Validating Self-reported Turnout by Linking Public Opinion Surveys with Administrative Records

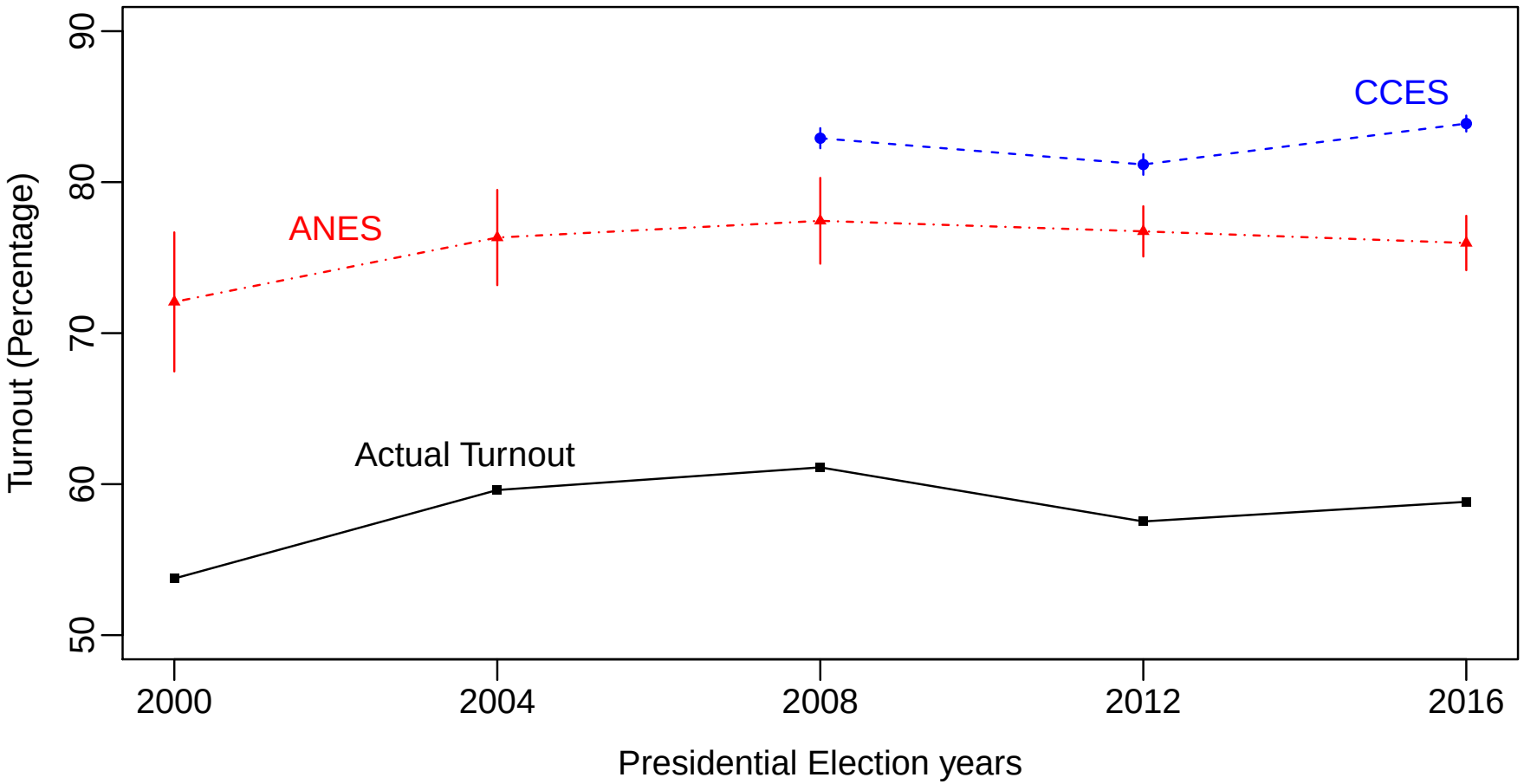
Ted Enamorado Kosuke Imai

Princeton University

Institute for Policy Research
Northwestern University

January 24, 2018

Bias of Self-reported Turnout



- Where does this gap come from?
- Nonresponse, Misreporting, Mobilization

Turnout Validation Controversy

- The Help America Vote Act of 2002 $\rightsquigarrow$ Development of systematically collected and regularly updated nationwide voter registration records
- 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 several thousand survey respondents in 180 million registered voters (less than 0.001%) $\rightsquigarrow$ finding needles in a haystack
- Problems: false matches and false non-matches

Methodological Motivation

- In any given project, social scientists often rely on multiple data sets
- Cutting-edge empirical research often merges large-scale administrative records with other types of data
- 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 data sets is an uncertain process
  $\rightsquigarrow$ quantify uncertainty and error rates
- **Solution:** Probabilistic Model

Overview of the Talk

1. Turnout validation for the 2016 Cooperative Congressional Election Study
2. Probabilistic method of record linkage and fastLink
3. Simulation study to compare fastLink with deterministic methods
4. Empirical findings:
   - fastLink recovers the actual turnout
   - Bias of self-reported turnout is largely driven by misreporting
   - those who are educated, rich, and interested in politics tend to misreport
   - fastLink performs at least as well as a state-of-art proprietary method

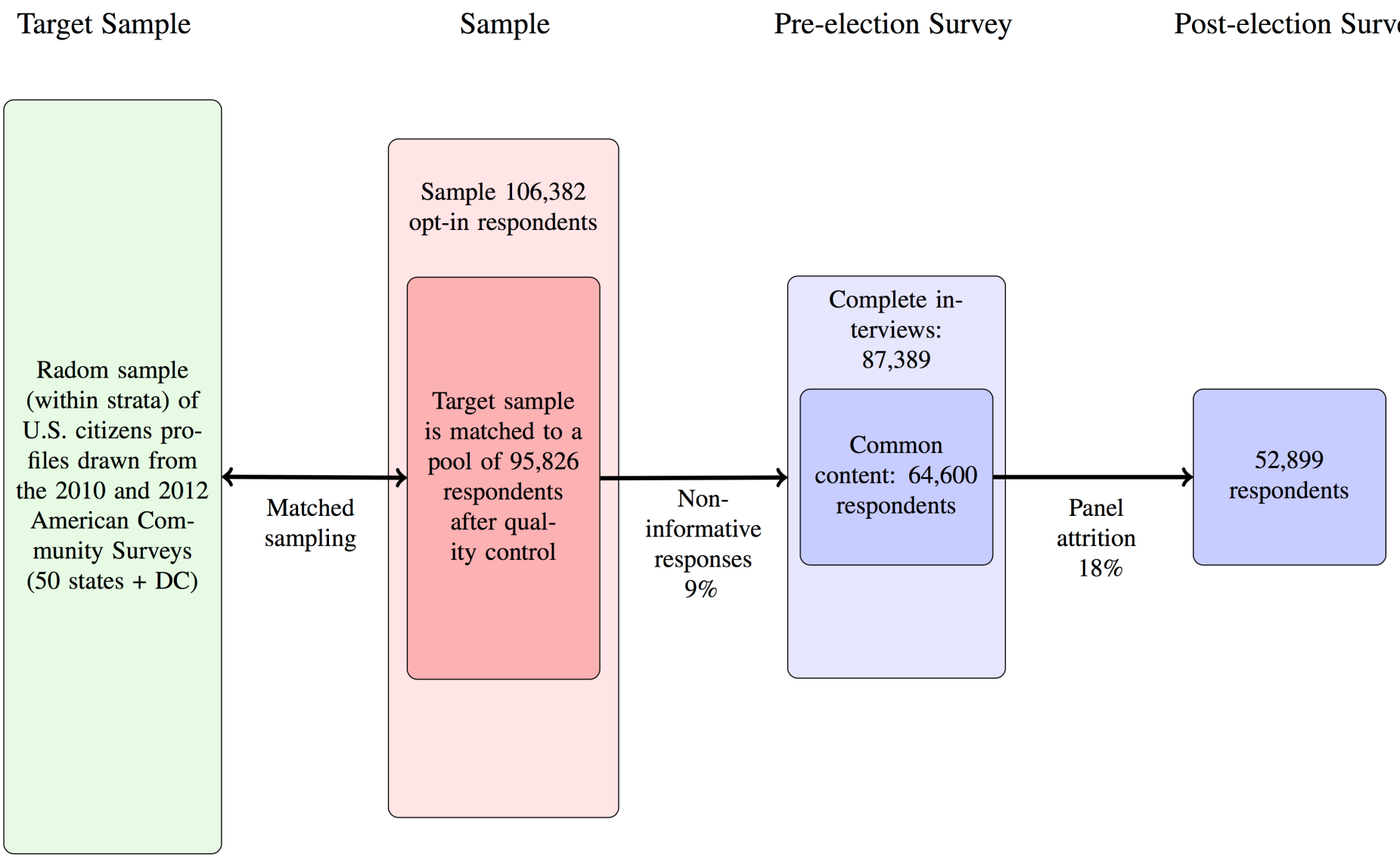
The 2016 US Presidential Election

- Donald Trump's surprising victory $\rightsquigarrow$ failure of polling
- Non-response and social desirability biases as possible explanations

- Two validation exercises:
  1. The 2016 American National Election Study (ANES)
  2. The 2016 Cooperative Congressional Election Study (CCES)
- We merge the survey data with a nationwide voter file
- We only report the preliminary results from the CCES validation today

- The voter file was obtained in July 2017 from L2, Inc.
  - total of 182 million records
  - 8.6 million "inactive" voters

CCES Sampling Design

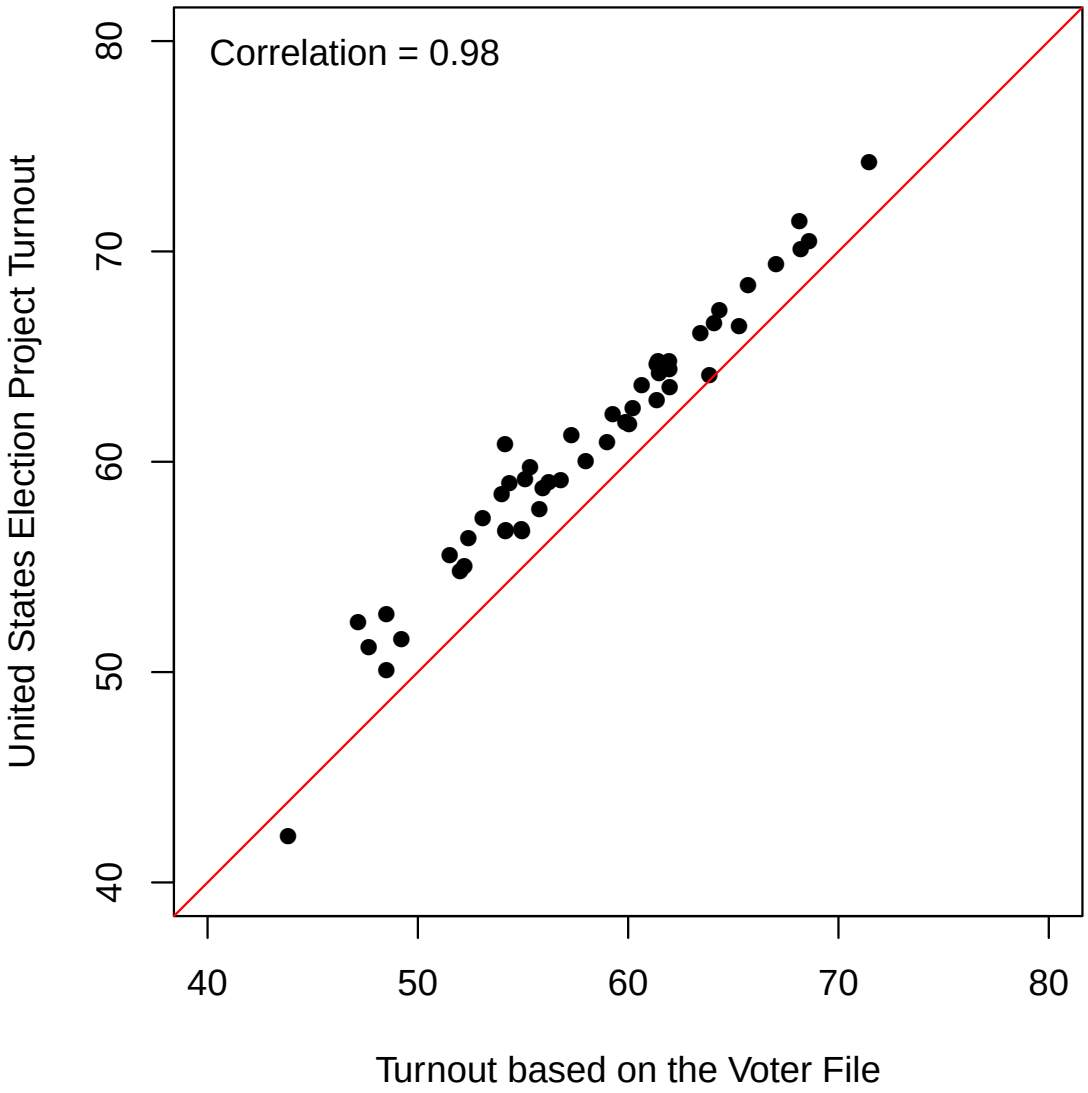


Bias of Self-reported Turnout and Registration Rates

	CCES	Election project	Voter file all	Voter file active	CPS
Turnout rate (%)	83.88 (0.27)	58.83	57.55		61.38 (1.49)
Registration rate (%)	91.99 (0.20)		80.37	76.57	70.34 (1.40)
Target pop. size (millions of voters)	224.10	232.40	227.60	227.60	224.10

- All results are based on the CCES pre-validation survey weights
- Target population
  - US citizens of voting age in 50 states plus Washington DC
  - Election project: cannot adjust for overseas population
  - Voter file: the deceased and out-of-state movers (after the election) are removed

Election Project vs. Voter File



Preprocessing

- We merge 64,600 CCES respondents with the nationwide voter file using name, age, gender, and address

- Standardization:
  - Name: first, middle, and last name
  - Missing (2.7%), Use of initials (5.9%), Complete (91.4%)
  - Address: house number, street name, zip code, and apartment number
  - Missing (11.6%), P.O. Box (2.6%), Complete (85.9%)

- Blocking:
  - Direct comparison $\rightsquigarrow$ 18 trillion pairs
  - Blocking by gender and state $\rightsquigarrow$ 102 blocks
  - Block size: from 3 million (WY/Male) to 25 billion pairs (CA/Male)
  - Apply the merge algorithm within each block

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, YouGov)
- 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): $\boldsymbol{\gamma}(i, j)$

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

- 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):

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

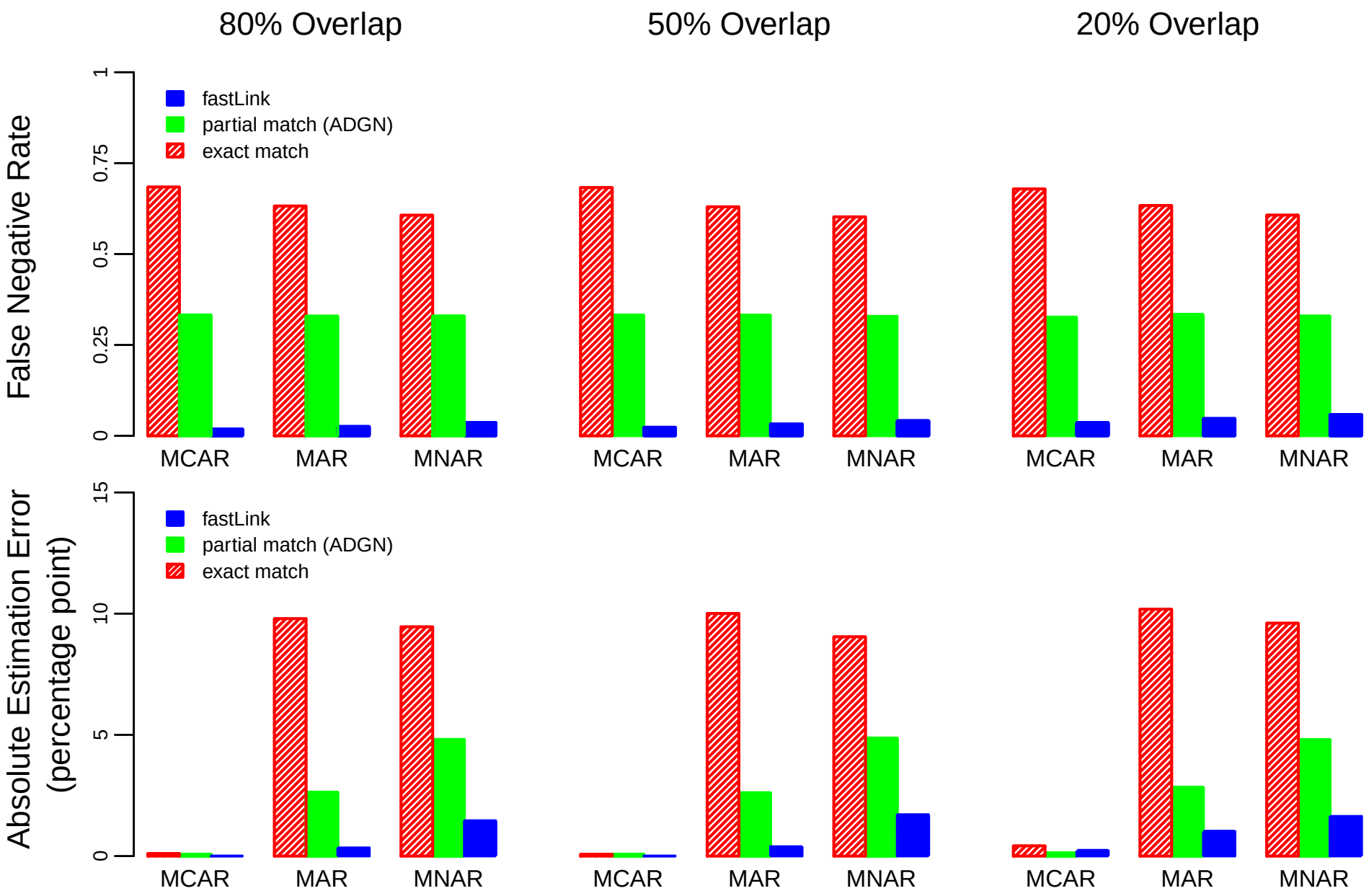
and $h_k^{(i,j)} = \mathbf{1}\{\gamma_k(i,j) > 0\}\, 2^{\gamma_k(i,j)+(k-1)\times L_k}$

- $\mathbf{H}_k$ 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 $T \ll \prod_{k=1}^{K} L_k$

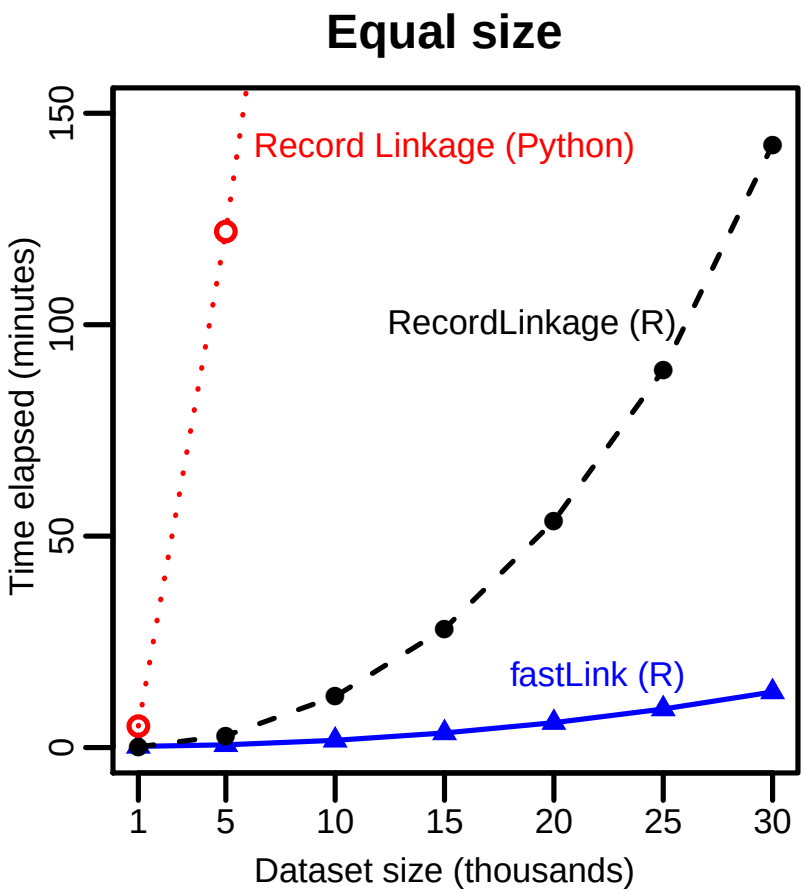
Simulation Studies

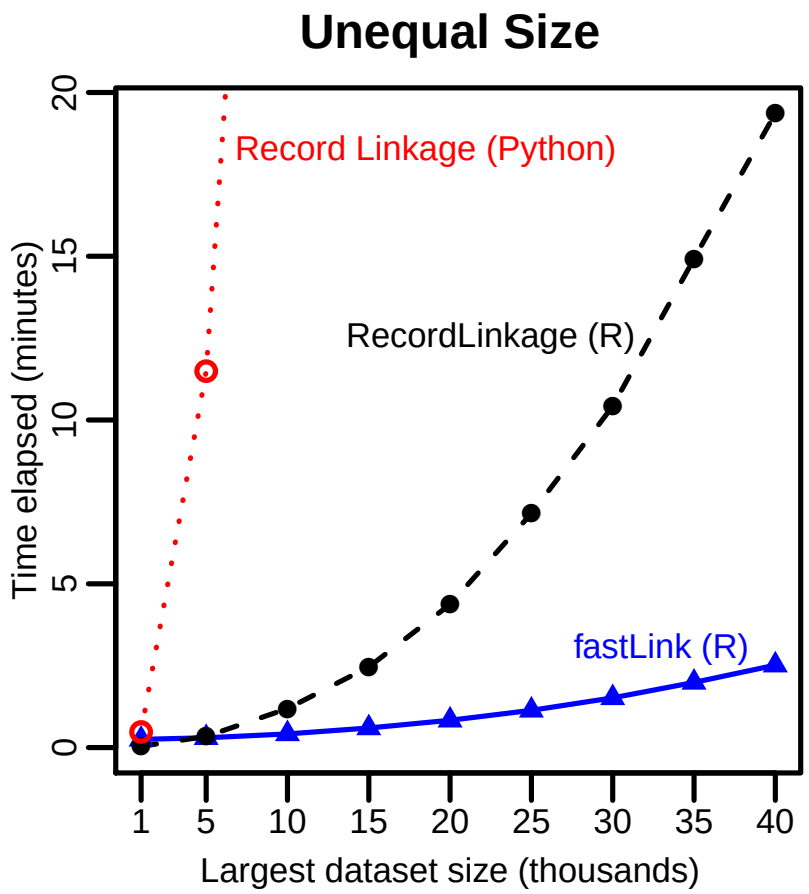
- 2006 voter files from California (female only; 8 million records)
- Validation data: records with no missing data (340k records)
- Linkage fields: first name, middle name, last name, date of birth, address (house number and street name), and zip code
- 2 scenarios:
  1. Unequal size: 1:100, 10:100, 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



Runtime Comparisons





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

Merge Procedure and Results

- Use of three agreement levels for string variables and age
- Merge process:
  1. within-block merge
  2. remove within-state matches (posterior match prob. > 0.75)
  3. across-state merge (exact match on gender, names, age)
- Our analysis uses posterior match probability as well as pre-validation CCES sampling weights
- Match rate as an estimate of registration rate:

fastLink		Voter file		
Pre-election	Post-election	all	active	CPS
66.60	70.52	80.37	76.57	70.34
(0.18)	(0.19)			(1.40)

Turnout Validation Results

- Comparison with actual turnout rates:

fastLink		Actual turnout	
Pre-election	Post-election	Voter file	Election Project
54.11 (0.31)	55.67 (0.37)	57.55	58.83

- Validated turnout rates by response categories:

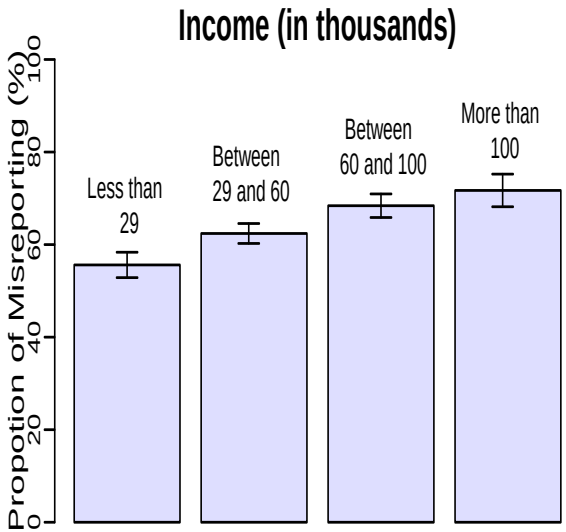
		Registered		Post-election
	Not registered	Did not Vote	Voted	attrition
fastLink	16.37 (0.84)	10.15 (0.73)	73.05 (0.28)	24.02 (0.60)
N	4684	3237	44796	11701

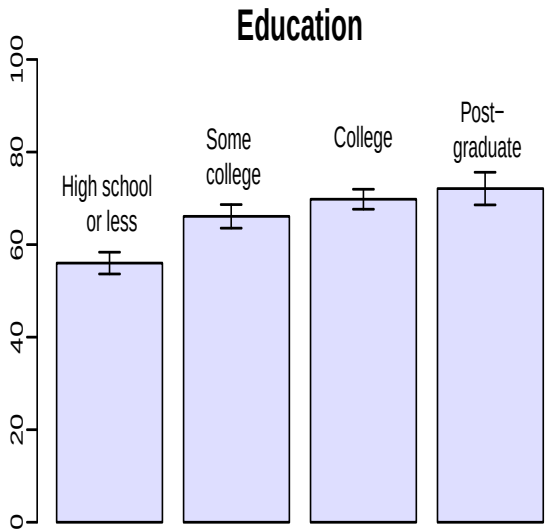
Do Voters Misreport Turnout?

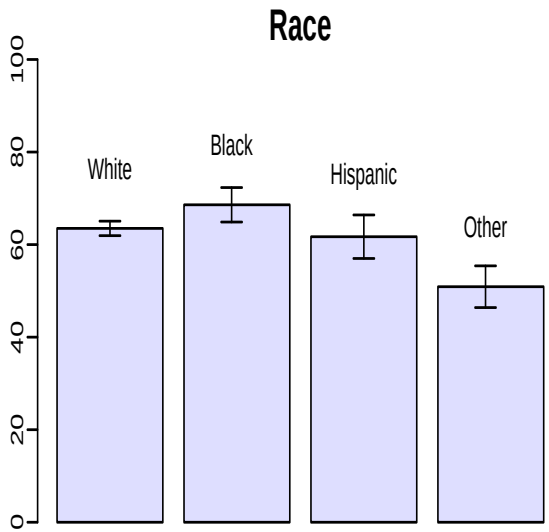
- Berent, Krosnick, and Lupia (2016) argue that voters don't misreport:
  - Poor quality of voter files and difficulty of merging
  - Failure to match survey respondents who actually voted
  - Results in a lower validated turnout rate
- As evidence, BKL show:
  1. the match rate is lower than the registration rate
  2. matched voters do not lie

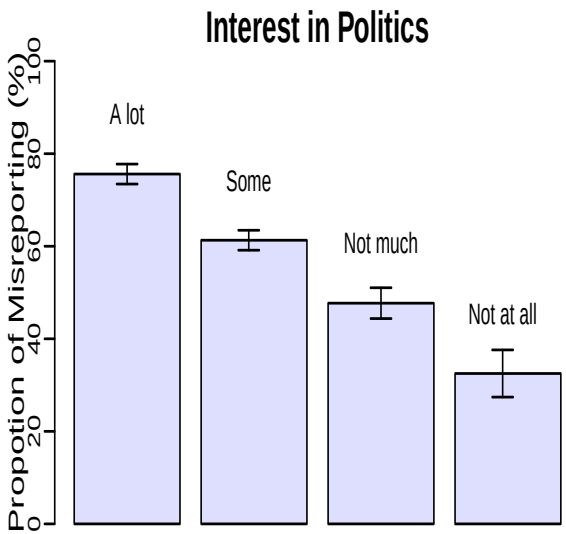
- Our match rate is lower than the registration rate based on voter file
- However, we find that matched non-voters do lie at a high rate:
  - matched respondents who voted: 93.8% (s.e. = 0.36, N=32,841)
  - matched respondents who did not vote: 43.9% (s.e. = 1.50, N=3,618)
- Regression analysis:
  - binary response: respondent said voted but validation found otherwise
  - age, marriage, education, gender, race, income, partisanship, interests in politics, church attendance, ideology

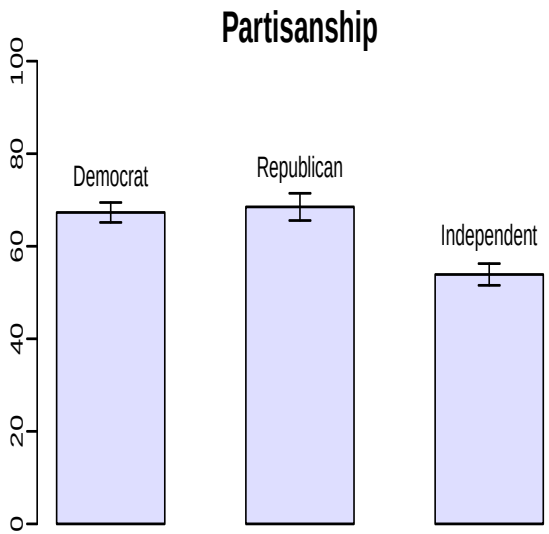
Who Misreports?

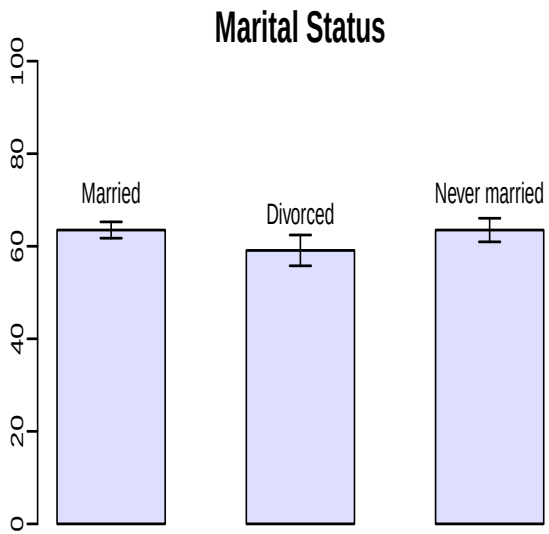








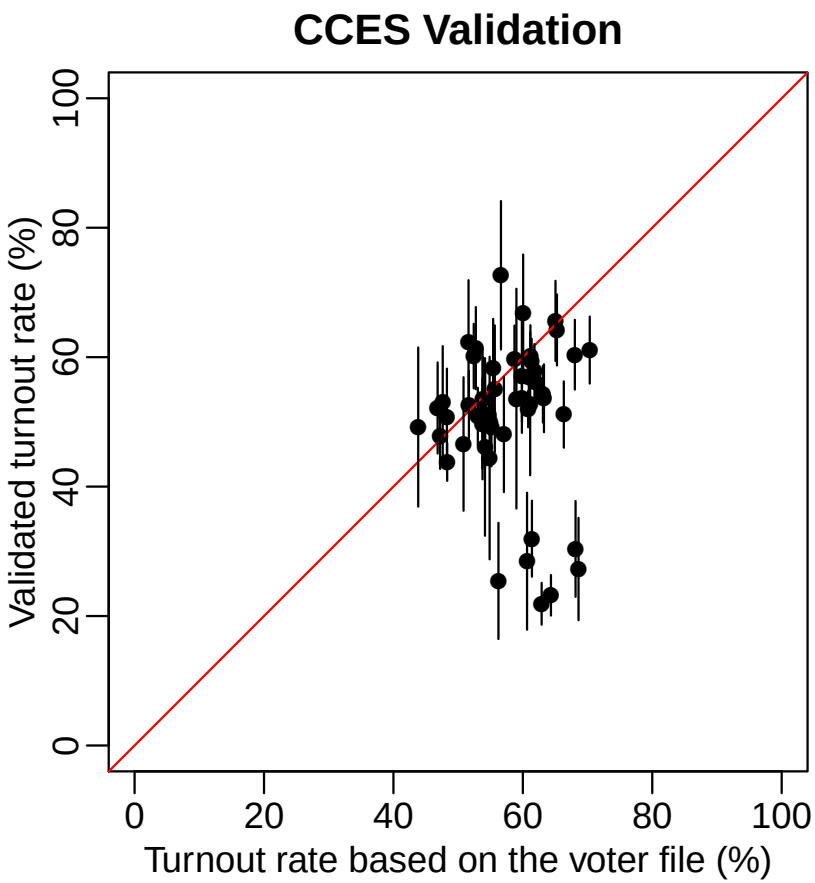


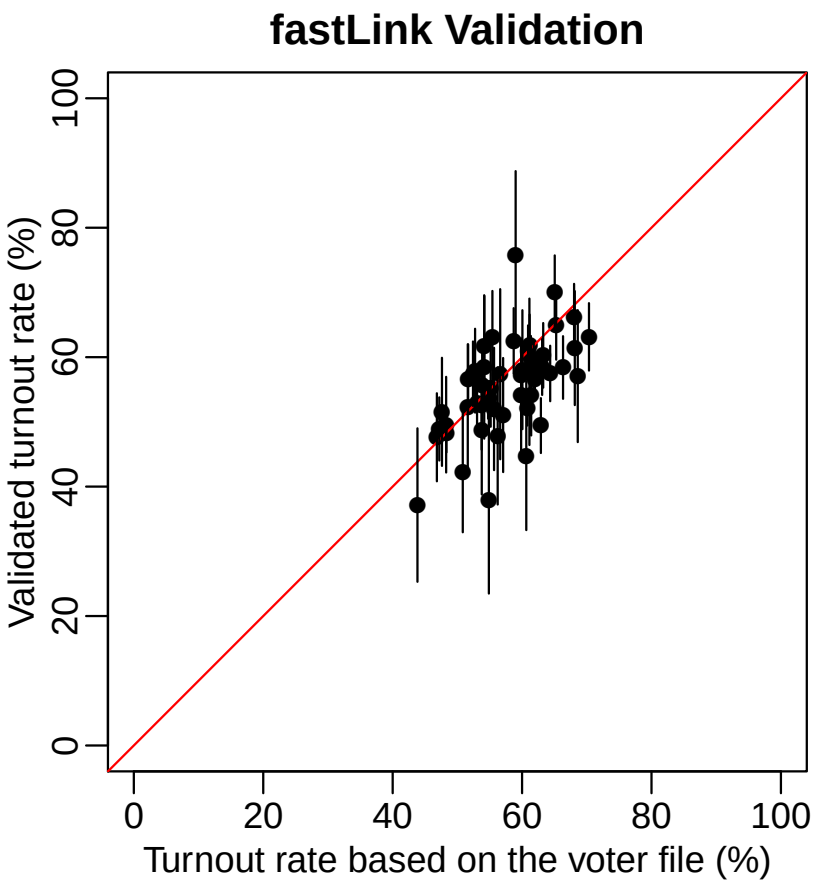


Comparison with CCES Turnout Validation

		Common matches	CCES only	fastLink only	Overall
Validated Turnout	L2	73.08 (0.34)	7.70 (0.19)	25.72 (0.42)	54.11 (0.31)
	CCES	71.01 (0.35)	9.84 (0.23)	0.00	49.87 (0.34)
Proportion of Misreporting	L2	4.95 (0.17)	12.80 (0.26)	8.48 (0.32)	6.33 (0.16)
	CCES	6.58 (0.19)	5.12 (0.19)	25.84 (0.46)	27.35 (0.29)
Number of respondents		34627	7877	8394	64600

State-level Comparison





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
- We offer a fast, principled, and scalable probabilistic merging method
- Open-source software `fastLink` available at CRAN
- Application: controversy regarding bias in self-reported turnout
  - Previous turnout validations relied upon proprietary algorithms
  - We merge CCES with a nationwide voter file using fastLink
  - fastLink yields high-quality matches and recovers actual turnout rate
  - Bias is driven by misreporting rather than nonresponse
  - fastLink outperforms a state-of-art proprietary method
- Turnout validation results for the 2016 ANES forthcoming