Advanced Techniques for Privacy-Preserving Linking of Multiple Large Databases

Dinusha Vatsalan

Work done in collaboration with:
Prof. Peter Christen, Prof. Vassilios S. Verykios, Dr. Christine M. O’Keefe, Prof. Erhard Rahm, and Thilina Ranbaduge

Funded by the ARC Discovery Projects DP130101801 and DP160101934, and Universities Australia and the German Academic Exchange Service (DAAD)

Research School of Computer Science, College of Engineering and Computer Science, The Australian National University, Australia
Multi-party privacy-preserving record linkage (MP-PPRL)

MP-PPRL techniques
- Bloom filter-based approximate matching for MP-PPRL (AM-BF)
- Counting Bloom filter-based matching for MP-PPRL (AM-CBF)
  - Improved communication patterns
  - Incremental clustering-based subset matching for MP-PPRL (AM-Clus)

Bloom filter-based data masking for different data types

Outlook to future research directions
# Privacy-Preserving Record Linkage (PPRL) – An Example

## Health database

<table>
<thead>
<tr>
<th>PID</th>
<th>Surname</th>
<th>Given_name</th>
<th>Age</th>
<th>Postcode</th>
<th>Sex</th>
<th>Pressure</th>
<th>Stress</th>
<th>Last_visited</th>
<th>Reason_of_visit</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1209</td>
<td>Robertt</td>
<td>Peter</td>
<td>41</td>
<td>2617</td>
<td>m</td>
<td>140/90</td>
<td>high</td>
<td>25 days ago</td>
<td>chest pain</td>
</tr>
<tr>
<td>P4204</td>
<td>Miller</td>
<td>Amelia</td>
<td>39</td>
<td>2415</td>
<td>f</td>
<td>120/80</td>
<td>high</td>
<td>61 days ago</td>
<td>headache</td>
</tr>
<tr>
<td>P4894</td>
<td>Siemen</td>
<td>Jeff</td>
<td>30</td>
<td>2602</td>
<td>m</td>
<td>110/80</td>
<td>normal</td>
<td>15 days ago</td>
<td>checkup</td>
</tr>
</tbody>
</table>

## Social security database

<table>
<thead>
<tr>
<th>ID</th>
<th>First_name</th>
<th>Last_name</th>
<th>DOB</th>
<th>Gender</th>
<th>Postcode</th>
<th>Loan_type</th>
<th>Period</th>
<th>Amount</th>
<th>Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>6723</td>
<td>Peter</td>
<td>Robert</td>
<td>20.06.1972</td>
<td>M</td>
<td>2617</td>
<td>Mortgage</td>
<td>20</td>
<td>350,000</td>
<td>130,000</td>
</tr>
<tr>
<td>8345</td>
<td>Miller</td>
<td>Roberts</td>
<td>11.10.1979</td>
<td>M</td>
<td>2602</td>
<td>Personal</td>
<td>5</td>
<td>10,000</td>
<td>1,900</td>
</tr>
<tr>
<td>9241</td>
<td>Amelia</td>
<td>Millar</td>
<td>06.01.1974</td>
<td>F</td>
<td>2415</td>
<td>Mortgage</td>
<td>30</td>
<td>475,000</td>
<td>154,250</td>
</tr>
</tbody>
</table>

## Bank database

<table>
<thead>
<tr>
<th>SSN</th>
<th>Title</th>
<th>Last_name</th>
<th>First_name</th>
<th>Age</th>
<th>Postcode</th>
<th>Employment</th>
<th>Income</th>
<th>Benefits</th>
<th>Payment</th>
</tr>
</thead>
<tbody>
<tr>
<td>490814</td>
<td>Mrs</td>
<td>Amilia</td>
<td>Smith</td>
<td>39</td>
<td>2642</td>
<td>Teacher</td>
<td>60,000</td>
<td>Child care</td>
<td>45,000</td>
</tr>
<tr>
<td>581233</td>
<td>Mr</td>
<td>Peter</td>
<td>Roberts</td>
<td>42</td>
<td>2627</td>
<td>Engineer</td>
<td>110,000</td>
<td>Family tax</td>
<td>50,000</td>
</tr>
<tr>
<td>932389</td>
<td>Mr</td>
<td>William</td>
<td>Smith</td>
<td>69</td>
<td>3205</td>
<td>Retired</td>
<td>-</td>
<td>Pension</td>
<td>35,000</td>
</tr>
</tbody>
</table>
Three-party protocols

- Use a linkage unit (LU) to conduct or facilitate linkage

Two-party protocols

- Only the two database owners participate in the linkage

Multi-party protocols

- Linking records from multiple databases (with or without a LU)
- Additional challenge of scalability and privacy (due to collusions)
Multi-Party PPRL Techniques

- Private blocking techniques
  - A family of Bloom filter-based private blocking techniques developed recently for multi-party PPRL to reduce complexity (Ranbaduge et al. 2014, 2015, 2016)

- Private matching and classification techniques
  - All existing techniques for multi-party PPRL either support exact matching only or are applicable to categorical data only (E.g. Lai et al. 2006 proposed Bloom filter-based exact matching technique – EM-BF)

- Our techniques
  - Bloom filter-based approximate matching – AM-BF \(O(n^p)\) complexity
  - Counting Bloom filter-based approximate matching – AM-CBF \(O(n^r)\) complexity
  - Clustering-based subset matching – AM-Clus \(O(n^2)\) complexity

\(P\) databases each containing \(n\) records; \(r\) is the ring size for AM-CBF (with \(r < p\))
The similarity calculation can be distributed among $p$ parties:

$$\text{Dice}_{\text{sim}}(b_1, ..., b_p) = \frac{p \times \sum c_i}{\sum_i x_i}$$

<table>
<thead>
<tr>
<th></th>
<th>P₁</th>
<th>P₂</th>
<th>P₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_1$</td>
<td>1 1 0</td>
<td>1 1 0 0</td>
<td>0 1 1 1</td>
</tr>
<tr>
<td>$b_2$</td>
<td>1 1 0 0</td>
<td>1 1 0 0</td>
<td>0 0 1</td>
</tr>
<tr>
<td>$b_3$</td>
<td>1 0 0 0</td>
<td>1 1 1 0</td>
<td>0 0 1</td>
</tr>
<tr>
<td>$\sum$</td>
<td>1 0 0 0</td>
<td>1 1 0 0</td>
<td>0 0 1</td>
</tr>
</tbody>
</table>

- $c_1 = 1$
- $c_2 = 2$
- $c_3 = 1$

$x_1 = 6$
$x_2 = 5$
$x_3 = 5$

secure summation = 16

$\text{sim} = \frac{3 \times 4}{16} = 0.75$
Multi-Party Counting Bloom Filter-based Approximate Matching (AM-CBF)

- An integer array of length \( l \) containing counts of values in each bit position \( \beta \), \( 1 \leq \beta \leq l \) over \( p \) Bloom filters.

- The similarity of \( p \) Bloom filters can be calculated given a counting Bloom filter \( cbf \):

\[
\text{Dice}_\text{sim}(b_1, \ldots, b_p) = \frac{p \times |\beta: 1 \leq \beta \leq l \text{ and } cbf(\beta) = p|}{\sum_{\beta=1}^{l} cbf(\beta)}
\]

\[
\text{sim}(b_1, b_2, b_3) = \frac{3 \times |1, 4, 5, 9|}{(3+2+0+3+3+1+0+1+3)}
\]

\[
= \frac{3 \times 4}{16} = 0.75
\]
AM-CBF (contd..)

Counting Bloom filters

Party $p_1$
\[
\begin{array}{cccccc}
1 & 1 & 0 & 1 & 1 & 0 \\
3 & 1 & 0 & 4 & 2 & 1 \\
4 & 2 & 0 & 5 & 3 & 1 \\
\end{array}
\]

Party $p_2$
\[
\begin{array}{cccccc}
1 & 1 & 0 & 1 & 1 & 0 \\
4 & 2 & 0 & 5 & 3 & 1 \\
5 & 3 & 0 & 6 & 4 & 1 \\
\end{array}
\]

Linkage Unit
\[
\begin{array}{cccccc}
3 & 1 & 0 & 4 & 2 & 1 \\
5 & 3 & 0 & 6 & 4 & 1 \\
2 & 2 & 0 & 2 & 0 & 0 \\
\end{array}
\]

sim = $2 \times \frac{5}{11} = 0.9$
The naïve (NAI) comparison in multi-party linkage is exponential in $p$ and size of datasets even with a blocking technique.

<table>
<thead>
<tr>
<th>Dataset / block size</th>
<th>p=3</th>
<th>p=5</th>
<th>p=7</th>
<th>p=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000 / 10</td>
<td>$10^6$</td>
<td>$10^8$</td>
<td>$10^{10}$</td>
<td>$10^{13}$</td>
</tr>
<tr>
<td>10,000 / 100</td>
<td>$10^8$</td>
<td>$10^{12}$</td>
<td>$10^{16}$</td>
<td>$10^{22}$</td>
</tr>
<tr>
<td>10,000 / 1,000</td>
<td>$10^{10}$</td>
<td>$10^{16}$</td>
<td>$10^{22}$</td>
<td>$10^{31}$</td>
</tr>
<tr>
<td>100,000 / 10</td>
<td>$10^7$</td>
<td>$10^9$</td>
<td>$10^{11}$</td>
<td>$10^{14}$</td>
</tr>
<tr>
<td>100,000 / 100</td>
<td>$10^9$</td>
<td>$10^{13}$</td>
<td>$10^{17}$</td>
<td>$10^{23}$</td>
</tr>
<tr>
<td>100,000 / 1,000</td>
<td>$10^{11}$</td>
<td>$10^{17}$</td>
<td>$10^{23}$</td>
<td>$10^{32}$</td>
</tr>
</tbody>
</table>

However, most comparisons are between true non-matches (class imbalance problem) and a true matching set must match between any subset of parties.
Sequential Communication using a LU

Ring 1

- \( p_1 \) matches ring 1
- \( p_2 \)

Ring 2

- \( p_3 \) matches rings 1,2
- \( p_4 \)

Ring 3

- \( p_5 \) matches rings 1,2,3
- \( p_6 \)

Ring 4

- \( p_7 \)
- \( p_8 \) matches

Linkage Unit (LU)
Ring by Ring Communication without using a LU

Phase 1

Ring 1
- $p_1$ 1 $p_2$
- $p_2$ 2 $p_3$
- $p_3$ 3 $p_1$

Ring 2
- $p_4$ 1a $p_5$
- $p_5$ 2a $p_6$
- $p_6$ 3a $p_4$

Ring 3
- $p_7$ 1b $p_6$
- $p_6$ 2b $p_8$
- $p_8$ 3b $p_7$

Phase 2

matches ring 1 $\rightarrow$ matches ring 2 $\rightarrow$ matches ring 3

matches
Datasets: North Carolina Voters Registration datasets
- \( p = [3,5,7,10] \) each containing up to 1 million records
- \( p = 26 \) each containing up to 5 million records
- Non-corrupted and corrupted (using GeCo) datasets

Prototypes implemented in Python 2.7

Scalability measures: Runtime
Linkage quality measures: F-score (harmonic mean of precision and recall, where precision is true matches/matches and recall = true matches/ all true matches
Privacy measures: disclosure risk under the worst case assumption (Vatsalan et al. 2014)
Experimental Evaluation (AM-CBF)

Scalability

The number of candidate sets for linkage

<table>
<thead>
<tr>
<th>Dataset size (in thousands)</th>
<th>5K</th>
<th>10K</th>
<th>50K</th>
<th>100K</th>
<th>500K</th>
<th>1,000K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of comparisons</td>
<td>$10^3$</td>
<td>$10^4$</td>
<td>$10^5$</td>
<td>$10^6$</td>
<td>$10^7$</td>
<td>$10^8$</td>
</tr>
</tbody>
</table>

Privacy

Comparison of disclosure risk

<table>
<thead>
<tr>
<th>Number of parties (p)</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disclosure risk ($DR_{Mean}$)</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Disclosure risk ($DR_{Mark}$)</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Legend:
- No-mod, CBF, $DR_{Mean}$
- Mod, CBF, $DR_{Mean}$
- No-mod, CBF, $DR_{Mark}$
- Mod, CBF, $DR_{Mark}$
- No-mod, BF, $DR_{Mean}$
- Mod, BF, $DR_{Mean}$
- No-mod, BF, $DR_{Mark}$
- Mod, BF, $DR_{Mark}$
Large-scale Sub-set Matching (AM-Clus)

\[ D_1^M \quad D_2^M \quad \text{Iteration 1} \quad D_3^M \quad \text{Iteration 2} \quad D_4^M \quad \text{Iteration 3} \]

- \( r_{1,1} \) to \( r_{2,1} \) with weight 0.75
- \( r_{1,2} \) to \( r_{2,2} \) with weight 0.95
- \( r_{1,3} \) to \( r_{2,3} \) with weight 0.8
- \( r_{1,4} \) to \( r_{2,1} \) with weight 0.9

Mapping & merging:
- \( D_1^M \) to \( D_2^M \)
- \( D_2^M \) to \( D_1^M \)
- \( D_1^M \) to \( D_3^M \)
- \( D_3^M \) to \( D_1^M \)
- \( D_1^M \) to \( D_4^M \)
- \( D_4^M \) to \( D_1^M \)
- \( D_1^M \) to \( D_2^M \)
- \( D_2^M \) to \( D_1^M \)

Weights:
- 0.75
- 0.9
- 0.85
- 0.9
- 0.8
- 0.9
- 0.75
- 0.85
- 0.75
- 0.9
Scalability

Total runtime required of linkage for different number of parties

Results for linking large datasets

F-measure for linking 26 NCVR datasets

- Precision
- Recall
- Fscore
Comparative Evaluation

**Runtime comparison**

Comparison of runtime for different number of parties

**Linkage quality comparison**

Comparison of F-measure of record linkage for different number of parties
Numerical Bloom Filter

\[
\begin{align*}
v_2 &= 26 \\ L_2 &= [24, 25, 26, 27, 28] \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1
\end{align*}
\]

\[
\begin{align*}
v_1 &= 25 \\ L_1 &= [23, 24, 25, 26, 27] \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1
\end{align*}
\]

\[
\begin{align*}
v_3 &= 27 \\ L_3 &= [25, 26, 27, 28, 29] \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1
\end{align*}
\]

\[
\begin{align*}
sim_M(v_1, v_2) &= \frac{2 \times 7}{(8 + 9)} \\
&= 0.82
\end{align*}
\]

\[
\begin{align*}
sim_M(v_1, v_3) &= \frac{2 \times 6}{(8 + 9)} \\
&= 0.71
\end{align*}
\]
Numerical Bloom Filter (contd..)
Conclusions and Future Work

PPRL research gaps

Privacy Aspects
- Efficient two-party PPRL
- Efficient masking
- Multi-party PPRL
- Other adversary models

Linkage Techniques
- Efficient private blocking
- Real-time linkage
- Different data types
- Advanced classification

Theoretical Analysis
- Theoretical privacy assessment
- Theoretical analysis for Big Data

Evaluation
- Evaluation framework
- Privacy measures
- Comprehensive evaluation
- Clerical review in PPRL

Practical Aspects
- Realistic datasets
- A language for PPRL
- Use case scenarios and applications
- Big Data frameworks

- Contributed
- Partially contributed
- Future work
Thank You!