Efficient Privacy-Aware Record Integration

Reviewed by Liyue Fan

1 Summary

This paper studies the problem of private record linkage (PRL). The authors proposed a three-step framework: 1) a third party generates the blocks using publicly available data, such as voter’s registration list; 2) both data holders partition their data sets into blocks and party A shares the blocks with encrypted records with party B, while the block size is protected by differential privacy; 3) party B performs linking with SMC and shared with party A the matched records.

2 Strengths

1. The third party could derive the blocks with publicly available data, which doesn’t incur any information leakage on either party’s data set.

2. To protect party A’s block size, they propose to perturb it by Laplace mechanism with positive distribution mean. This helps reduce the number of suppressed/removed records due to negative perturbation noise, thus reducing the loss of matched pairs.

3. The SMC protocol for string attributes is explicitly defined in this paper. String attributes are transformed into a bit vector with respect to bigrams and then encrypted.

3 Weaknesses

1. Generating blocks with publicly available data is arguable. First of all, the data sets to link may not have the same domain or distribution as the available, public data set. Second of all, the blocking quality cannot be guaranteed without considering the data set’s distribution.

2. The idea of positive mean Laplace distribution is helpful in reducing record loss. On the other hand, it also introduces many excessive computations for [fake, original] pairs due to the larger, positive perturbation noise. In experiment Section6.2, each data holder has 5000 records partitioned into 500 or more blocks. With 10 records per block on average, the number of fake records introduced for differential privacy could easily dominate the pairwise computation in SMC protocol.
3. Overall, there’s not much technical innovation compared with their previous work [1]. For the SMC step, the conversion of strings to bit vectors is based on bloom filters, which is the state-of-the-art for PRL proposed by [2].

4 Discussion

I think this is a system/application paper considering the completeness of the linkage protocol and the lack of new research contributions.

References
