Record linkage approaches
in pSCANNER

Toan Ong, PhD
Assistant Professor
Department of Pediatrics
University of Colorado, Anschutz Medical Campus
Outline

• Problem

• Challenges

• Record linkage solutions

• Looking ahead
Problem

• Data for analysis are distributed across different institutions
• Horizontally partitioned data

[Diagram: Depiction of data partitioning with de-duplication and enrichment processes]

• Vertically partitioned data
Example

• John is a severe chronic asthma patient who received care at both health institution A, B, and C in Colorado

• Mary is a mild asthma patient who received care at only at A

• What is the prevalence of severe asthma among patient with asthma?

prevalence = John + John + John / (3 Johns + Mary) = 75%

Instead of 50%
Definition

• Record linkage: The process of linking records that represent the same entity in one or more databases

  ❖ Objective:
  • Data completeness
  • Data de-duplication

• Privacy-preserving record linkage (PPRL): record linkage without revealing clear-text linkage data using data encryption
Challenges

• A universally shared identifier does not exist
• Clear-text linkage variables (SSN, first and last name, DOB...) are HIPAA-protected information
• Linkage data have errors (e.g., typographical errors)
• Attack to decrypt hashed data
• Lack of gold-standard linked data to test record linkage methods
• Difficult to perform linkage verification
Linkage variables

- Social security number
- First name
- Last name
- Date of birth

1. Seeded HashID of (First Name + Last Name + Date of Birth),
2. Seeded HashID of (Date of Birth + SSN),
3. Seeded HashID of (Last Name + SSN), or
4. Seeded HashID of (Three Letter First Name + Three Letter Last Name + Soundex First Name + Soundex Last Name + Date of Birth + SSN).

Abel et al., 2015
Record linkage approach

Data partner 1 ➔ Trusted third party (TTP) ➔ Data partner 2 ➔ Data partner 3 ➔ Data partner 4

Hashed data
Record linkage approach

[Diagram showing interactions between Data partner 1, 2, 3, and 4 with an annotation for Garbled circuit]
Record linkage methods

• Deterministic:
  • A linkage is determined by exact matching of hash value
  ⇒ intolerant to errors in linkage data

1. Seeded HashID of (First Name + Last Name + Date of Birth),
2. Seeded HashID of (Date of Birth + SSN),
3. Seeded HashID of (Last Name + SSN), or
4. Seeded HashID of (Three Letter First Name + Three Letter Last Name + Soundex First Name + Soundex Last Name + Date of Birth + SSN).

Abel et al., 2015
Record linkage methods

• Probabilistic PPRL:

Bloom filters

Similarity score

Dice coefficient = \( \frac{2|X \cap Y|}{|X| + |Y|} \)

Johnson et al., 2010
Probabilistic

- Effective to link data with errors
- Compatible with both TTP or pair-wise approach
- Efficiency can be improved by effective data blocking strategies
Examples of data errors

- Findings from verifying real data:
  - Typos in the value of the linkage variables
  - Nick name
  - Middle name
  - Maiden name included in last name (two-word names)
  - Prefixes and suffixes
Linkage performance (synthetic data)

• Synthetic datasets:
  • 10K records each
  • Corrupted data
  • 6K overlapping records

<table>
<thead>
<tr>
<th>Approach</th>
<th>Method</th>
<th># TP</th>
<th># FP</th>
<th># FN</th>
<th>Run time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTP</td>
<td>Deterministic</td>
<td>4,607</td>
<td>0</td>
<td>1,393</td>
<td>47</td>
</tr>
<tr>
<td>TTP</td>
<td>Probabilistic</td>
<td>5,757</td>
<td>1</td>
<td>243</td>
<td>1,038</td>
</tr>
<tr>
<td>Pairwise (GC)</td>
<td>Deterministic</td>
<td>4,067</td>
<td>0</td>
<td>1,393</td>
<td>13,647</td>
</tr>
<tr>
<td>Pairwise (GC)</td>
<td>Probabilistic</td>
<td>5,948</td>
<td>16</td>
<td>52</td>
<td>30,285</td>
</tr>
</tbody>
</table>
Linkage performance (synthetic data)

- Probabilistic pair-wise

<table>
<thead>
<tr>
<th>Blocking variable</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>Run time</th>
</tr>
</thead>
<tbody>
<tr>
<td>LN</td>
<td>4643</td>
<td>15</td>
<td>1357</td>
<td>7,978</td>
</tr>
<tr>
<td>YOB</td>
<td>4842</td>
<td>6</td>
<td>1158</td>
<td>16,275</td>
</tr>
<tr>
<td>MOB+DOB</td>
<td>5407</td>
<td>3</td>
<td>593</td>
<td>6,030</td>
</tr>
<tr>
<td>Combined</td>
<td>5948</td>
<td>16</td>
<td>52</td>
<td>30,285</td>
</tr>
</tbody>
</table>
### Testing on real data

<table>
<thead>
<tr>
<th>Site ID</th>
<th>Patients</th>
<th>Site ID</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
<th>Value 5</th>
<th>Value 6</th>
<th>Value 7</th>
<th>Value 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22639</td>
<td>1</td>
<td>NULL</td>
<td>29</td>
<td>25</td>
<td>27</td>
<td>12</td>
<td>10</td>
<td>12</td>
<td>436</td>
</tr>
<tr>
<td>2</td>
<td>111743</td>
<td>2</td>
<td>29</td>
<td>NULL</td>
<td>754</td>
<td>85</td>
<td>50</td>
<td>36</td>
<td>64</td>
<td>93</td>
</tr>
<tr>
<td>3</td>
<td>75167</td>
<td>3</td>
<td>25</td>
<td>754</td>
<td>NULL</td>
<td>74</td>
<td>43</td>
<td>17</td>
<td>47</td>
<td>65</td>
</tr>
<tr>
<td>4</td>
<td>95217</td>
<td>4</td>
<td>27</td>
<td>85</td>
<td>74</td>
<td>NULL</td>
<td>898</td>
<td>186</td>
<td>641</td>
<td>185</td>
</tr>
<tr>
<td>5</td>
<td>86514</td>
<td>5</td>
<td>12</td>
<td>50</td>
<td>898</td>
<td>NULL</td>
<td>901</td>
<td>264</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>78555</td>
<td>6</td>
<td>10</td>
<td>36</td>
<td>17</td>
<td>186</td>
<td>901</td>
<td>NULL</td>
<td>232</td>
<td>43</td>
</tr>
<tr>
<td>7</td>
<td>80268</td>
<td>7</td>
<td>12</td>
<td>64</td>
<td>47</td>
<td>641</td>
<td>264</td>
<td>232</td>
<td>NULL</td>
<td>122</td>
</tr>
<tr>
<td>8</td>
<td>104286</td>
<td>8</td>
<td>436</td>
<td>93</td>
<td>65</td>
<td>185</td>
<td>68</td>
<td>43</td>
<td>122</td>
<td>NULL</td>
</tr>
<tr>
<td>9</td>
<td>107734</td>
<td>9</td>
<td>24</td>
<td>105</td>
<td>41</td>
<td>52</td>
<td>28</td>
<td>20</td>
<td>42</td>
<td>66</td>
</tr>
<tr>
<td>10</td>
<td>72295</td>
<td>10</td>
<td>50</td>
<td>129</td>
<td>58</td>
<td>351</td>
<td>165</td>
<td>54</td>
<td>167</td>
<td>393</td>
</tr>
<tr>
<td>11</td>
<td>279329</td>
<td>11</td>
<td>54</td>
<td>167</td>
<td>109</td>
<td>424</td>
<td>210</td>
<td>164</td>
<td>524</td>
<td>319</td>
</tr>
<tr>
<td>12</td>
<td>108557</td>
<td>12</td>
<td>17</td>
<td>33</td>
<td>22</td>
<td>65</td>
<td>50</td>
<td>41</td>
<td>46</td>
<td>81</td>
</tr>
<tr>
<td>13</td>
<td>73248</td>
<td>13</td>
<td>7</td>
<td>51</td>
<td>37</td>
<td>44</td>
<td>27</td>
<td>18</td>
<td>32</td>
<td>37</td>
</tr>
<tr>
<td>14</td>
<td>92804</td>
<td>14</td>
<td>11</td>
<td>24</td>
<td>22</td>
<td>109</td>
<td>73</td>
<td>76</td>
<td>189</td>
<td>57</td>
</tr>
<tr>
<td>15</td>
<td>238683</td>
<td>15</td>
<td>48</td>
<td>120</td>
<td>57</td>
<td>140</td>
<td>122</td>
<td>73</td>
<td>138</td>
<td>171</td>
</tr>
<tr>
<td>16</td>
<td>425518</td>
<td>16</td>
<td>55</td>
<td>134</td>
<td>92</td>
<td>112</td>
<td>109</td>
<td>66</td>
<td>122</td>
<td>241</td>
</tr>
<tr>
<td>17</td>
<td>171553</td>
<td>17</td>
<td>40</td>
<td>100</td>
<td>61</td>
<td>67</td>
<td>57</td>
<td>46</td>
<td>74</td>
<td>162</td>
</tr>
<tr>
<td>18</td>
<td>208741</td>
<td>18</td>
<td>57</td>
<td>221</td>
<td>112</td>
<td>76</td>
<td>47</td>
<td>35</td>
<td>85</td>
<td>194</td>
</tr>
<tr>
<td>19</td>
<td>125578</td>
<td>19</td>
<td>24</td>
<td>92</td>
<td>31</td>
<td>97</td>
<td>82</td>
<td>47</td>
<td>89</td>
<td>105</td>
</tr>
<tr>
<td>20</td>
<td>385862</td>
<td>20</td>
<td>83</td>
<td>940</td>
<td>566</td>
<td>223</td>
<td>223</td>
<td>126</td>
<td>193</td>
<td>322</td>
</tr>
<tr>
<td>21</td>
<td>42050</td>
<td>21</td>
<td>15</td>
<td>19</td>
<td>19</td>
<td>28</td>
<td>15</td>
<td>2</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>22</td>
<td>72934</td>
<td>22</td>
<td>11</td>
<td>705</td>
<td>916</td>
<td>30</td>
<td>28</td>
<td>16</td>
<td>30</td>
<td>63</td>
</tr>
</tbody>
</table>
Progress

• Methods
  • Deterministic PPRL
  • Probabilistic PPRL
  • Deterministic garbled circuit
  • Probabilistic garbled circuit

• Conferences
  • Challenge workshop at the Academy health concordium
  • AMIA record linkage panel

• Grant
  • PCORI letter of intent submitted
Next steps

• Test on real data
  • Using VA datasets (IRB protocol approved)
  • Using USC data (IRB protocol approved)

• Establish pSCANNER protocol for expert determination on record linkage methods

• Link data based on practical use cases
  • Linkage among pSCANNER sites
  • CDRN-PPRN linkage
Team

• Daniella Meeker, Ph.D.
• Lucila Ohno-Machado, MD, Ph.D.
• Xiaoqian Jiang, Ph.D.
• Feng Chen, Ph.D.
• Jason Doctor, Ph.D.
• Michael Kahn, MD, Ph.D.

• Lisa Schilling, MD, MSPH
• Michael E. Matheny, MD, MS, MPH
• Jaideep Vaidya, Ph.D.
• Shuang Wang Ph.D.
• Ibrahim Lazrig, Ph.D. candidate
• Dax Westerman, MS
• Tara Knight, Ph.D.
• Thank you. Questions.