Privacy-Preserving Temporal Record Linkage

Thilina Ranbaduge and Peter Christen

thilina.ranbaduge@anu.edu.au and peter.christen@anu.edu.au

Research School of Computer Science, College of Engineering & Computer Science, The Australian National University, Canberra

What is Temporal Record Linkage?

- In data integration different databases often need to be linked to identify the sets of matching records that refer to the same entity
- **Temporal record linkage** matches records in databases while considering temporal information in those records

Record ID	Entity ID	First name	Last name	Street address	Timestamp
r_1	e_1	Anne	Miller	161 Main Road, Sydney	2002-09-11
r_2	e_2	Anne	Smith	43 Town Place, Sydney	2005-05-23
r_3	e_2	Ann	Miller	43 Town Place, Sydney	2006-11-05
r_4	e_1	Anne	Smith	23 Town Place, Sydney	2007-04-10
r_5	e_2	Ann	Miller	12 Main Road, Sydney	2007-12-21
r_6	e_3	Anne	Miller	12 Main Road, Sydney	2010-02-11

Problem and Challenges

Problem

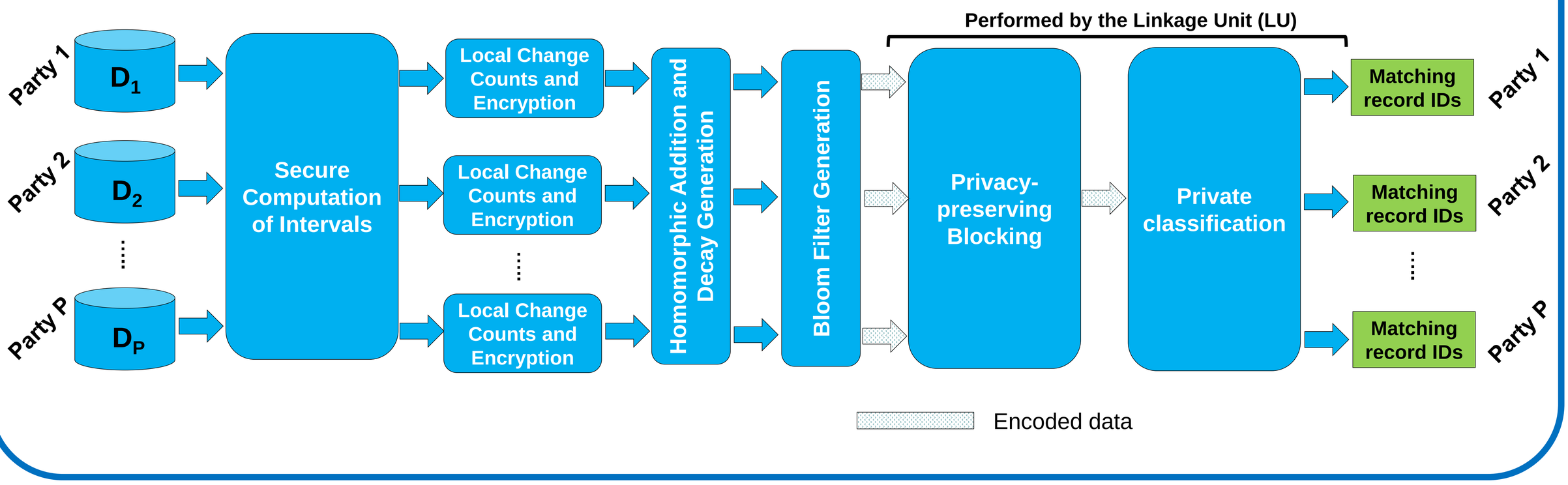
- Linkage is often based on personal identifiers of entities
- Due to **privacy and confidentiality concerns**, organizations are often not willing or allowed to share or reveal their sensitive data

Challenges

- Secure use of temporal information in records for similarity computations
- Secure comparison of records without revealing any attribute values
- Scale to large databases and increasing number of database owners

Privacy-preserving Temporal Record Linkage

- We propose a novel linkage protocol that performs temporal linkage across different databases while preserving privacy
- A **homomorphic encryption based secure multiparty computation** protocol is used to learn **the decay probabilities** that entities change their attribute values over a period of time
- Similarities between record pairs are then adjusted according to the attribute decay values



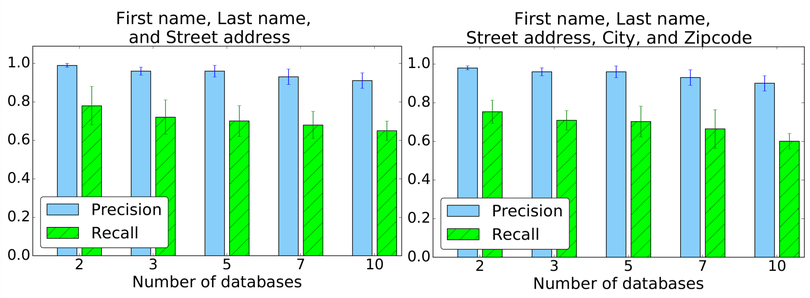
Discussion and Future Work

- 25 North Carolina Voter Registration (NCVR) datasets each with more than 7 million records
- Experiments with different number of parties and attribute combinations, including *First name, Last name, Street, City,* and *Zip code*

Runtime

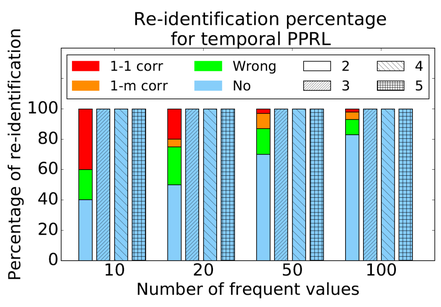
Number of records	Privacy-preserving temporal record linkage		
	Decay generation	Bloom filter generation	Comparison
100,000	4,018.2 sec	201.4 sec	595.2 sec
500,000	4,053.7 sec	998.7 sec	2,899.7 sec
1,000,000	4,105.8 sec	2,156.2 sec	5,819.4 sec
5,000,000	4,567.3 sec	9,876.8 sec	25,923.5 sec

Linkage Quality



- 8% to 20% average improvement in precision with a slight reduction in recall compared to the linkage process that does not use any temporal information

Privacy



Frequency based attack (Christen et al., 2017)

Measures percentage of correct

- 1-1 matches
- 1-m matches
- Wrong matches
- No matches

Conclusion

- Secure protocol for temporal record linkage that uses:
  - Secure multiparty computation for learning decay probabilities across parties
  - Masking of Bloom filter encodings based on temporal decay probabilities
- Our approach is scalable to large number of parties and database sizes while providing privacy for individuals encoded in the Bloom filters

Future Works

- Privacy-preserving active learning for training data generation
- Secure techniques for encoding attribute values
- Secure protocols for performing linkage with a third party
- Secure protocols that incorporate temporal information in the blocking process
- Secure linkage protocol for dynamic data

This research is funded by the Australian Research Council under Discovery Project DP130101801 and DP160101934.

Australian National University

ANU College of Engineering & Computer Science