ENCODING OF NUMERICAL DATA FOR PRIVACY-PRESERVING RECORD LINKAGE

Lea Demelius
Overview

- Introduction
  - Privacy-Preserving Record Linkage
  - Bloom filters and Cryptographic Longterm Keys
  - Why numerical encoding?
- Methods
  - Experimental setup
  - Numerical encoding method
- Results
Privacy-preserving Record Linkage
Bloom filters

\[
\text{Dice coefficient} = \frac{q \cdot h}{a+b} = \frac{2 \cdot 6}{7+7} = 0.857
\]
Cryptographic Longterm Keys (CLKs)

= record-level Bloom filters

EUPID
(European Patient Identity Service)
An additional identifier: **Place of birth**

Geocoordinates:

- independent of language
- stable
- allow distance measurements
Methods

• Python
• synthetic data (3000 records)
• comparison of 3 methods:
  • string encoding
  • numerical encoding
  • string encoding with shortened geocoordinates
Numerical encoding
Numerical encoding

∀i ∈ [0, 2b]:

\[ x_i = g + (i - b) \cdot d_{intv} \]

- **g**: geocoordinate
- **d_{intv}**: interval width
- **2b + 1**: length of list
Numerical encoding

45.2

[43.2 43.7 44.2 44.7 45.2 45.7 46.2 46.7 47.2]

45.1

[43.1 43.6 44.1 44.6 45.1 45.6 46.1 46.6 47.1]
Numerical encoding

\[
\text{new } x_i = \begin{cases} 
  x_i, & x_i \mod d_{\text{intv}} = 0 \\
  x_i - (x_i \mod d_{\text{intv}}), & x_i \mod d_{\text{intv}} < \frac{d_{\text{intv}}}{2} \\
  x_i + (d_{\text{intv}} - (x_i \mod d_{\text{intv}}), & x_i \mod d_{\text{intv}} \geq \frac{d_{\text{intv}}}{2}
\end{cases}
\]
Numerical encoding

longitude

latitude

+180° -180°

85° 90° 85°
Numerical encoding

\( f_{\text{final}}_i = \begin{cases} 
\text{min} + \text{new}_x_i \mod \text{max}, & \text{new}_x_i > \text{max} \\
\text{new}_x_i \mod \text{max}, & \text{new}_x_i < \text{min} \\
\text{new}_x_i, & \text{else}
\end{cases} \)

\( f_{\text{final}}_i = \begin{cases} 
2 \cdot \text{max} - \text{new}_x_i, & \text{new}_x_i > \text{max} \\
2 \cdot \text{min} - \text{new}_x_i, & \text{new}_x_i < \text{min} \\
\text{new}_x_i, & \text{else}
\end{cases} \)
### Results

<table>
<thead>
<tr>
<th>Test run No.</th>
<th>Test run description</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>string encoding</td>
<td>0.863</td>
<td>0.977</td>
<td>0.917</td>
</tr>
<tr>
<td>2</td>
<td>numerical encoding</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td>3</td>
<td>string encoding with shortened geocoordinates</td>
<td>0.960</td>
<td>0.990</td>
<td>0.975</td>
</tr>
</tbody>
</table>
Results

Numerical Encoding using CLKs leads to:

• better quality
• more privacy
ENCODING OF NUMERICAL DATA FOR PRIVACY-PRESERVING RECORD LINKAGE

Lea Demelius