Probabilistic Estimates of Attribute Statistics and Match Likelihood for People Entity Resolution

Xin Wang, Ang Sun, Hakan Kardes, Siddharth Agrawal, Lin Chen, Andrew Borthwick
Our Mission

• Gather 20 billion raw records about people
  • Publicly Available
  • White Page (phone records, credit card headers)
  • Property Record
  • Court Record (criminal, civil, marriage/divorce)
  • Social Media
  • News
  • Professional

• Conflate all the records about the same person together

• Create a graph of 250 million profiles:
  • One profile for everybody in US
<table>
<thead>
<tr>
<th>Person</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naveen Jain</td>
<td>Chief Executive Officer</td>
</tr>
<tr>
<td>John Arnold</td>
<td></td>
</tr>
</tbody>
</table>

**John Arnold**

[Wikipedia](https://en.wikipedia.org) John Arnold was an English watchmaker and inventor. John Arnold was the first to design a watch that was both practical and accurate, and also brought the term “Chronometer” into use in its modern sense.
Our Approach

- Formulate the problem as a Graph Partition task
  - 7 billion nodes (each record as a node)
  - Weights on edges are similarity scores
    - from Machine Learning based models
  - Cluster graph into 313.9 million clusters

The Challenges

- Most graph partition algorithm can’t be scratch to such a scale
  - Dynamic Blocking:
    - Iteratively divide the graph into smaller subgraphs
- Limited resources:
  - 88 node hadoop cluster for multiple monthly builds
- Number of clusters in a sub graph unknown
- People records are ambiguous by nature
Low probability for two records with a common name in a big city to be about the same person
Two records with a Common name in a small town are more likely to be about the same person
Combining evidence from multiple locations increases the match likelihood
Incorporating other demographic information also helps with matching two records
Approximate Match Likelihood of Two Records with Demographics

- Demographic information we can use:
  - Name Frequency
  - Population of US
  - Population of a shared location
    - Can be a city, zip-code, county, MSA, state, or distance based

\[
\left( 1 - \frac{p_r(l)}{p_r(U S')} \right)^{m-1}
\]
Approximate Match Likelihood of Two Records with Demographics

- Demographic information we can use:
  - Name Frequency
  - Population of US
  - Population of a shared location
    - Can be a city, zip-code, county, MSA, state, or distance based
  - Birthday/Age information

\[
P_{nbp} = \left(1 - \min_l \frac{p_r(l)}{p_r(U,S)}\right)^m \frac{\Delta_b}{R_b} - 1
\]

\[
\Delta_b(x, y) = \min_{i,j} \{D_B(b_i, b_j) \mid b_i \in \bar{b}_x, b_j \in \bar{b}_y\}
\]
Approximate Match Likelihood of Two Records with Demographics

Patricia Johnson
- 312 Main St, Oberlin, OH
- 227 56th St, New York, NY 10022
- 2 Beechwood Way, Scarborough, NY
- 1502 SE 5th St Bellevue, WA
- DOB: 1974

Patricia Johnson
- BA, Oberlin College, 96
- Worked: Morgan Stanley, NY
- Worked: IBM Armonk, NY
- Worked: Microsoft, Redmond, WA
Approximate Match Likelihood of Two Records with Demographics

- Multiple Regions and multiple location matches in each region:
  - Name Frequency of a Region
  - Population of a Region
    - State, MSA
  - Population of a shared location
    - Can be a city, zip-code, county, MSA, state, or distance based
  - Birthday/Age information

\[ P_{rnbp} = \max_r P(r) \]

\[ P(r) = \begin{cases} 
(1 - \frac{\min_{l \in r} p_r(l)}{p_r(r)})^{\Phi_n(r) \frac{\Delta b}{R_b} - 1}, & \frac{\min_{l \in r} p_r(l)}{p_r(r)} < 1 \\
(1 - \frac{\Delta b}{R_b})^{\Phi_n(r) - 1}, & \frac{\min_{l \in r} p_r(l)}{p_r(r)} = 1 
\end{cases} \]  

(4)
Approximate Match Likelihood of Two Records with Demographics

\[
\phi(r) = \begin{cases} 
  e^{-\frac{\min_{l \in r} p_r(l)}{p_r(r)} \Phi_n(r) \frac{\Delta_b}{R_b}} - 1, & \frac{\min_{l \in r} p_r(l)}{p_r(r)} < 1 \\
  e^{-\frac{\Delta_b}{R_b} (\Phi_n(r) - 1)}, & \frac{\min_{l \in r} p_r(l)}{p_r(r)} = 1
\end{cases}
\]

\[
\lambda_r = \begin{cases} 
  \frac{\min_{l \in r} p_r(l)}{p_r(r)} \Phi_n(r), & \frac{\min_{l \in r} p_r(l)}{p_r(r)} < 1 \\
  \Phi_n(r) - 1, & \frac{\min_{l \in r} p_r(l)}{p_r(r)} = 1
\end{cases}
\]

\[
P_{lrbp} = \sum_{i=1}^{k} \frac{\prod_{j=1, j \neq i}^{k} \lambda_j}{\prod_{j=1, j \neq i}^{k} (\lambda_j - \lambda_i)} e^{-\lambda_i \frac{\Delta_b}{R_b}}
\]
How do we get the demographic statistics?

1. Population
   • US Population
   • State, MSA
   • County, City, Zipcode

2. Name Frequencies
   • US, State, MSA
   • Different Combination of Name Components
Data Sources and Their Record Counts

<table>
<thead>
<tr>
<th>Source</th>
<th>Count</th>
<th>Source</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>75,848,150</td>
<td>C</td>
<td>100,601,282</td>
</tr>
<tr>
<td>D</td>
<td>75,826,071</td>
<td>F</td>
<td>86,646,948</td>
</tr>
<tr>
<td>L</td>
<td>85,982,872</td>
<td>G</td>
<td>88,455,393</td>
</tr>
<tr>
<td>H</td>
<td>219,054,407</td>
<td>N</td>
<td>107,238,935</td>
</tr>
<tr>
<td>Q</td>
<td>215,271,940</td>
<td>T</td>
<td>262,904,192</td>
</tr>
<tr>
<td>X</td>
<td>515,176,119</td>
<td>E</td>
<td>228,545,777</td>
</tr>
<tr>
<td>M</td>
<td>423,808,772</td>
<td>Y</td>
<td>258,533,490</td>
</tr>
<tr>
<td>Z</td>
<td>255,112,376</td>
<td>V</td>
<td>304,804,288</td>
</tr>
<tr>
<td>I</td>
<td>909,702,398</td>
<td>B</td>
<td>7,677,583</td>
</tr>
</tbody>
</table>
Gaussian Truth Model For Estimating Name Frequencies
Implementation of GTM for Name Frequency Truth Estimation
Experimental Results:

Name Frequency Estimates (First Last)

<table>
<thead>
<tr>
<th>Config</th>
<th>Relative Error</th>
<th>Total Count</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>95280.7</td>
<td>5.190E8</td>
<td>5.289%</td>
</tr>
<tr>
<td>Config 4</td>
<td>91702.1</td>
<td>4.576E8</td>
<td>9.682%</td>
</tr>
</tbody>
</table>

Contribution of the Demographic Based Likelihood Feature

<table>
<thead>
<tr>
<th>Model</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>With ML Features</td>
<td>0.996</td>
<td>0.895</td>
</tr>
<tr>
<td>Without ML Feature</td>
<td>0.996</td>
<td>0.844</td>
</tr>
</tbody>
</table>
Q & A
Thank You