Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records

Kosuke Imai
Princeton University
Talk at SOSC Seminar
Hong Kong University of Science and Technology
June 14, 2017
Joint work with Ted Enamorado and Ben Fifield
Motivation

- In any given project, social scientists often rely on multiple data sets.
- We can easily merge data sets if there is a common unique identifier. E.g., use the `merge` function in R or Stata.
- How should we merge data sets if no unique identifier exists? Must use variables: names, birthdays, addresses, etc.
- Variables often have measurement error and missing values. Cannot use exact matching.
- What if we have millions of records? Cannot merge “by hand.”
- Merging is an uncertain process. Quantify uncertainty and error rates.
- **Solution:** Probabilistic Model
Data Merging Can be Consequential

- Turnout validation for the American National Election Survey
  - 2012 Election: self-reported turnout (78%) $\gg$ actual turnout (59%)

- Ansolabehere and Hersh (2012, *Political Analysis*): “electronic validation of survey responses with commercial records provides a far more accurate picture of the American electorate than survey responses alone.”

- Berent, Krosnick, and Lupia (2016, *Public Opinion Quarterly*): “Matching errors ... drive down “validated” turnout estimates. As a result, ... the apparent accuracy [of validated turnout estimates] is likely an illusion.”

- Challenge: Find 2500 survey respondents in 160 million registered voters (less than 0.001%) $\rightsquigarrow$ finding needles in a haystack

- Problem: match $\neq$ registered voter, non-match $\neq$ non-voter
Many social scientists use deterministic methods:
- match “similar” observations (e.g., Ansolabehere and Hersh, 2016; Berent, Krosnick, and Lupia, 2016)
- proprietary methods (e.g., Catalist)

Problems:
1. not robust to measurement error and missing data
2. no principled way of deciding how similar is similar enough
3. lack of transparency

Probabilistic model of record linkage:
- originally proposed by Fellegi and Sunter (1969, JASA)
- enables the control of error rates

Problems:
1. current implementations do not scale
2. missing data treated in ad-hoc ways
3. does not incorporate auxiliary information
The Fellegi-Sunter Model

- Two data sets: \( A \) and \( B \) with \( N_A \) and \( N_B \) observations
- \( K \) variables in common
- We need to compare all \( N_A \times N_B \) pairs
- Agreement vector for a pair \((i,j)\): \( \gamma(i,j) \)

\[
\gamma_k(i,j) = \begin{cases} 
0 & \text{different} \\
1 & \text{similar} \\
L_k - 2 & \text{identical} \\
\end{cases}
\]

- Latent variable:

\[
M_{i,j} = \begin{cases} 
0 & \text{non-match} \\
1 & \text{match} \\
\end{cases}
\]

- Missingness indicator: \( \delta_k(i,j) = 1 \) if \( \gamma_k(i,j) \) is missing
How to Construct Agreement Patterns

- Jaro-Winkler distance with default thresholds for string variables

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Data set A</th>
<th>Data set B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>James</td>
<td>Michael</td>
</tr>
<tr>
<td></td>
<td></td>
<td>V</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Smith</td>
<td>Martinez</td>
</tr>
<tr>
<td></td>
<td></td>
<td>780</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Devereux St.</td>
<td>16th St.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>John</td>
<td>James</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Martin</td>
<td>Smith</td>
</tr>
<tr>
<td></td>
<td></td>
<td>780</td>
<td>780</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Devereux St.</td>
<td>Devereux St.</td>
</tr>
</tbody>
</table>

Agreement patterns

- $A.1 - B.1$
  - 0 0 0 0 0
- $A.1 - B.2$
  - 2 NA 2 2 1
- $A.2 - B.1$
  - 0 NA 1 0 0
- $A.2 - B.2$
  - 0 NA 0 2 1
Independence assumptions for computational efficiency:

1. Independence across pairs
2. Independence across variables: $\gamma_k(i, j) \perp \perp \gamma_{k'}(i, j) \mid M_{ij}$
3. Missing at random: $\delta_k(i, j) \perp \perp \gamma_k(i, j) \mid M_{ij}$

Nonparametric mixture model:

$$\prod_{i=1}^{N_A} \prod_{j=1}^{N_B} \left\{ \sum_{m=0}^{1} \lambda^m (1 - \lambda)^{1-m} \prod_{k=1}^{K} \left( \prod_{\ell=0}^{L_k-1} \pi_{km\ell}^{1\{\gamma_k(i, j) = \ell\}} \right)^{1-\delta_k(i,j)} \right\}$$

where $\lambda = P(M_{ij} = 1)$ is the proportion of true matches and $\pi_{km\ell} = Pr(\gamma_k(i, j) = \ell \mid M_{ij} = m)$

Fast implementation of the EM algorithm (R package fastLink)

EM algorithm produces the posterior matching probability $\xi_{ij}$

Deduping to enforce one-to-one matching

1. Choose the pairs with $\xi_{ij} > c$ for a threshold $c$
2. Use Jaro’s linear sum assignment algorithm to choose the best matches
Controlling Error Rates

1 False negative rate (FNR):

\[
\frac{\text{#true matches not found}}{\text{# true matches in the data}} = \frac{P(M_{ij} = 1 \mid \text{unmatched})P(\text{unmatched})}{P(M_{ij} = 1)}
\]

2 False discovery rate (FDR):

\[
\frac{\text{# false matches found}}{\text{# matches found}} = P(M_{ij} = 0 \mid \text{matched})
\]

- We can compute FDR and FNR for any given posterior matching probability threshold $c$
Computational Improvements via Hashing

- Sufficient statistics for the EM algorithm: number of pairs with each observed agreement pattern
- $H_k$ maps each pair of records (keys) in linkage field $k$ to a corresponding agreement pattern (hash value):

$$H = \sum_{k=1}^{K} H_k$$

where

$$H_k = \begin{bmatrix}
h_{k}^{(1,1)} & h_{k}^{(1,2)} & \ldots & h_{k}^{(1,N_2)} \\
\vdots & \vdots & \ddots & \vdots \\
h_{k}^{(N_1,1)} & h_{k}^{(N_1,2)} & \ldots & h_{k}^{(N_1,N_2)}
\end{bmatrix}$$

and

$$h_{k}^{(i,j)} = 1 \{ \gamma_k(i,j) > 0 \} 2^{\gamma_k(i,j)+(k-1)\times L_k}$$

- $H_k$ is a sparse matrix, and so is $H$
- With sparse matrix, lookup time is $O(T)$ where $T$ is the number of unique patterns observed $T \ll \prod_{k=1}^{K} L_k$
Simulation Studies

- 2006 voter files from California (female only; 8 million records)
- Validation data: records with no missing data (340k records)
- Linkage fields: first name, middle name, last name, date of birth, address (house number and street name), and zip code
- 2 scenarios:
  1. Unequal size: 1:100, 10:100, and 50:100, larger data 100k records
  2. Equal size (100k records each): 20%, 50%, and 80% matched
- 3 missing data mechanisms:
  1. Missing completely at random (MCAR)
  2. Missing at random (MAR)
  3. Missing not at random (MNAR)
- 3 levels of missingness: 5%, 10%, 15%
- Noise is added to first name, last name, and address
- Results below are with 10% missingness and no noise
Error Rates and Estimation Error for Turnout

80% Overlap

50% Overlap

20% Overlap

False Negative Rate

Absolute Estimation Error (percentage point)

Kosuke Imai (Princeton)  Merging Large Data Sets  HKUST (June 14, 2017)
Accuracy of Estimated Error Rates

80% Overlap

50% Overlap

20% Overlap

False Discovery Rate

False Negative Rate

Kosuke Imai (Princeton)

Merging Large Data Sets

HKUST (June 14, 2017) 12 / 24
Runtime Comparisons

- No blocking, single core (parallelization possible with fastLink)

CCES respondents are matched with DIME donors (2010, 2012).

Use of a proprietary method, treating non-matches as non-donors.

Donation amount coarsened and small noise added.

4,432 (8.1%) matched out of 54,535 CCES respondents.

Discrepancies between self-reports and donation records:

1. 25% of self-reported donors are matched.
2. 54% of those who reported $300 or more donation are matched.
3. Democratic self-identified donors are better matched than Republicans.

We asked YouGov to apply *fastLink* for merging the two data sets.

We signed the NDA form ⇝ no coarsening, no noise.
Merging Process

- DIME: 5 million unique contributors
- CCES: 51,184 respondents (YouGov panel only)
- Exact matching: 0.33% match rate
- Blocking: 140 blocks using state and gender, followed by $k$-means
- Linkage fields: first name, middle name, last name, address (house number, street name), zip code
- Took 2.5 hours using a dual-core laptop
- Examples from the output of one block:

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>Middle</td>
<td>Last</td>
</tr>
<tr>
<td>agree</td>
<td>agree</td>
<td>agree</td>
</tr>
<tr>
<td>similar</td>
<td>NA</td>
<td>Agree</td>
</tr>
<tr>
<td>agree</td>
<td>NA</td>
<td>Agree</td>
</tr>
</tbody>
</table>
## Merge Results

<table>
<thead>
<tr>
<th>Match rate</th>
<th>Threshold</th>
<th>Liberal</th>
<th>Moderate</th>
<th>Strict</th>
<th>Proprietary</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td></td>
<td>9.61%</td>
<td>9.33%</td>
<td>8.74%</td>
<td>8.96%</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>8.61</td>
<td>8.45</td>
<td>8.11</td>
<td>8.25</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>10.74</td>
<td>10.31</td>
<td>9.46</td>
<td>9.75</td>
</tr>
<tr>
<td>All</td>
<td>FDR</td>
<td>1.36</td>
<td>0.79</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>0.87</td>
<td>0.53</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>1.80</td>
<td>1.03</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>FNR</td>
<td>29.58</td>
<td>31.26</td>
<td>35.18</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>10.60</td>
<td>11.91</td>
<td>15.21</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>40.97</td>
<td>42.88</td>
<td>47.16</td>
<td></td>
</tr>
</tbody>
</table>

- Estimated proportion of true matches:
  - 12.67% (All), 8.73% (Female), 16.95% (Male)

- Proportion of self-identified donors (over $200):
  - 10.46% (All), 7.71% (Female), 13.55% (Male)
Correlations with Self-reports and Matching Probabilities

**Common matches**
- Corr = 0.74
- N = 3520

**fastLink only matches**
- Corr = 0.52
- N = 830

**Proprietary only matches**
- Corr = 0.33
- N = 1064
Post-merge Analysis

1. Merged variable as the outcome
   - Assumption: No omitted variable for merge $Z_i^* \perp \!\!\!\!\perp X_i \mid (\delta, \gamma)$
   - Posterior mean of merged variable: $\zeta_i = \sum_{j=1}^{N_B} \xi_{ij} Z_j / \sum_{j=1}^{N_B} \xi_{ij}$
   - Regression:
     $$\mathbb{E}(Z_i^* \mid X) = \mathbb{E}\{\mathbb{E}(Z_i^* \mid \gamma, \delta, X_i) \mid X_i\} = \mathbb{E}(\zeta_i \mid X_i)$$

2. Merged variable as a predictor
   - Linear regression:
     $$Y_i = \alpha + \beta Z_i^* + \eta^\top X_i + \epsilon_i$$
   - Additional assumption: $Y_i \perp \!\!\!\!\perp (\delta, \gamma) \mid Z^*, X$
   - Weighted regression:
     $$\mathbb{E}(Y_i \mid \gamma, \delta, X_i) = \alpha + \beta \mathbb{E}(Z_i^* \mid \gamma, \delta, X_i) + \eta^\top X_i + \mathbb{E}(\epsilon_i \mid \gamma, \delta, X_i)$$
     $$= \alpha + \beta \zeta_i + \eta^\top X_i$$
Predicting Ideology using Contribution Status

- Hill and Huber regresses ideology score (−1 to 1) on the indicator variable for being a donor (merging indicator), turnout, and demographic variables.
- We use the weighted regression approach.

<table>
<thead>
<tr>
<th></th>
<th>Republicans</th>
<th>Democrats</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>fastLink</td>
</tr>
<tr>
<td>Contributor dummy</td>
<td>0.080</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>2012 General vote</td>
<td>0.095</td>
<td>0.094</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>2012 Primary vote</td>
<td>0.094</td>
<td>0.096</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.009)</td>
</tr>
</tbody>
</table>
We are merging two national voter files (2015 and 2016) with 160 million voters each!

We report the 20-state merge results today
  - Almost all merging is done within each state
  - But, some people move across states!
    \[ \approx 7.5 \text{ million cross-state movers between 2014 and 2015} \]

IRS Statistics of Income Migration Data
  - 9.2\% of residents moved to new address in same state
  - 1.6\% moved to a new state
  - Popular move: New York \(\rightarrow\) Florida, followed by California \(\rightarrow\) Texas

Linkage fields: first name, middle name, last name, date/year/month of birth, gender, house number (within-state only), street name (within-state only), date of registration (within-state only)
Incorporating Auxiliary Information on Migration

- Five-step process for across-state merge:
  1. Within-state estimation on random sample of each state
  2. Apply to full state to find non-movers and within-state movers
  3. Subset out successful matches
  4. Cross-state estimation on random sample to find cross-state movers
  5. Apply estimates to each cross-state pair

- Use of prior distribution
  1. Within-state merge:

\[ P(M_{ij} = 1) \approx \frac{\text{non-movers} + \text{in-state movers}}{N_A \times N_B} \]

\[ P(\gamma_{\text{address}}(i,j) = 0 \mid M_{ij} = 1) \approx \frac{\text{in-state movers}}{\text{in-state movers} + \text{non-movers}} \]

  2. Across-state merge:

\[ P(M_{ij} = 1) \approx \frac{\text{outflow from state } A \text{ to state } B}{N_A^* \times N_B^*} \]
## Merge Results

<table>
<thead>
<tr>
<th>Match rate</th>
<th>Threshold</th>
<th>Liberal</th>
<th>Moderate</th>
<th>Strict</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>89.45%</td>
<td>88.77%</td>
<td>88.40%</td>
<td>62.49%</td>
<td></td>
</tr>
<tr>
<td>Within-state</td>
<td>88.43%</td>
<td>88.21%</td>
<td>88.13%</td>
<td>62.47%</td>
<td></td>
</tr>
<tr>
<td>Across-state</td>
<td>1.02%</td>
<td>0.56%</td>
<td>0.27%</td>
<td>0.01%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FDR</th>
<th>All</th>
<th>0.26%</th>
<th>0.06%</th>
<th>0.01%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-state</td>
<td>0.12%</td>
<td>0.02%</td>
<td>0.01%</td>
<td></td>
</tr>
<tr>
<td>Across-state</td>
<td>0.14%</td>
<td>0.04%</td>
<td>0.01%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FNR</th>
<th>All</th>
<th>10.55%</th>
<th>11.23%</th>
<th>11.60%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-state</td>
<td>10.03%</td>
<td>10.67%</td>
<td>11.03%</td>
<td></td>
</tr>
<tr>
<td>Across-state</td>
<td>0.52%</td>
<td>0.55%</td>
<td>0.57%</td>
<td></td>
</tr>
</tbody>
</table>
Movers Found

Match Rates for Cross–State Movers

IRS Moving Probabilities for Cross–State Movers

- Recover intra-Northeast migration \((VT \rightarrow NH, ME \rightarrow NH)\)
- Recover intra-Midwest/Rockies migration \((NE \rightarrow IA, ID \rightarrow UT)\)
Concluding Remarks

- Merging data sets is critical part of social science research
  - merging can be difficult when no unique identifier exists
  - large data sets make merging even more challenging
  - yet merging can be consequential

- Merging should be part of replication archive

- We offer a fast, principled, and scalable merging method that can incorporate auxiliary information

- Pre-release of open-source software fastLink available upon request

- More applications under way:
  - Merging voter files over time and across states
  - Merging ANES/CCES with voter files