incurs infeasible computation time while providing highest security. Several methods are excluded from the survey, such as [3, 13, 4], due to reported poor performance.

A few other surveys are mentioned in the paper [5, 14, 6, 15].

2 Discussion

Question: can we go farther than where this paper has gone?

- No blocking step is considered for any proposed method;

\(^1\)http://www.app.sboe.state.nc.us/data
• This is not fair comparison for Embedding method due to its contractive property. How to design the evaluation?

• How to perform security analysis when only two parties are involved? (They see the encoded records and know the protocol)

• What are methods to compare in our survey, considering some have been reported to perform poorly?

• What is the scope of our comparison: full linkage process, or mapping only?

References


