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
  $\rightsquigarrow$ e.g. Use the `merge` function in **R** or Stata
- How should we merge data sets if no unique identifier exists?
  $\rightsquigarrow$ must use variables: names, birthdays, addresses, etc.
- Variables often have measurement error and missing values
  $\rightsquigarrow$ cannot use exact matching
- What if we have millions of records?
  $\rightsquigarrow$ cannot merge "by hand"
- Merging is an uncertain process
  $\rightsquigarrow$ 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

Probabilistic Model of Record Linkage

- 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: $\mathcal{A}$ and $\mathcal{B}$ with $N_{\mathcal{A}}$ and $N_{\mathcal{B}}$ observations
- K variables in common
- We need to compare all $N_{\mathcal{A}} \times N_{\mathcal{B}}$ pairs
- Agreement vector for a pair (i, j): $\boldsymbol{\gamma}(i, j)$

$$\gamma_k(i,j) \;=\; \begin{cases} \quad 0 & \text{different} \\ \quad 1 & \\ \quad \vdots & \text{similar} \\ L_k - 2 & \\ L_k - 1 & \text{identical} \end{cases}$$

- 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

	Name			Address	
	First	Middle	Last	House	Street
Data set $\mathcal{A}$					
1	James	V	Smith	780	Devereux St.
2	John	NA	Martin	780	Devereux St.
Data set $\mathcal{B}$					
1	Michael	F	Martinez	4	16th St.
2	James	NA	Smith	780	Dvereuux St.
Agreement patterns					
$\mathcal{A}.1 - \mathcal{B}.1$	0	0	0	0	0
$\mathcal{A}.1 - \mathcal{B}.2$	2	NA	2	2	1
$\mathcal{A}.2 - \mathcal{B}.1$	0	NA	1	0	0
$\mathcal{A}.2 - \mathcal{B}.2$	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:

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

  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 ξ_{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
- $\mathbf{H}_k$ maps each pair of records (keys) in linkage field k to a corresponding agreement pattern (hash value):

$$\mathbf{H} \;=\; \sum_{k=1}^{K} \mathbf{H}_k \quad \text{where} \quad \mathbf{H}_k = \begin{bmatrix} h_k^{(1,1)} & h_k^{(1,2)} & \dots & h_k^{(1,N_2)} \\ \vdots & \vdots & \ddots & \vdots \\ h_k^{(N_1,1)} & h_k^{(N_1,2)} & \dots & h_k^{(N_1,N_2)} \end{bmatrix}$$

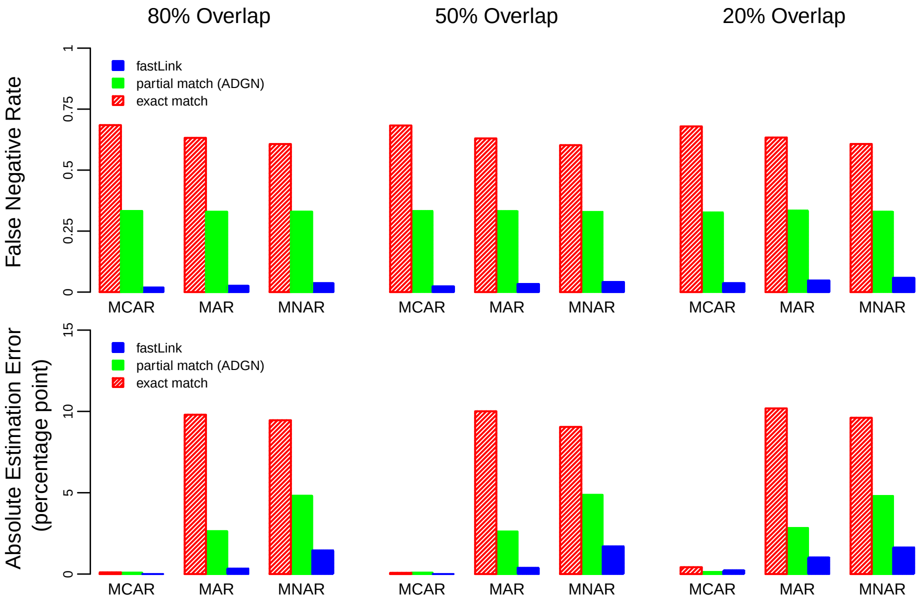
and $h_k^{(i,j)} = \mathbf{1}\left\{\gamma_k(i,j) > 0\right\} 2^{\gamma_k(i,j)+(k-1)\times L_k}$

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

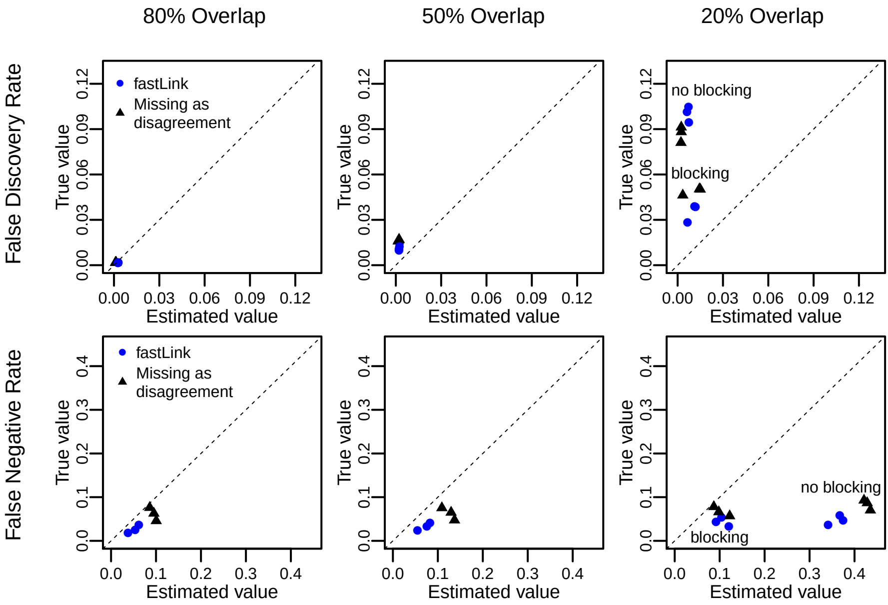
Simulation Studies

- 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

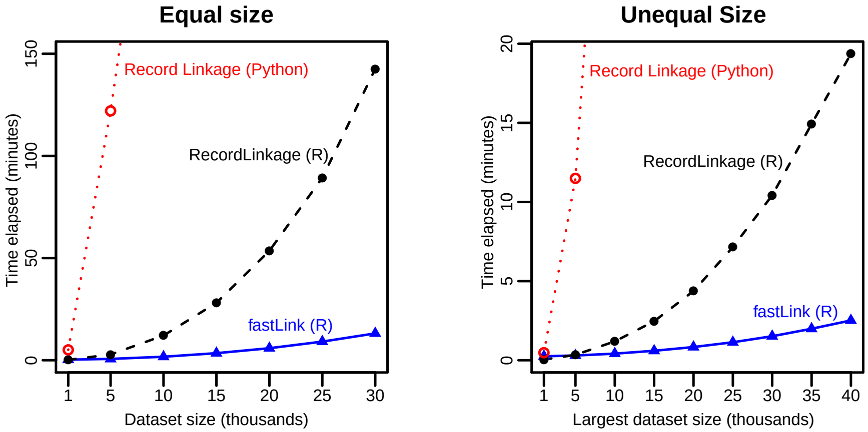
Error Rates and Estimation Error for Turnout



Accuracy of Estimated Error Rates



Runtime Comparisons



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

Application ❶: Merging Survey with Administrative Record

- Hill and Huber (2017, *Political Behavior*) study differences between donors and non-donors among CCES (2012) respondents
- 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:

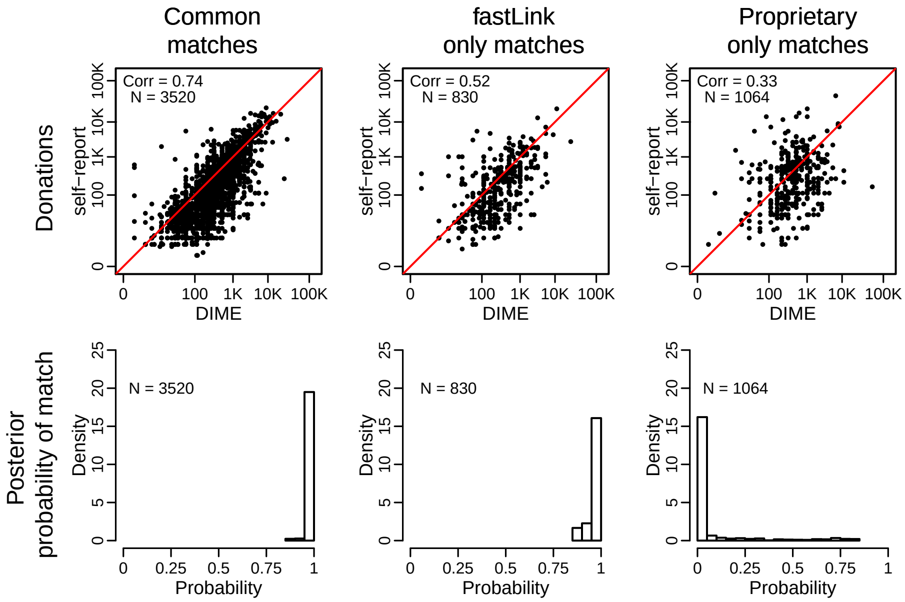
Name			Address			
First	Middle	Last	Street	House	Zip	Posterior
agree	agree	agree	agree	agree	agree	1.00
similar	NA	Agree	similar	agree	agree	0.93
agree	NA	Agree	disagree	disagree	NA	0.01

Merge Results

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

- 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



Post-merge Analysis

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

$$\mathbb{E}(Z_i^* \mid \mathbf{X}) \;=\; \mathbb{E}\{\mathbb{E}(Z_i^* \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i) \mid \mathbf{X}_i\} \;=\; \mathbb{E}(\zeta_i \mid \mathbf{X}_i)$$

2. Merged variable as a predictor
   - Linear regression:

$$Y_i \;=\; \alpha + \beta Z_i^* + \eta^\top \mathbf{X}_i + \epsilon_i$$

   - Additional assumption: $Y_i \perp\!\!\!\perp (\boldsymbol{\delta}, \boldsymbol{\gamma}) \mid \mathbf{Z}^*, \mathbf{X}$
   - Weighted regression:

$$\begin{aligned}\mathbb{E}(Y_i \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i) &= \alpha + \beta \mathbb{E}(Z_i^* \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i) + \eta^\top \mathbf{X}_i + \mathbb{E}(\epsilon_i \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{X}_i) \\ &= \alpha + \beta \zeta_i + \eta^\top \mathbf{X}_i\end{aligned}$$

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

	Republicans		Democrats	
	Original	`fastLink`	Original	`fastLink`
Contributor dummy	0.080	0.046	−0.180	−0.165
	(0.016)	(0.015)	(0.008)	(0.009)
2012 General vote	0.095	0.094	−0.060	−0.060
	(0.013)	(0.013)	(0.010)	(0.010)
2012 Primary vote	0.094	0.096	−0.019	−0.024
	(0.009)	(0.009)	(0.009)	0.008)

Application ❷: Merging National Voter Files

- 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!
    $\rightsquigarrow$ 7.5 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 $\longrightarrow$ Florida, followed by California $\longrightarrow$ 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:

$$P(M_{ij} = 1) \approx \frac{\text{non-movers} + \text{in-state movers}}{N_{\mathcal{A}} \times N_{\mathcal{B}}}$$

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

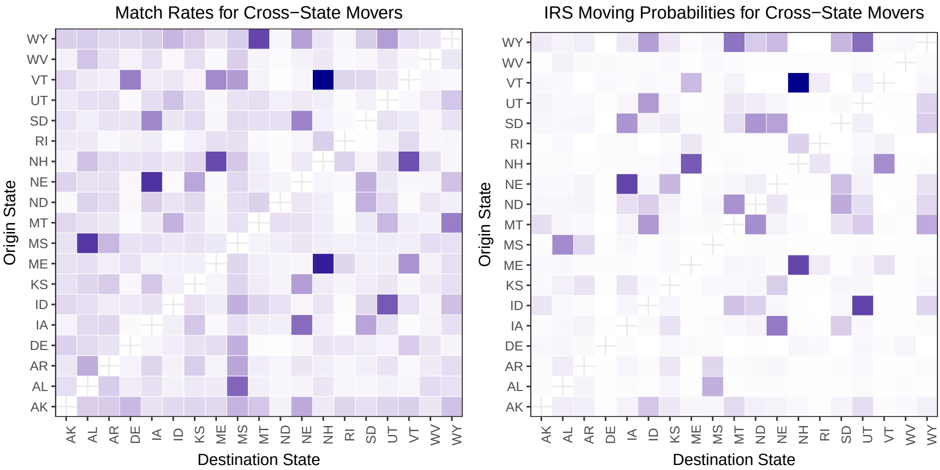
  2. Across-state merge:

$$P(M_{ij} = 1) \approx \frac{\text{outflow from state } \mathcal{A} \text{ to state } \mathcal{B}}{N^*_{\mathcal{A}} \times N^*_{\mathcal{B}}}$$

Merge Results

		Threshold			
		Liberal	Moderate	Strict	Exact
Match rate	All	89.45%	88.77%	88.40%	62.49%
	Within-state	88.43%	88.21%	88.13%	62.47%
	Across-state	1.02%	0.56%	0.27%	0.01%
FDR	All	0.26%	0.06%	0.01%	
	Within-state	0.12%	0.02%	0.01%	
	Across-state	0.14%	0.04%	0.01%	
FNR	All	10.55%	11.23%	11.60%	
	Within-state	10.03%	10.67%	11.03%	
	Across-state	0.52%	0.55%	0.57%	

Movers Found



- Recover intra-Northeast migration (VT → NH, ME → NH)
- Recover intra-Midwest/Rockies migration (NE → IA, ID → 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