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

Kosuke Imai

Princeton University

Talk at the Institute for Quantitative Social Science

Harvard University

March 29, 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.
- What if we have millions of records (e.g., voter files)?
  $\rightsquigarrow$ cannot merge "by hand", need for a scalable algorithm
- Variables often have measurement error and missing values
  $\rightsquigarrow$ cannot use exact matching
- Merging is an uncertain process
  $\rightsquigarrow$ quantify uncertainty and error rates
- Probabilistic model as a solution
- Initial motivation: merging national voter files

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 to large data sets
  2. missing data are treated as disagreements
  3. do not incorporate auxiliary information

The Fellegi and Sunter Model

- Two data sets: $\mathcal{D}_1$ and $\mathcal{D}_2$ with N_1 and N_2 observations
- $\mathbf{Z}$: K linkage variables in common
- Consider all $N_1 \times N_2$ pairs
- Agreement vector for a pair (i, j): $\boldsymbol{\gamma}(i, j)$

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

- Latent variable:

$$U(i,j) = \begin{cases} 0 & \text{non-match} \\ 1 & \text{match} \end{cases}$$

- Missingness indicator: $M_k(i,j) = 1$ if $\gamma_k(i,j)$ is missing

- Independence assumptions for computational efficiency:
    1. Independence across pairs
    2. Independence across variables: $\gamma_k(i,j) \perp\!\!\!\perp \gamma_{k'}(i,j) \mid U(i,j)$
    3. Missing at random: $M_k(i,j) \perp\!\!\!\perp \gamma_k(i,j) \mid U(i,j), \mathbf{Z}$

- Nonparametric mixture model:

$$\prod_{i=1}^{N_1}\prod_{j=1}^{N_2}\left\{\lambda \prod_{k=1}^{K} P(\gamma_k(i,j) \mid U(i,j)=1)^{1-M_k(i,j)} + (1-\lambda)\prod_{k=1}^{K} P(\gamma_k(i,j) \mid U(i,j)=0)^{1-M_k(i,j)}\right\}$$

where $\lambda = P(U(i,j)=1)$ is the proportion of true matches

- Fast implementation of the EM algorithm (**R** package fastLink)

Hashing

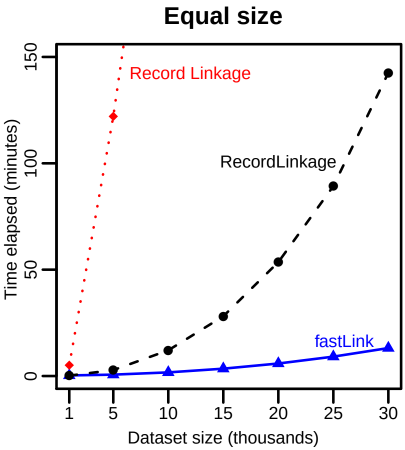
- Sufficient statistics for the EM algorithm: number of pairs with each *observed* agreement pattern
- $\mathbf{H}_k$ maps each pair of records (keys) in linkage field k to a corresponding agreement pattern (hash value):

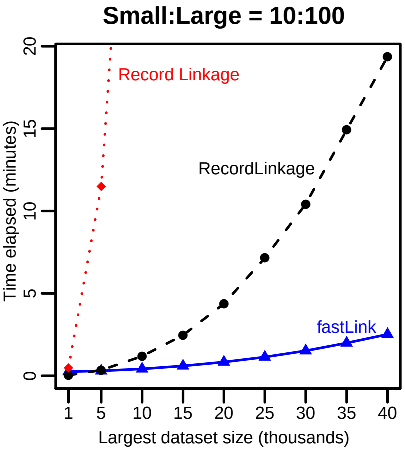
$$\mathbf{H} = \sum_{k=1}^{K} \mathbf{H}_k \quad \text{where} \quad \mathbf{H}_k = \begin{bmatrix} h_k^{(1,1)} & h_k^{(1,2)} & \dots & h_k^{(1,N_2)} \\ \vdots & \vdots & \ddots & \vdots \\ h_k^{(N_1,1)} & h_k^{(N_1,2)} & \dots & h_k^{(N_1,N_2)} \end{bmatrix}$$

  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 $T \ll \prod_{k=1}^{K} L_k$
- Use of many linkage fields $\rightsquigarrow$ min hashing and locally sensitive hashing

Runtime Comparison with Another **R** Package





- No blocking, single core (parallelization possible)
- `RecordLinkage` cannot merge two equal sized data sets of more than 30k observations on an ordinary laptop without blocking

Controlling Error Rates

1. False negative rate (FNR):

$$\frac{\#\text{true matches not found}}{\#\text{ true matches in the data}} = \frac{P(U(i,j)=1 \mid \text{unmatched})P(\text{unmatched})}{P(U(i,j)=1)}$$

2. False discovery rate (FDR):

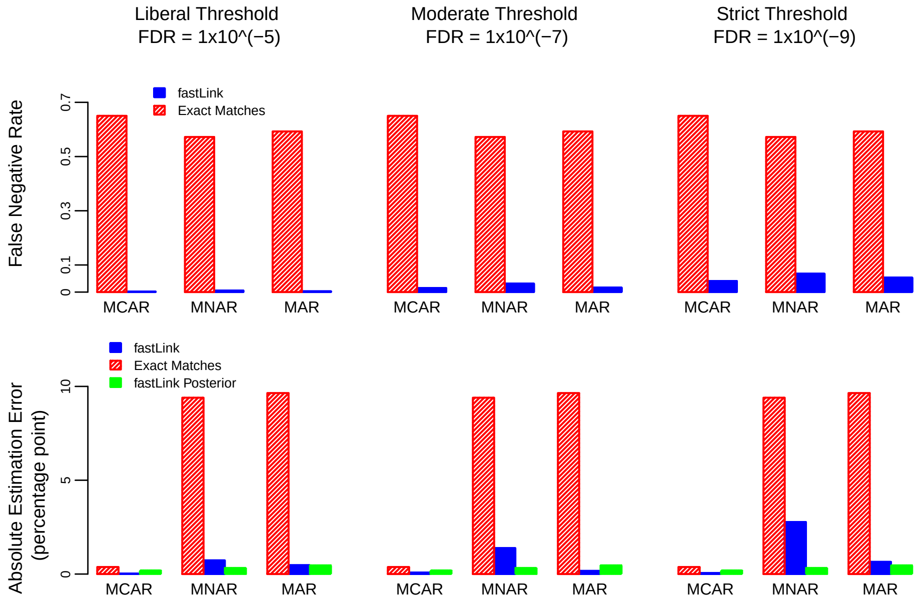
$$\frac{\#\text{ false matches found}}{\#\text{ matches found}} = P(U(i,j)=0 \mid \text{matched})$$

- We typically control FDR
- Simulation studies show FDR and FNR are accurately estimated

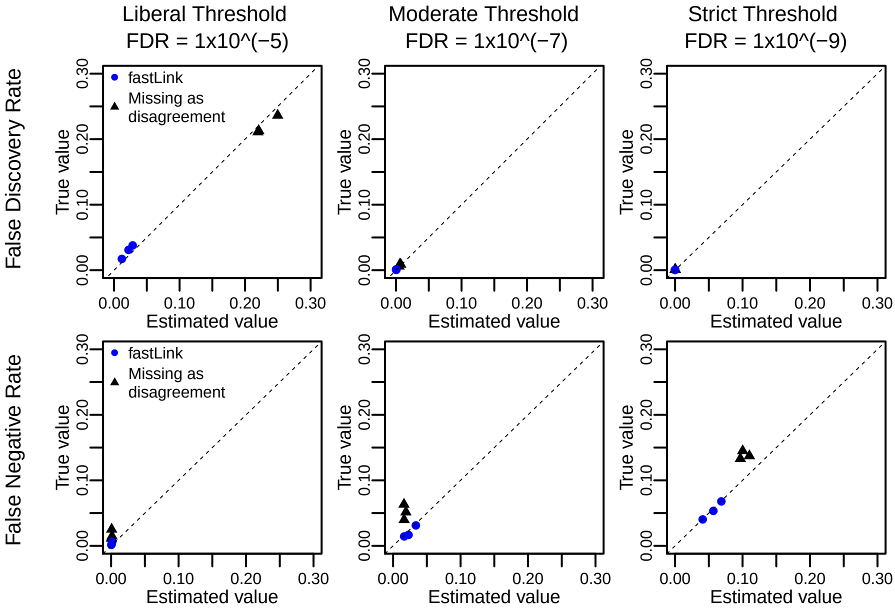
Simulation Studies

- 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. Equal size (25k records each): 20%, 50%, and 80% matched
  2. Unequal size: 1:100, 10:100, and 50:100
- 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: mild (1%), moderate (10%), severe (15%)
- Noise is added to first name, last name, and address
- Results below are with moderate missingness and no noise

Error Rates and Estimation Error for Turnout



Accuracy of Estimated Error Rates



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% (1%) of self-reported donors (non-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
- Clerical review examples:

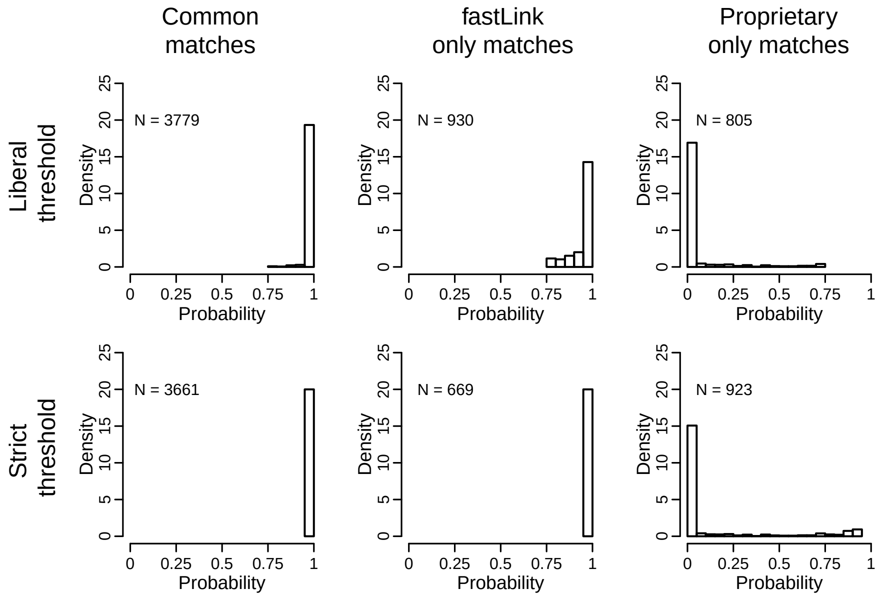
Name			Address				
First	Middle	Last	Street	House	Zip	FS weight	Posterior
2	2	2	2	2	2	38.86	1.00
1	NA	2	1	2	2	15.78	0.93
2	NA	2	0	0	NA	7.59	0.01

Merge Results

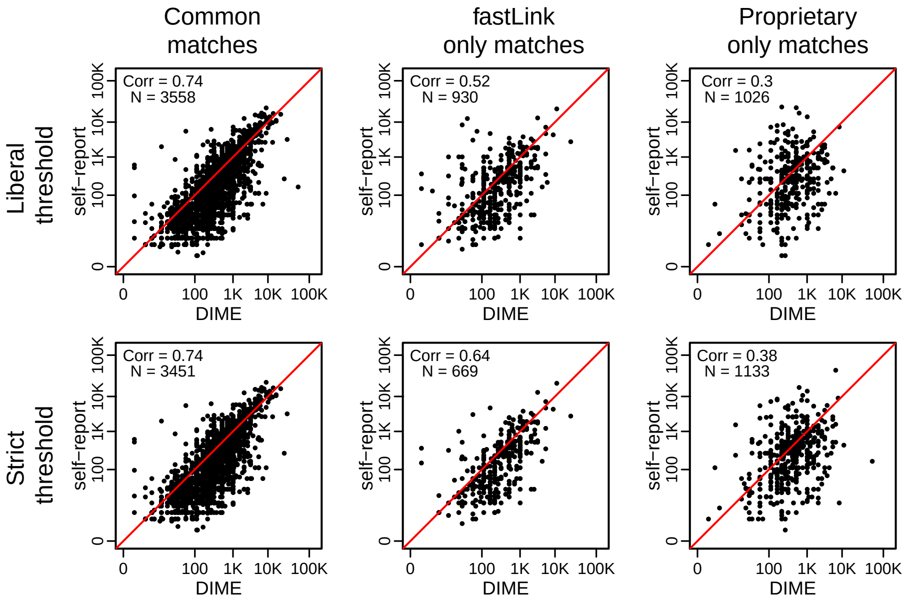
		Threshold			
		Liberal	Moderate	Strict	Proprietary
Match rate	All	9.61%	9.33%	8.74%	8.96%
	Female	8.61	8.45	8.11	8.25
	Male	10.74	10.31	9.46	9.75
FDR	All	1.36	0.79	0.21	
	Female	0.87	0.53	0.16	
	Male	1.80	1.03	0.27	
FNR	All	29.58	31.26	35.18	
	Female	10.60	11.91	15.21	
	Male	40.97	42.88	47.16	

- 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)

Posterior Probabilities of Matching



Correlations with Self-reported Donation (log scale)



Post-Mege Analysis

- Regression model of interest: $P(Y \mid M^*, \mathbf{X})$
- Assumptions:
  1. No omitted variable for merge: $M^* \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$
  2. No omitted variable for outcome: $Y \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, M^*$
- Weighted linear regression:

$$Y_i \;=\; \alpha + \beta W_i + \gamma^\top \mathbf{X}_i$$

where $W_i = \Pr(M_i^* = 1 \mid \mathbf{Z})$ is the posterior matching probability

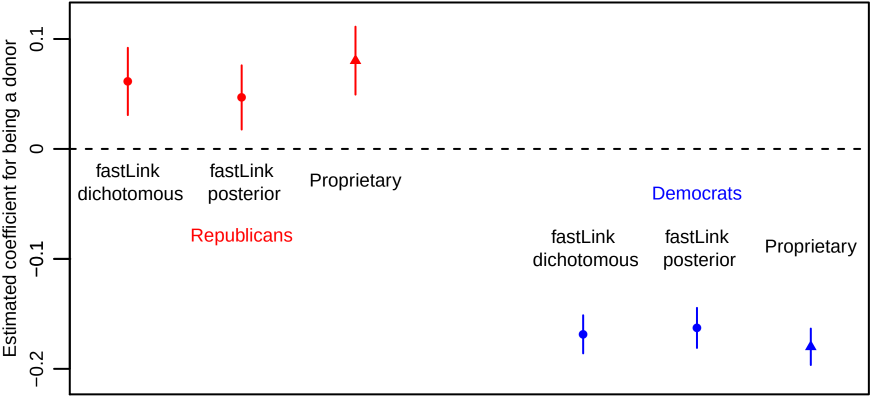
- Weighted maximum likelihood:

$$\mathcal{L} \;=\; W_i \log P(Y_i \mid M_i^* = 1, \mathbf{X}_i) + (1 - W_i) \log P(Y_i \mid M_i^* = 0, \mathbf{X}_i)$$

- Similarly, under $M \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}$, we estimate $\Pr(M_i^* = 1 \mid \mathbf{X}) = \mathbb{E}(W_i \mid \mathbf{X})$

Post-Mege Analysis Results

- 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 our merging indicator and posterior matching probability



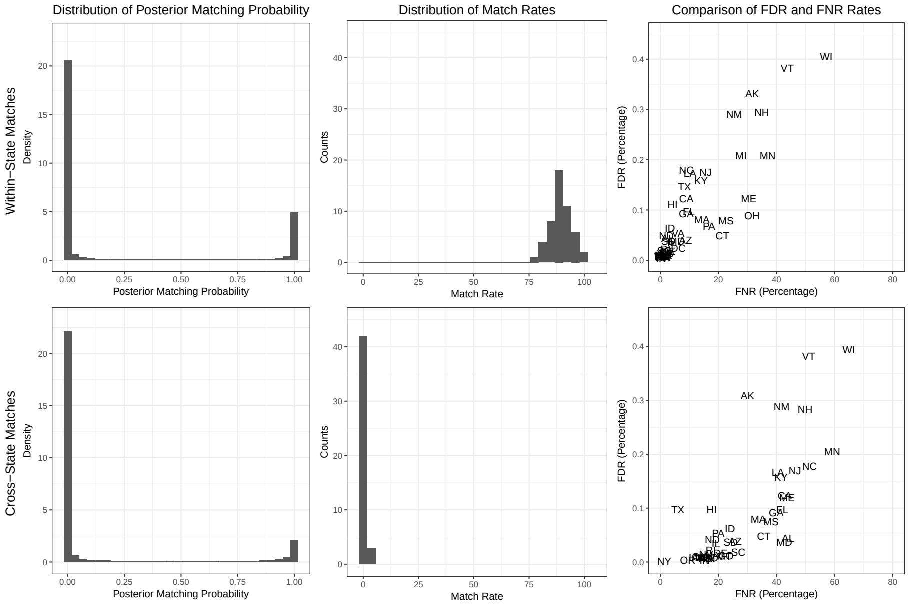
Application ❷: Merging National Voter Files

- Merge two national voter files (2015 and 2016) with 160 million voters each
    - Almost all merging is done within each state
    - But, some people move across states!
    - IRS Statistics of Income Migration Data
        - 9.2% of residents moved to new address in same state
        - 1.6% moved to a new state
        - New York $\longrightarrow$ Florida, followed by California $\longrightarrow$ Texas
- Three-step process for cross-state merge (blocking by gender):
    1. Within-state merge to find non-movers and within-state movers
    2. Subset out successful matches
    3. Run cross-state merge to find cross-state movers
- Linkage fields: first name, middle name, last name, date of birth, house number (within-state only), street name (within-state only), date of registration (within-state only)

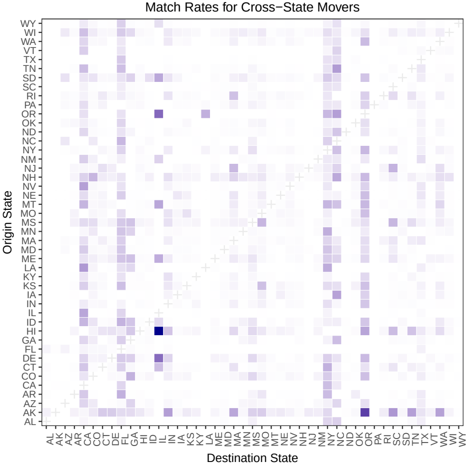
Merge Results

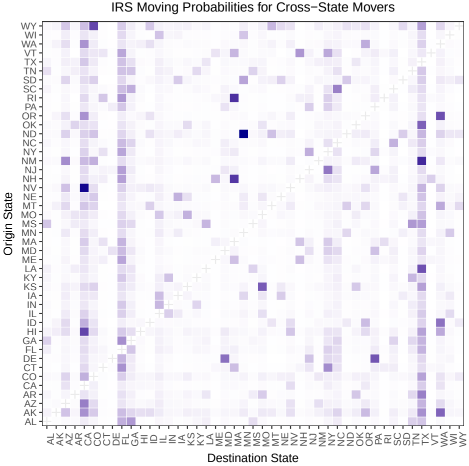
		Threshold			
		Liberal	Moderate	Strict	Exact
Match rate	All	95.44%	93.24%	90.86%	62.7%
	Within-state	90.32%	90.02%	89.45%	62.64%
	Across-state	5.12%	3.22%	1.41%	0.05%
FDR	All	1.8%	0.77%	0.14%	
	Within-state	1.17%	0.54%	0.1%	
	Across-state	0.62%	0.23%	0.04%	
FNR	All	15.87%	17.88%	20.76%	
	Within-state	9.73%	11.61%	14.32%	
	Across-state	6.14%	6.27%	6.44%	

Merge Results



Movers Found





- Recover the outflow of movers to California and Florida
- More difficulty finding movers to Texas
- IRS and match rate correlate at 0.29

Use of Auxiliary Information as Prior Distributions

- Within-state merge:

$$P(U(i,j)=1) \approx \frac{\text{non-movers} + \text{in-state movers}}{N_1 \times N_2}$$

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

- Across-state merge:

$$P(U(i,j)=1) \approx \frac{\text{outflow from county 1 to county 2}}{N_1^* \times N_2^*}$$

where N_j^* is the sample size data set j after removing in-state matches

- Conjugate priors with the above means and user-specified prior variances

Conclusions and Next Steps

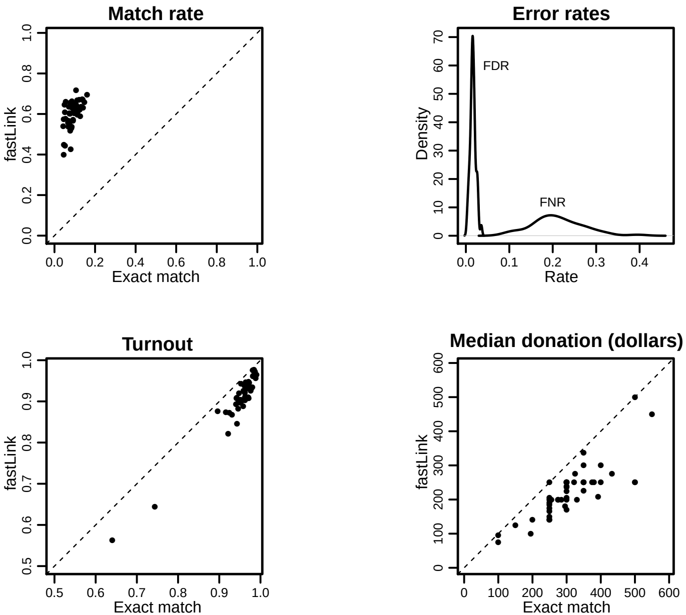
- Merging data sets is critical part of social science research
  - merging can be difficult when no unique identifier exists
  - large data sets make merging even more challenging
  - yet merging can be consequential
- Merging should be part of replication archive
- We offer a fast, principled, and scalable merging method that can incorporate auxiliary information
- Open-source software fastLink will be released soon
- More applications under way:
  - Merging CCES with voter files
  - Merging ANES with voter files
- Stochastic blocking, merging with more than two files over time
- Open problem: privacy-preserving record linkage

Extra Slides

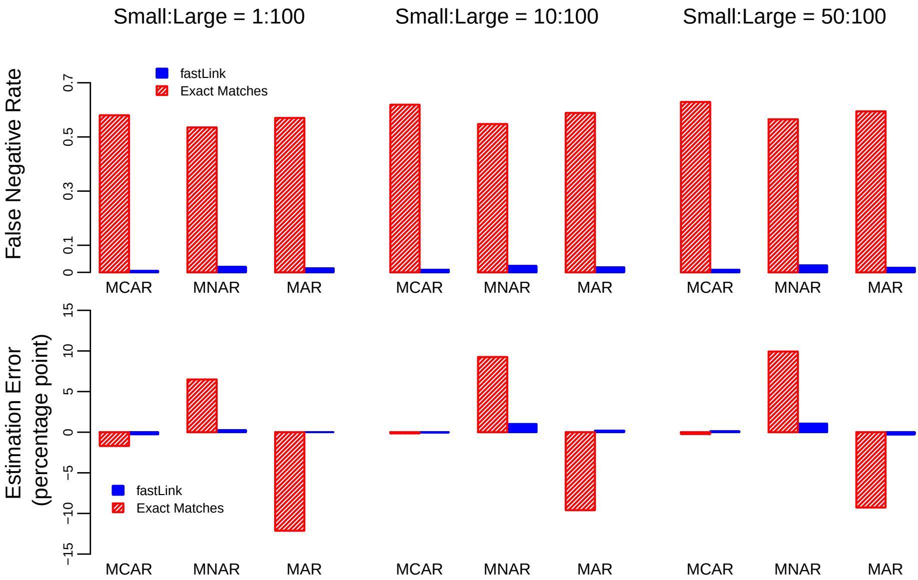
Application ❸: Merging Administrative Records

- Merge DIME (2012) with L2 Voter file (2014)
- Within-state merge for 50 states plus DC
- DIME: 5 million unique contributors
- Voter file: 160 million voters
- create 535 blocks (at most 500k records per block) using state and gender, followed by k-means on first name
- Linkage fields: first name, middle name, last name, address (house number, street name), zip code
- Took 30 hours using 360 cores (20 minutes per block with 60 cores)
- Challenges:
  - two big data sets with two years apart
  - at least 20% of contributors use P.O. Box as their address
  - no date of birth information in DIME

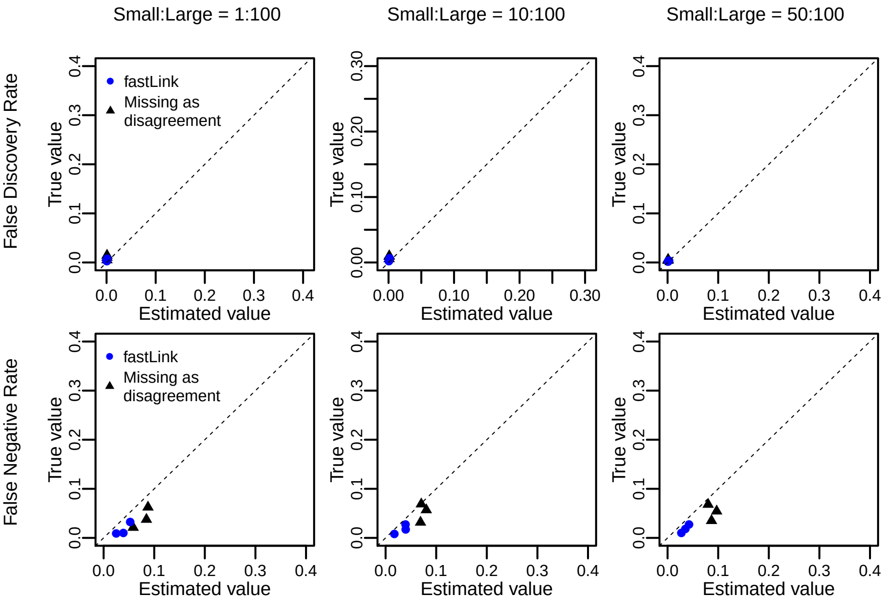
Empirical Results



Varying Data Set Sizes



Varying Data Set Sizes



Separate Results for Within- and Across-state Merge

		Threshold			
		Liberal	Moderate	Strict	Exact
Match rate	All	95.44%	93.24%	90.86%	62.7%
	Within-state	90.32%	90.02%	89.45%	62.64%
	Across-state	1.9%	1.2%	0.52%	0.02%
FDR	All	1.8%	0.77%	0.14%	
	Within-state	1.24%	0.56%	0.1%	
	Across-state	11.63%	6.71%	2.47%	
FNR	All	15.87%	17.88%	20.76%	
	Within-state	9.73%	11.61%	14.32%	
	Across-state	80.67%	87.17%	94.16%	