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

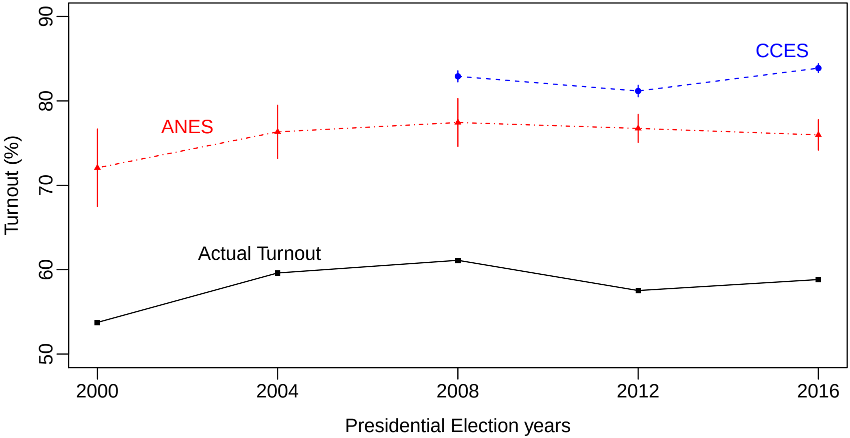
Ted Enamorado Kosuke Imai

Princeton University

Seminar at the Center for the Study of Democratic Politics
Princeton University

March 8, 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:
    - 2016 American National Election Study (ANES)
    - 2016 Cooperative Congressional Election Study (CCES)
2. Probabilistic method of record linkage and fastLink (with Ben Fifield)
3. Simulation study to compare fastLink with deterministic methods
    - fastLink effectively handles missing data and measurement error
4. Empirical findings:
    - fastLink recovers the actual turnout
    - clerical review helps with the ANES but not with the CCES
    - Bias of self-reported turnout appears to be largely driven by misreporting
    - fastLink performs at least as well as a state-of-art proprietary method

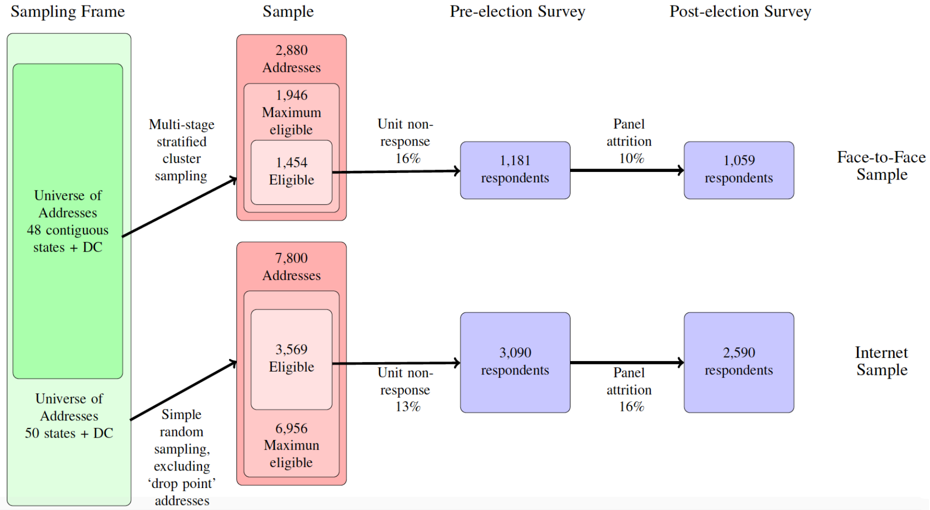
The 2016 US Presidential Election

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

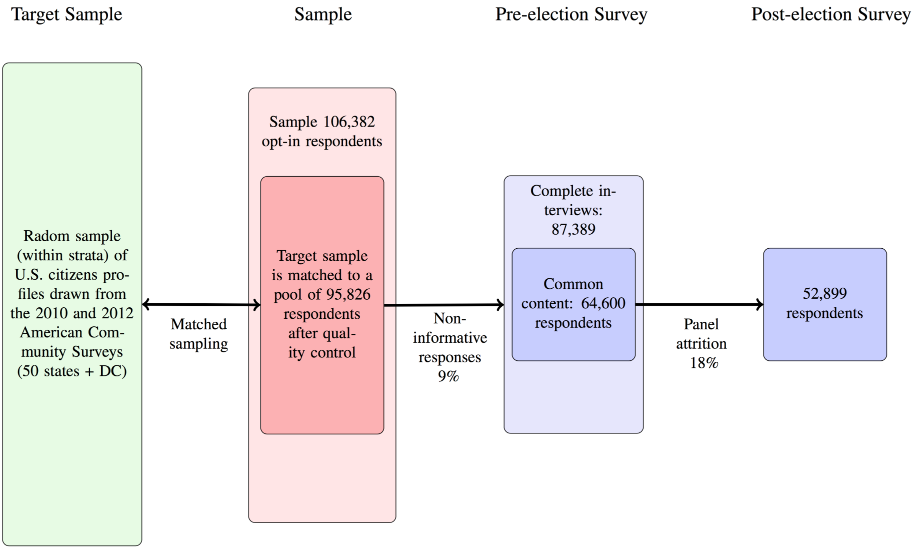
- Two validation exercises:
  1. The 2016 American National Election Study (ANES)
  2. The 2016 Cooperative Congressional Election Study (CCES)
- We merge the survey data with a nationwide voter file

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

ANES Sampling Design



CCES Sampling Design

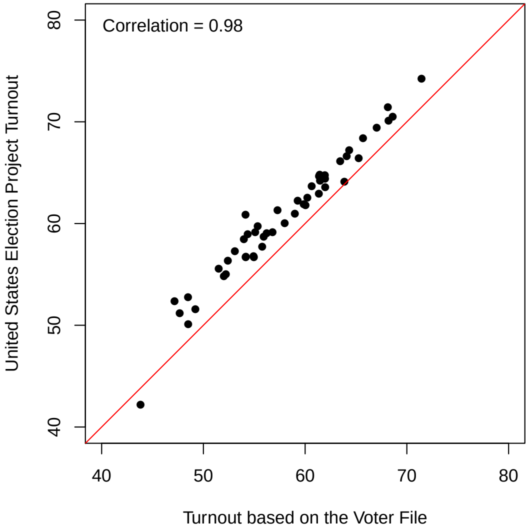


Bias of Self-reported Turnout and Registration Rates

	ANES	**CCES**	Election project	Voter files all	Voter files active	CPS
Turnout rate	75.96 (0.92)	83.79 (0.27)	58.83	57.55		61.38 (1.49)
Registration rate	89.18 (0.71)	91.93 (0.21)		80.37	76.57	70.34 (1.40)
Pop. size (millions)	224.10	224.10	232.40	227.60	227.60	224.10

- Based on the ANES sampling and CCES pre-validation weights
- Target population
  - ANES (face-to-face): US citizens of voting age in 48 states + DC
  - ANES (internet) / CCES: US citizens of voting age in 50 states + 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 with the nationwide voter file using name, age, gender, and address:
  1. 4,271 ANES respondents
  2. 64,600 CCES respondents
- Standardization:
  1. Name: first, middle, and last name
     - ANES: Missing (1.5%), Use of initials (0%), Complete (0.4%)
     - CCES: Missing (2.7%), Use of initials (5.9%), Complete (91.4%)
  2. Address: house number, street name, zip code, and apartment number
     - ANES: Complete (100%)
     - CCES: 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
    1. ANES: from 48k (HI/Female) to 108 million pairs (CA/Female)
    2. CCES: from 3 million (WY/Male) to 25 billion pairs (CA/Male)
  - Apply fastLink 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): $\gamma(i,j)$

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

- 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:

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

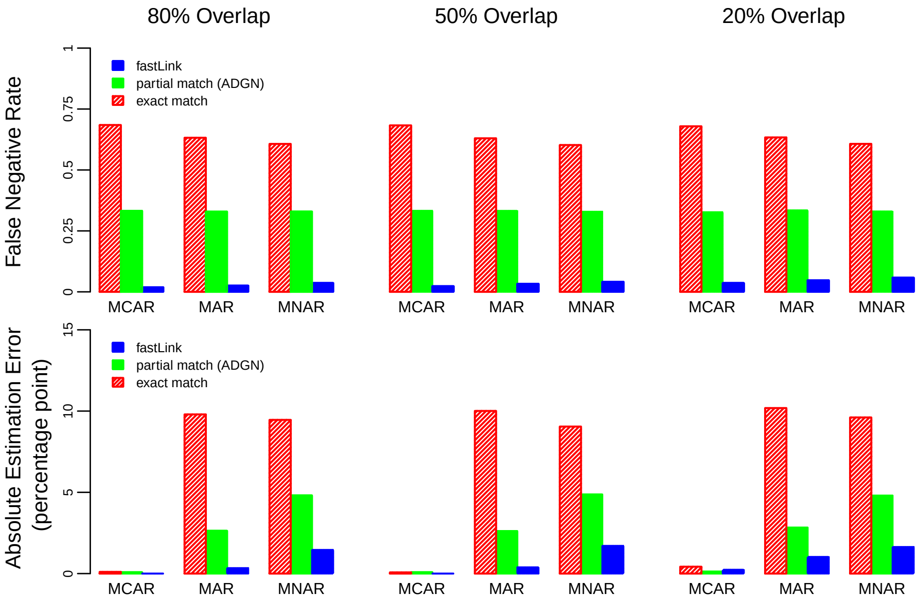
  where $\lambda = P(M_{ij}=1)$ is the proportion of true matches and $\pi_{km\ell} = \Pr(\gamma_k(i,j)=\ell \mid M_{ij}=m)$

- 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

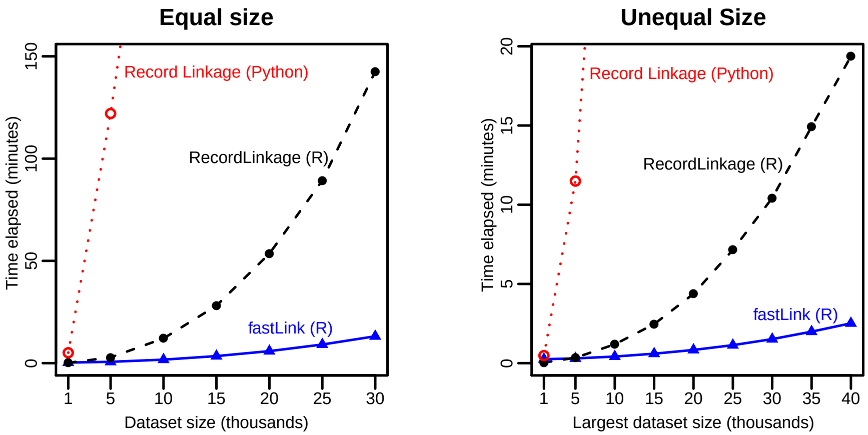
Simulation Studies

- 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)
  4. clerical review (for both matches and non-matches)

- Our analysis uses posterior match probability as well as ANES and CCES (pre-validation) sampling weights

Match Rate as an Estimate of Registration Rate

	Pre-election		Post-election		Registration rate		
		clerical		clerical	Voter file		
	fastLink	review	fastLink	review	all	active	CPS
ANES	76.54 (0.63)	68.79 (0.71)	77.15 (0.67)	69.85 (0.76)	80.37	76.57	70.34 (1.40)
CCES	66.60 (0.18)	58.59 (0.19)	70.52 (0.19)	63.57 (0.21)	80.37	76.57	70.34 (1.40)

- Registration rate is difficult to compute:
  - only some states classify voters as "active" or "inactive"
  - definition differs by states
- Clerical review
  - appears to work for the ANES
  - may have introduced false negatives for the CCES

Validated Turnout Rates

	Pre-election		Post-election		Actual turnout	
	fastLink	clerical review	fastLink	clerical review	Voter file	Election project
ANES	63.59 (0.91)	58.09 (0.93)	64.97 (0.96)	59.78 (1.00)	57.55	58.83
CCES	54.11 (0.31)	48.50 (0.31)	55.67 (0.37)	50.25 (0.37)	57.55	58.83

- fastLink plus clerical review works well for the ANES
- fastLink alone works better for the CCES

Validated Turnout by Response Category

		Not registered	Registered		Attrition
			Did not Vote	Voted	
ANES	fastLink	8.11	14.45	81.74	55.66
		(1.58)	(1.74)	(0.86)	(2.41)
	Clerical review	0.90	5.97	77.44	48.27
		(0.78)	(1.21)	(0.99)	(2.41)
CCES	fastLink	16.37	10.15	73.05	24.02
		(0.84)	(0.73)	(0.28)	(0.60)
	Clerical review	8.04	4.67	68.66	16.44
		(0.73)	(0.59)	(0.30)	(0.51)

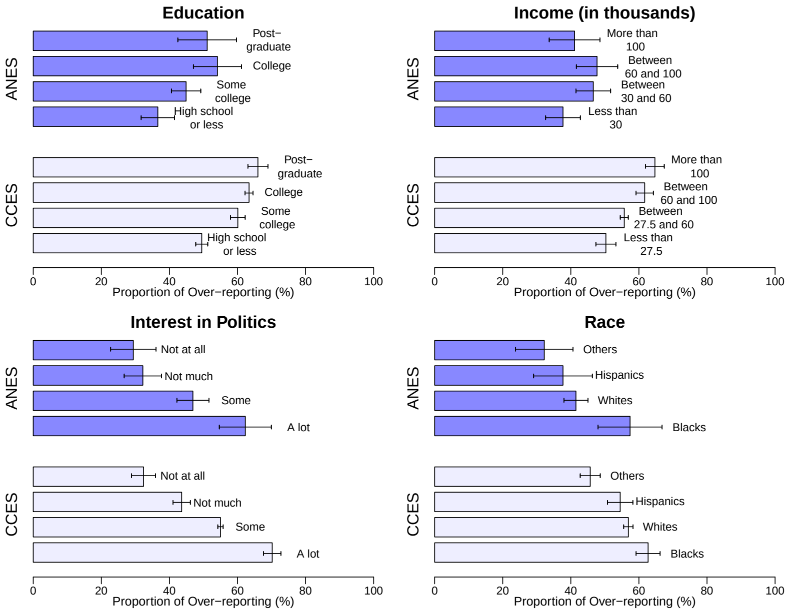
- Over-reporting is important: many are in the "Voted" category
- Attrition is a problem for the CCES, but not for the ANES

Do Voters Misreport Turnout?

- Berent, Krosnick, and Lupia (2016) argue that voters don't misreport:
  1. Poor quality of voter files and difficulty of merging
  2. Failure to match survey respondents who actually voted
  3. 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:
  1. matched respondents who voted:
     - ANES: 95.68% ($s.e.$=0.50, N=3,436)
     - CCES: 92.70% ($s.e.$=0.36, N=33,329)
  2. matched respondents who did not vote:
     - ANES: 33.66% ($s.e.$=3.01, N=378)
     - CCES: 43.49% ($s.e.$= 1.50, N=3,901)

Who Misreports?

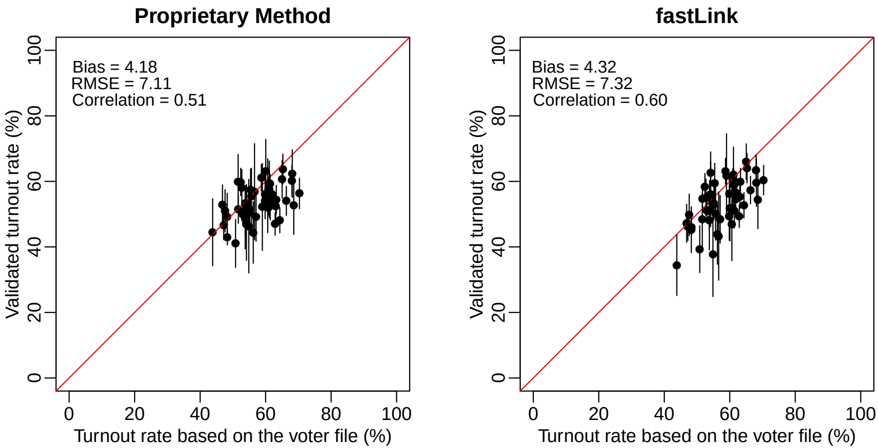


Comparison with CCES Turnout Validation

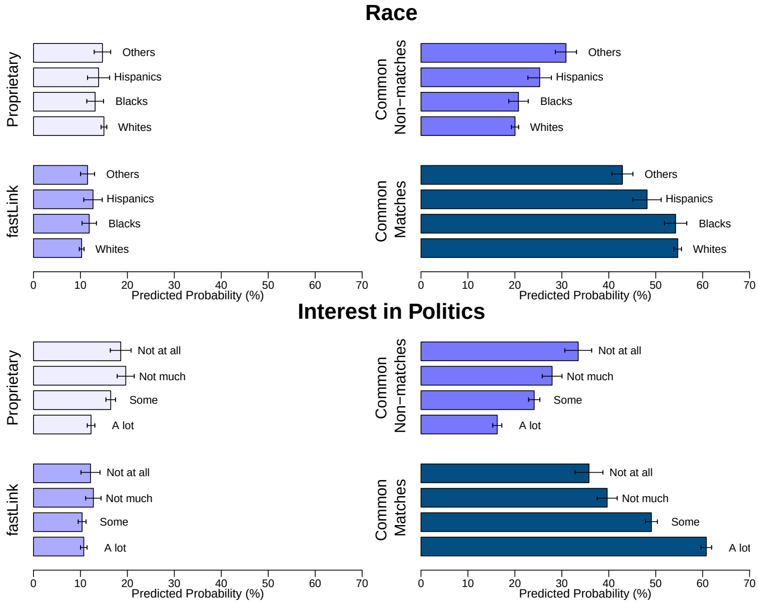
- Benchmark: 58.83 (election project) and 57.55 (voter file)

		Common matches	CCES only	fastLink only	Overall
Validated Turnout	L2	70.34	8.63	23.16	54.11
		(0.35)	(0.21)	(0.43)	(0.31)
	CCES	68.48	10.14	0.00	52.85
		(0.35)	(0.23)		(0.34)
Number of respondents		34,344	8,773	6,678	64,600

State-level Comparison



Predicting Match Type



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 ANES/CCES with a nationwide voter file using fastLink
  - fastLink yields high-quality matches and recovers actual turnout rate
  - Bias appears to be driven by misreporting rather than nonresponse
  - Probabilistic merge is robust to missing and invalid entries
  - Clerical review may introduce false negatives for messy data
  - fastLink performs as well as a state-of-art proprietary method