ECPR Methods Summer School: Big Data Analysis in the Social Sciences

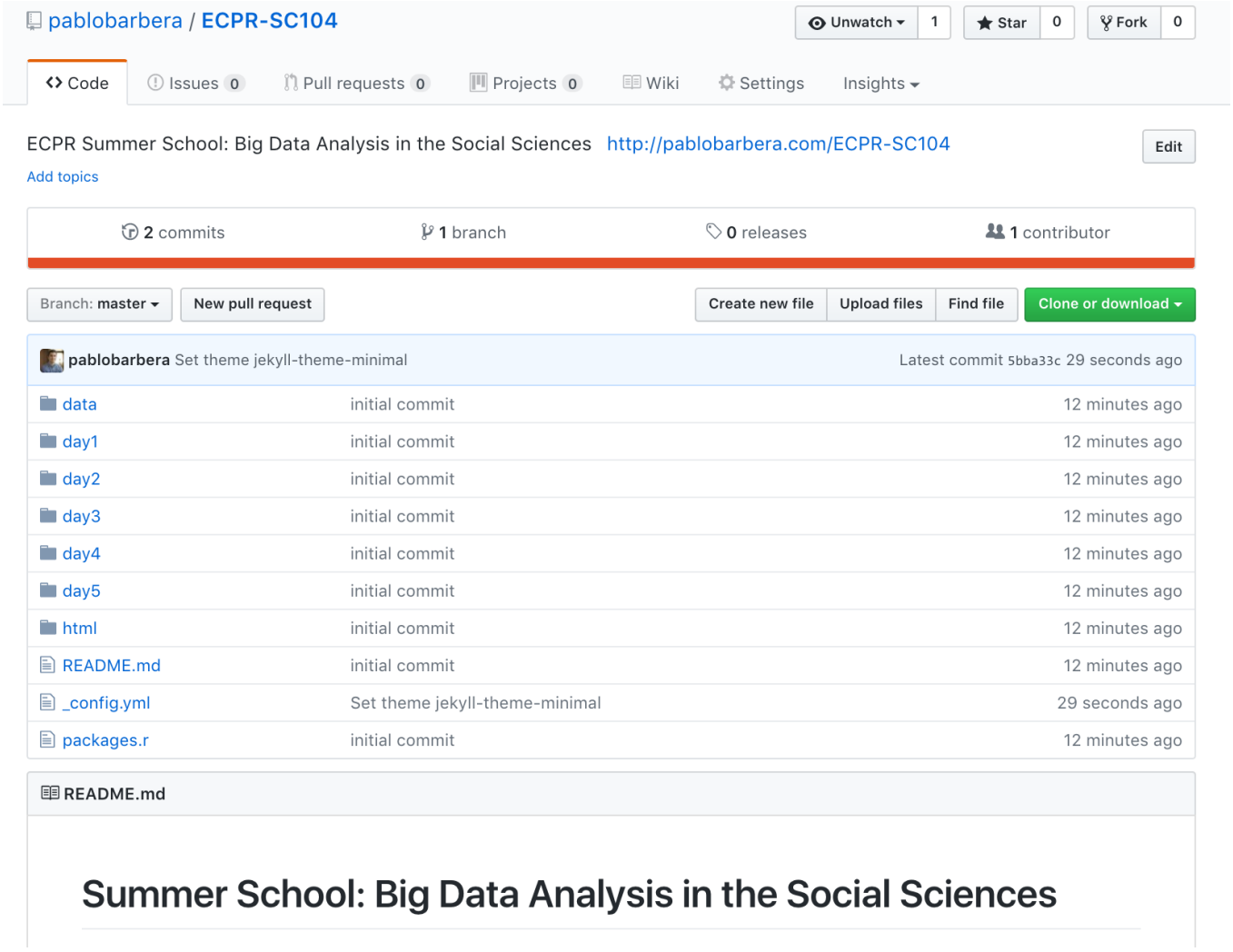
Pablo Barberá

School of International Relations
University of Southern California
pablobarbera.com

Networked Democracy Lab
www.netdem.org

Course website:
github.com/pablobarbera/ECPR-SC104

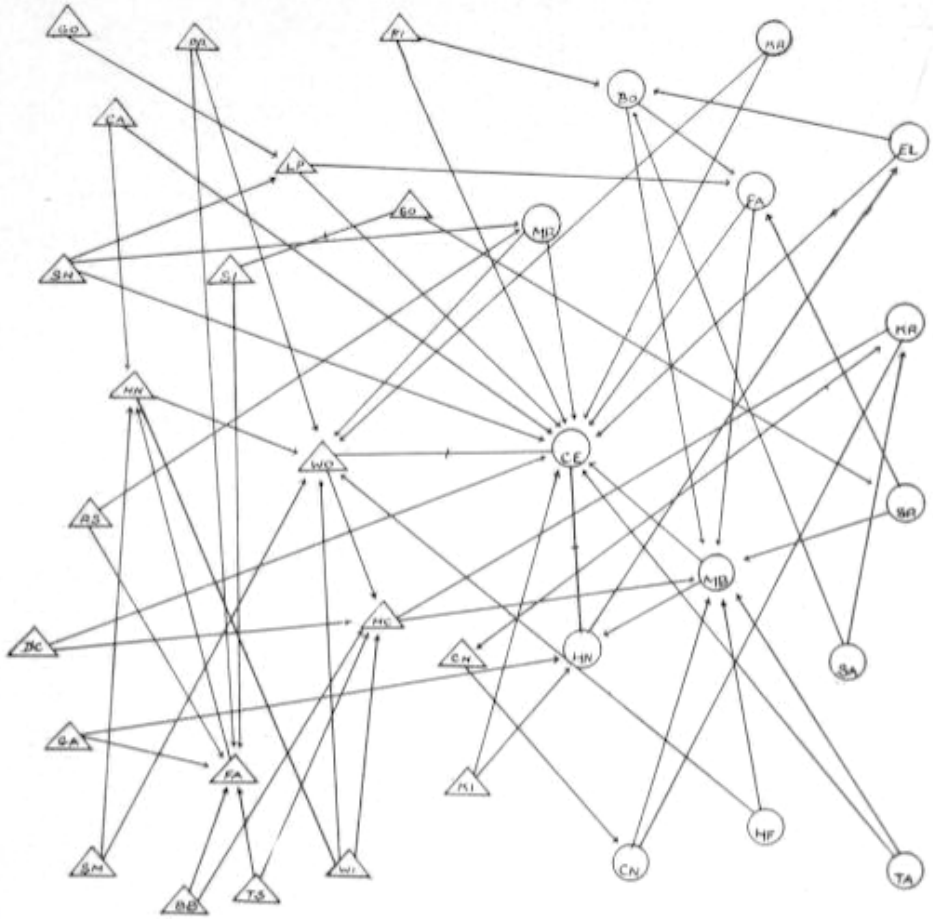
Course website



github.com/pablobarbera/ECPR-SC104

Discovery in large-scale networks

EVOLUTION OF GROUPS

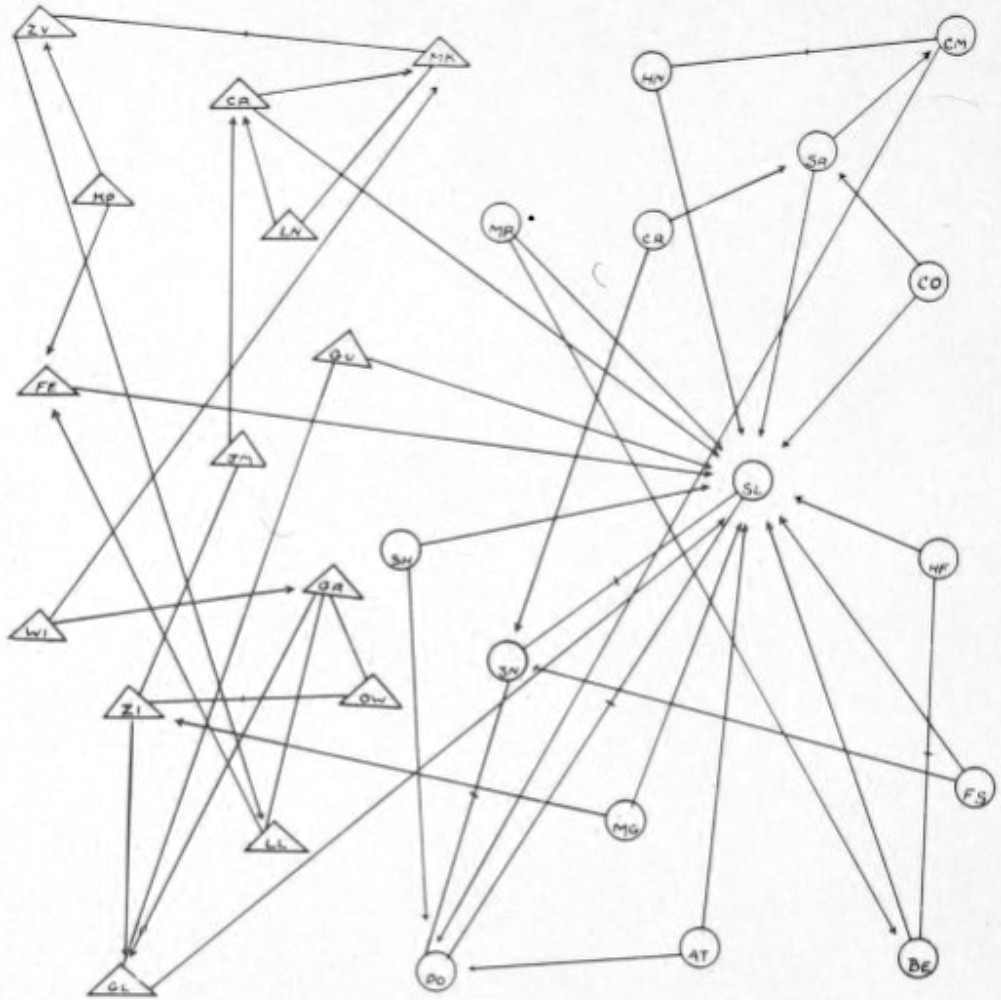


CLASS STRUCTURE, 1ST GRADE

21 boys and 14 girls. *Unchosen*, 18, GO, PR, CA, SH, FI, RS, DC, GA, SM, BB, TS, WI, KI, TA, HF, SA, SR, KR; *Pairs*, 3, EI-GO, WO-CE, CE-HN; *Stars*, 5, CE, WO, HC, FA, MB; *Chains*, 0; *Triangles*, 0; *Inter-sexual Attractions*, 22.

Moreno, “Who Shall Survive?” (1934)

EVOLUTION OF GROUPS

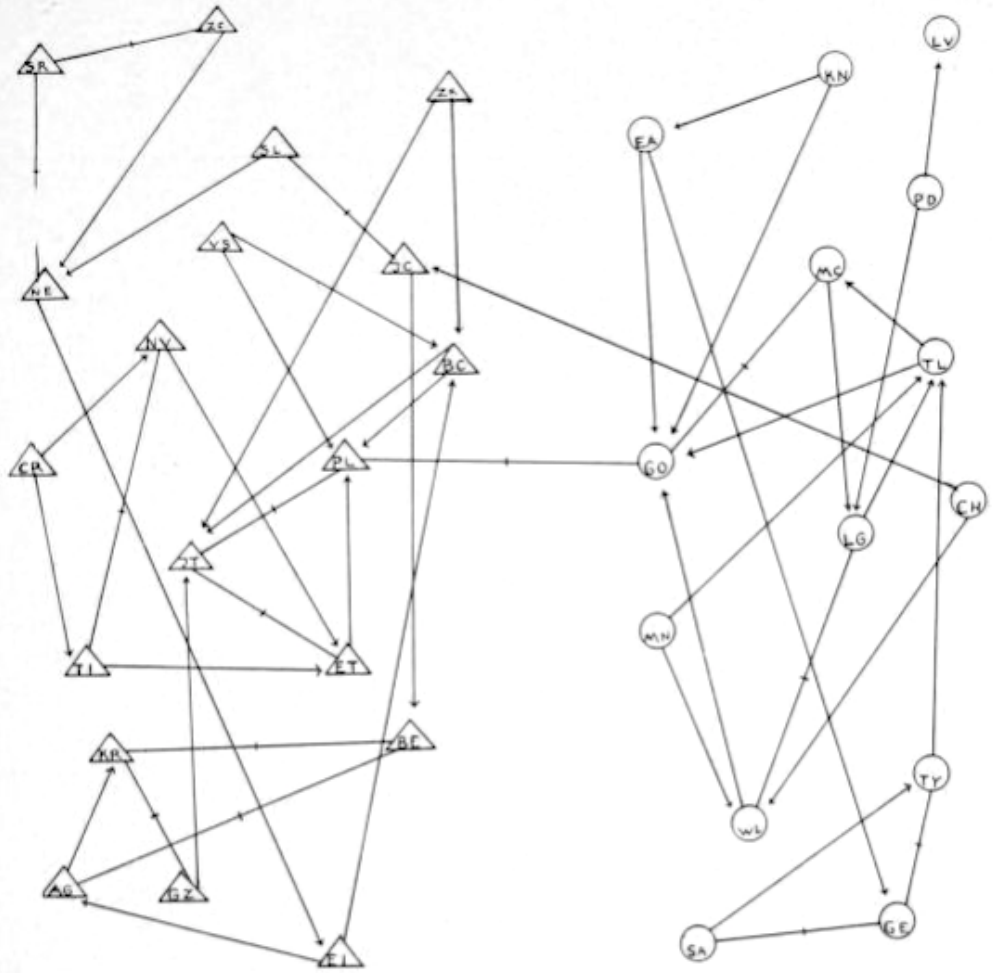


CLASS STRUCTURE, 2ND GRADE

14 boys and 14 girls. *Unchosen*, 9, WI, KP, MG, AT, FS, CN, CR, MR, SH; *Pairs*, 11, ZV-MK, MK-LN, OW-ZI, GR-LL, ZI-JM, HN-CM, SL-JN, JN-PO, PO-SL, HF-BE, GL-GU; *Stars*, 2, SL, PO; *Chains*, 0; *Triangles*, 1, SL-JN-PO; *Inter-sexual Attractions*, 5.

Moreno, “Who Shall Survive?” (1934)

EVOLUTION OF GROUPS

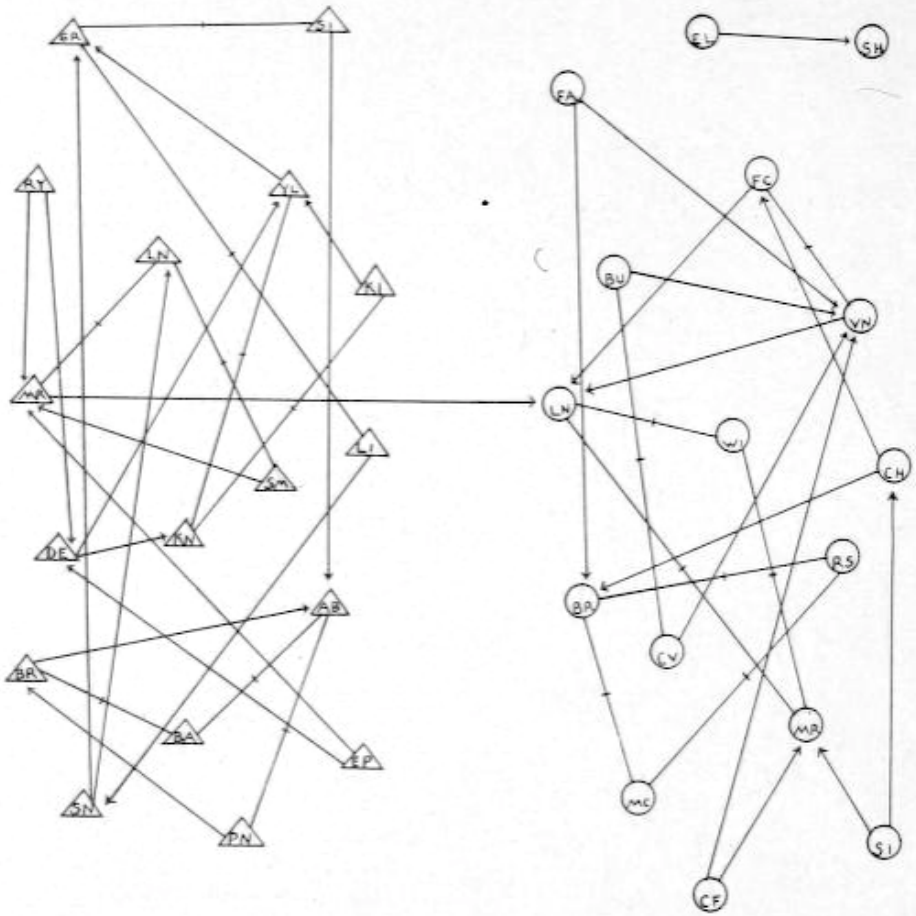


CLASS STRUCTURE, 3RD GRADE

19 boys and 14 girls. *Unchosen*, 7, VS, CR, CH, MN, PO, KN, ZK; *Pairs*, 14, SR-ZC, SR-NE, SL-JC, NV-TI, PL-JT, JT-ET, KR-BE, BE-AG, RR-GZ, PL-GO, GO-MC, WL-LG, SA-GE, GE-TY; *Stars*, 3, GO, PL, JT; *Chains*, 1, ET-JT-PL-GO-MC; *Triangles*, 0; *Inter-sexual Attractions*, 3.

Moreno, “Who Shall Survive?” (1934)

EVOLUTION OF GROUPS



CLASS STRUCTURE, 4TH GRADE

17 boys and 16 girls. *Unchosen*, 6, EP, RY, EL, FA, SI, CF; *Pairs*, 17, GR-SI, GR-LI, MR-LN, LN-SM, YL-KN, AB-BA, BA-BR, KI-KN, AB-PN, FC-VN, BU-CV, LN-WI, LN-MR, BR-MC, BR-RS, WI-MR, MC-RS; *Stars*, 2, LN, VN; *Chains*, 0; *Triangles*, 2, BR-RS-MC; LN-WI-MR; *Inter-sexual Attractions*, 1.

Moreno, “Who Shall Survive?” (1934)

Basic concepts

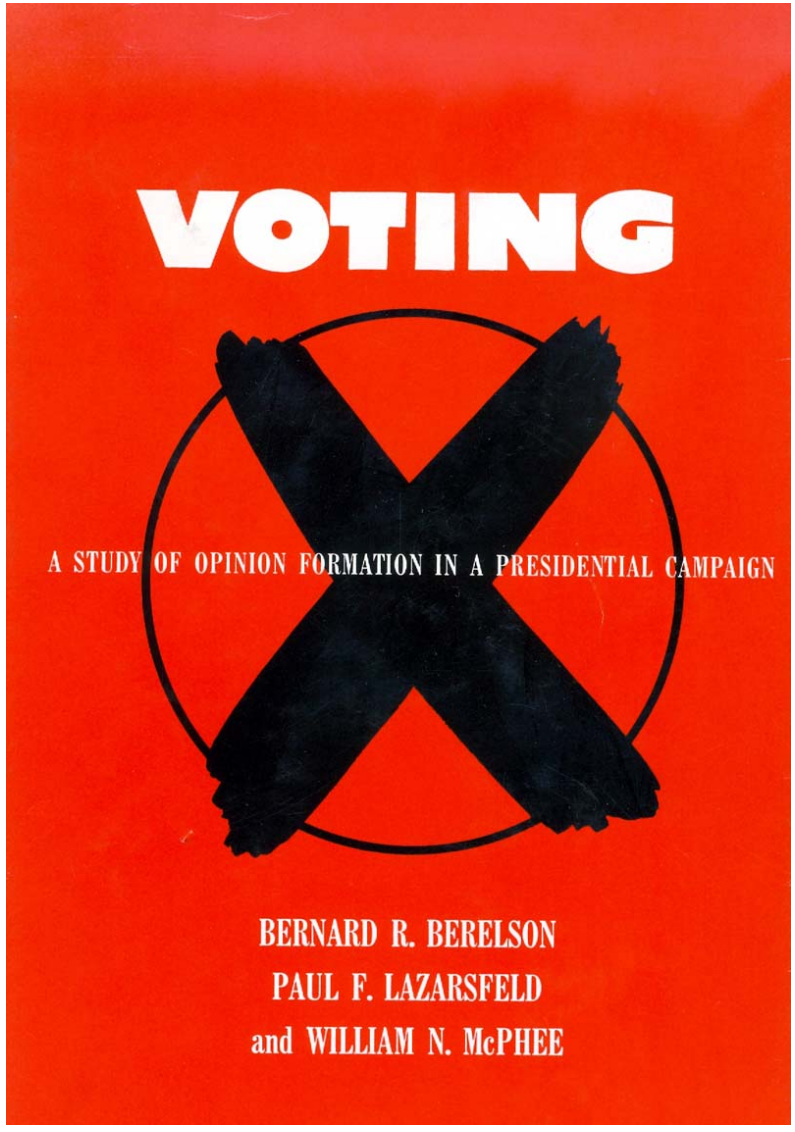
- Node (vertex): each of the units in the network
- Edge (tie): connection between nodes
  - Undirected: symmetric connection, represented by lines
  - Directed: imply direction, represented by arrows
- A network consists of a set of nodes and edges

Networks everywhere

- Classroom: students / friendships
- Twitter: users / retweets
- Academic literature: papers / citations
- Internet: websites / hyperlinks
- Trade: countries / trade flows
- Biology: neurons / connections
- Text: documents / cosine similarity

Political behavior is social

- Opinion formation as a *social process* (Berelson et al, 1954)



Political behavior is social

- Opinion formation as a *social process* (Berelson et al, 1954)
- *Voting is contagious* (Nickerson, 2008)

American Political Science Review Vol. 102, No. 1 February 2008
DOI: 10.1017/S0003055408080039

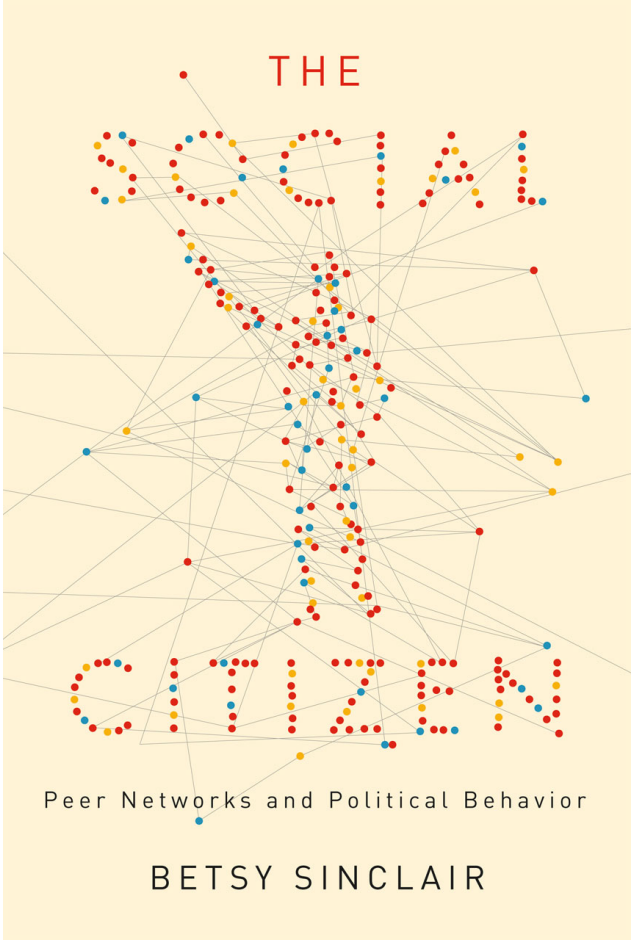
Is Voting Contagious? Evidence from Two Field Experiments

DAVID W. NICKERSON *University of Notre Dame*

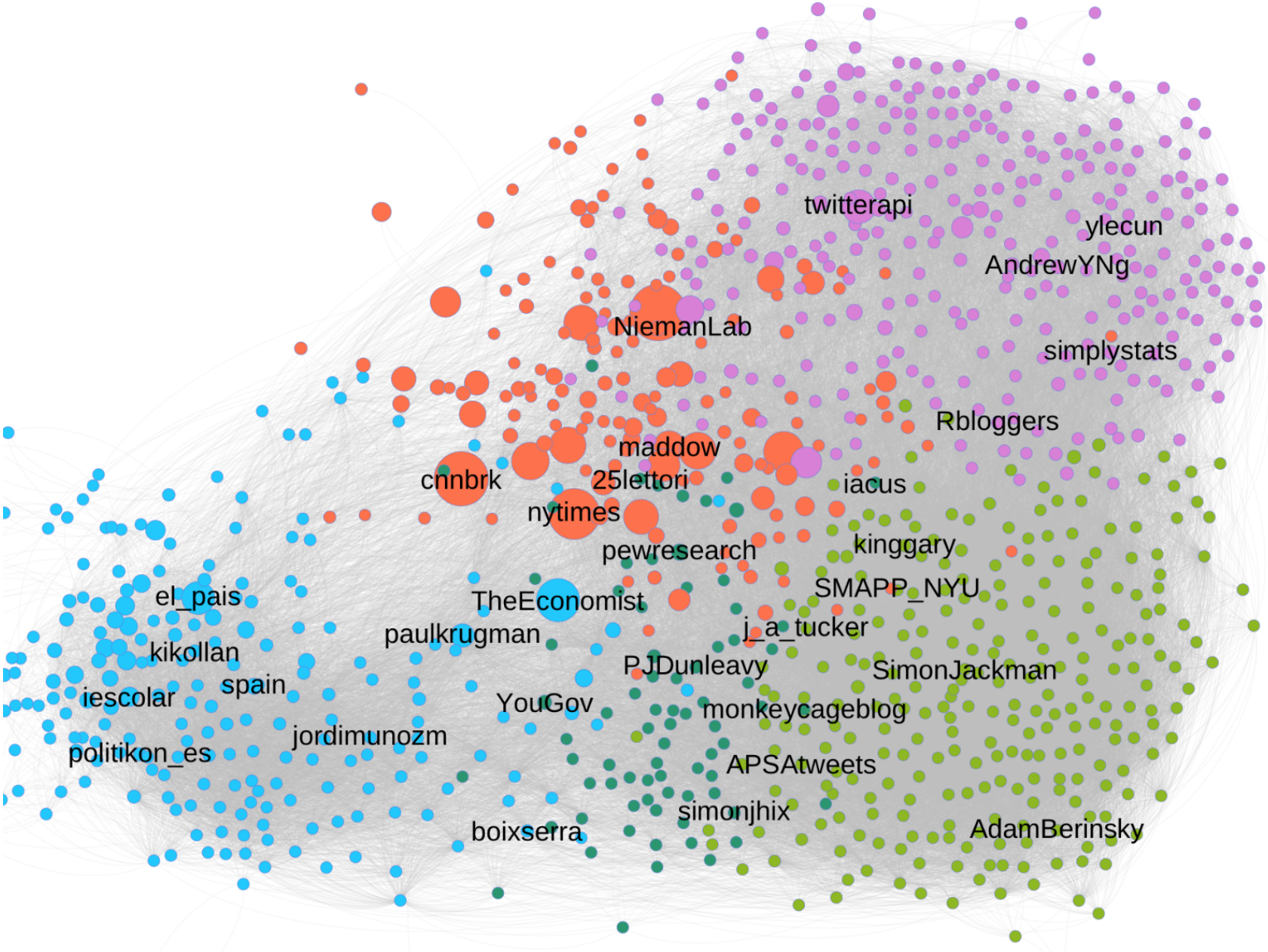
Members of the same household share similar voting behaviors on average, but how much of this correlation can be attributed to the behavior of the other person in the household? Disentangling and isolating the unique effects of peer behavior, selection processes, and congruent interests is a challenge for all studies of interpersonal influence. This study proposes and utilizes a carefully designed placebo-controlled experimental protocol to overcome this identification problem. During a face-to-face canvassing experiment targeting households with two registered voters, residents who answered the door were exposed to either a Get Out the Vote message (treatment) or a recycling pitch (placebo). The turnout of the person in the household not answering the door allows for contagion to be measured. Both experiments find that 60% of the propensity to vote is passed onto the other member of the household. This finding suggests a mechanism by which civic participation norms are adopted and couples grow more similar over time.

Political behavior is social

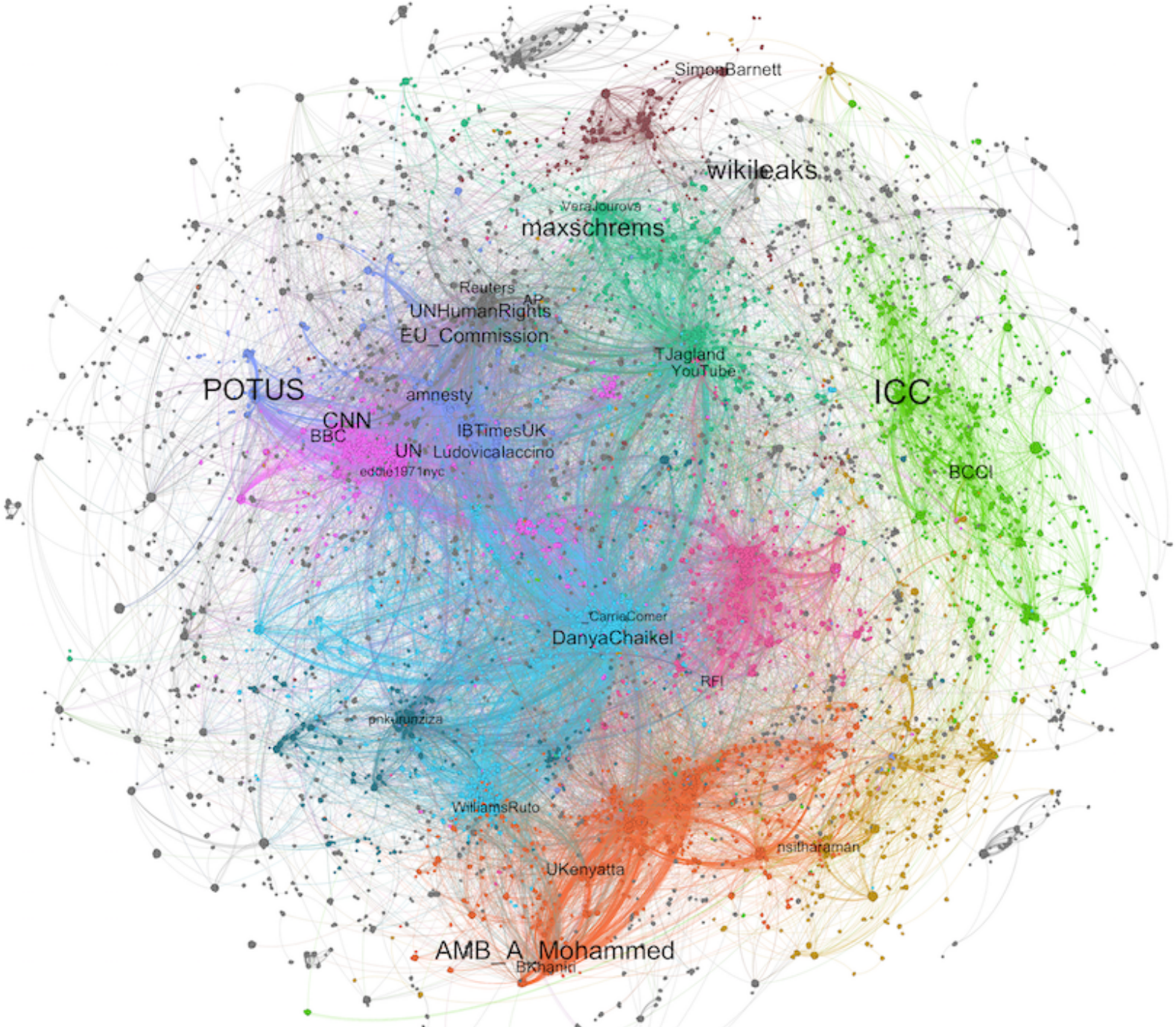
- Opinion formation as a *social process* (Berelson et al, 1954)
- *Voting is contagious* (Nickerson, 2008)
- The *social citizen* (Sinclair, 2012)



Latent structure of social networks



The dreaded *hairball*



Discovery in large-scale networks

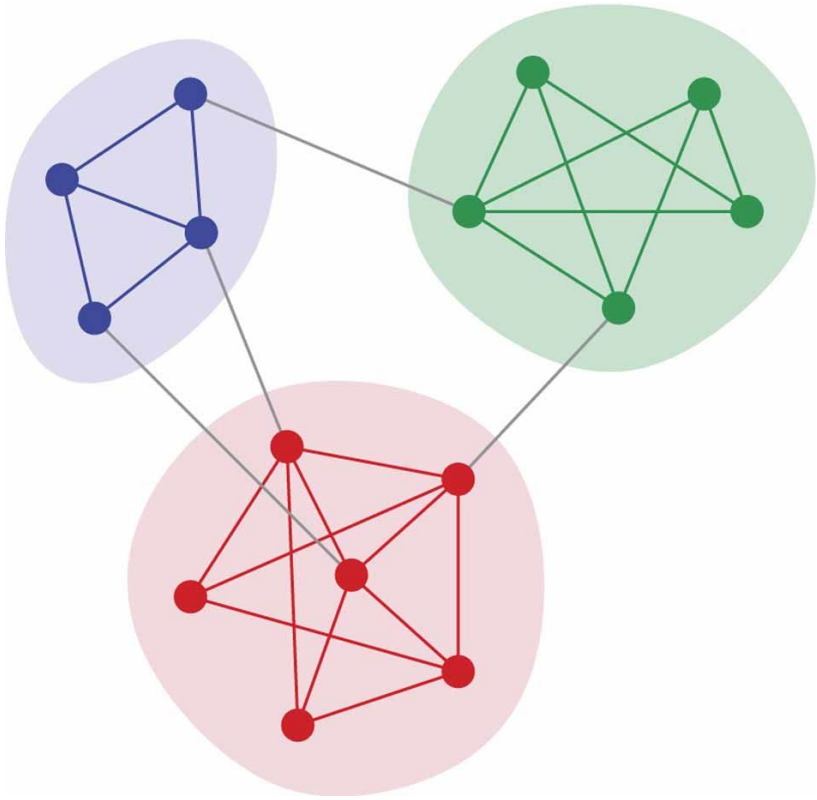
How to understand the structure of large-scale networks?

- Latent communities or clusters
  - **Community detection algorithms**
  - Finding groups of nodes that densely connected internally, more so than to the rest of the networks
  - Overlap with shared visible or latent similarities (homophily)
  - Also hierarchy: core-periphery detection
- Locating nodes on latent spaces
  - **Latent space models of networks**
  - Proximity on latent space (ideology) predicts existence of edges
  - Inference about latent positions based on multidimensional scaling of the adjacency matrix

Community detection

Community structure:

- Network nodes often cluster into tightly-knit groups with a high density of within-group edges and a lower density of between-group edges
- **Modularity score**: measures clustering of nodes compared to random network of same size
- Many different community detection algorithms based on different assumptions

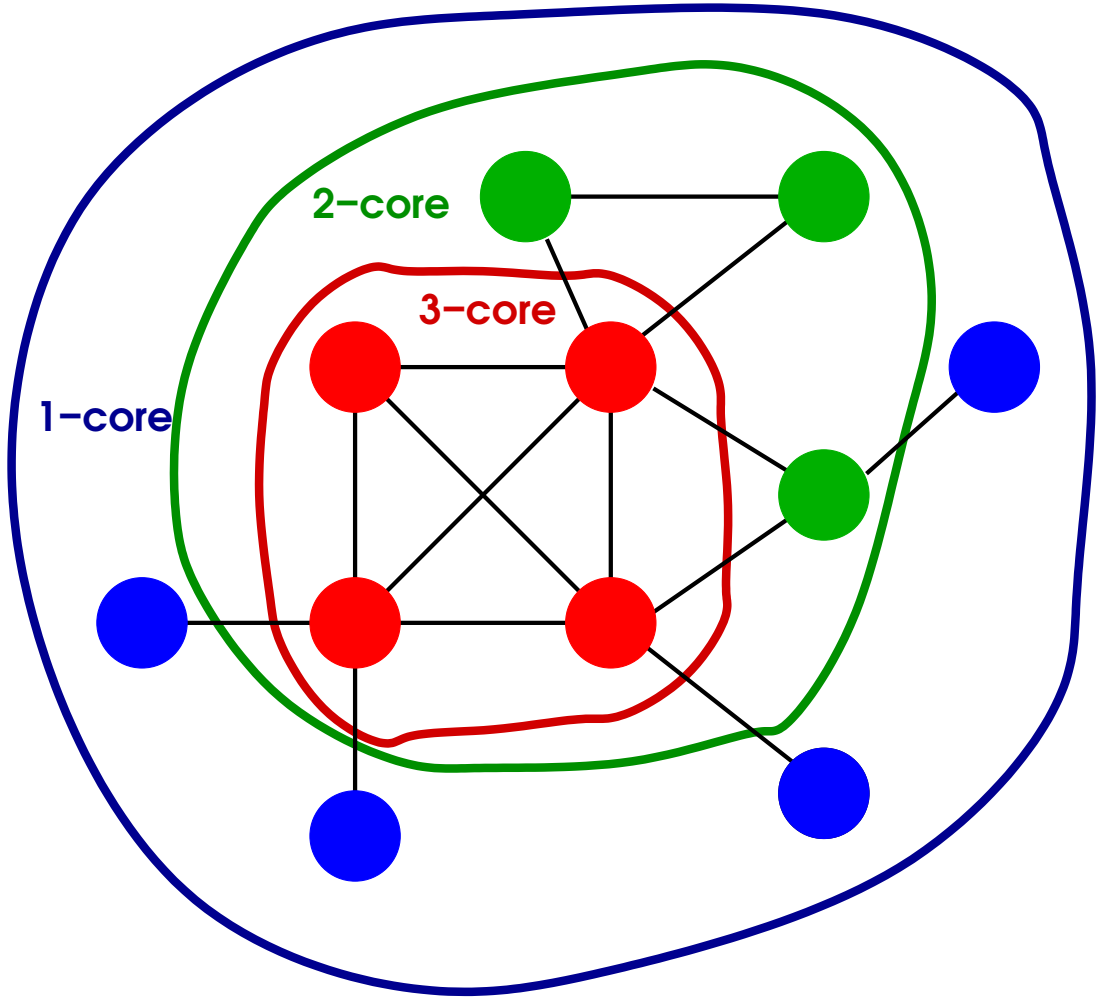


Source: Newman (2012)

Network hierarchy

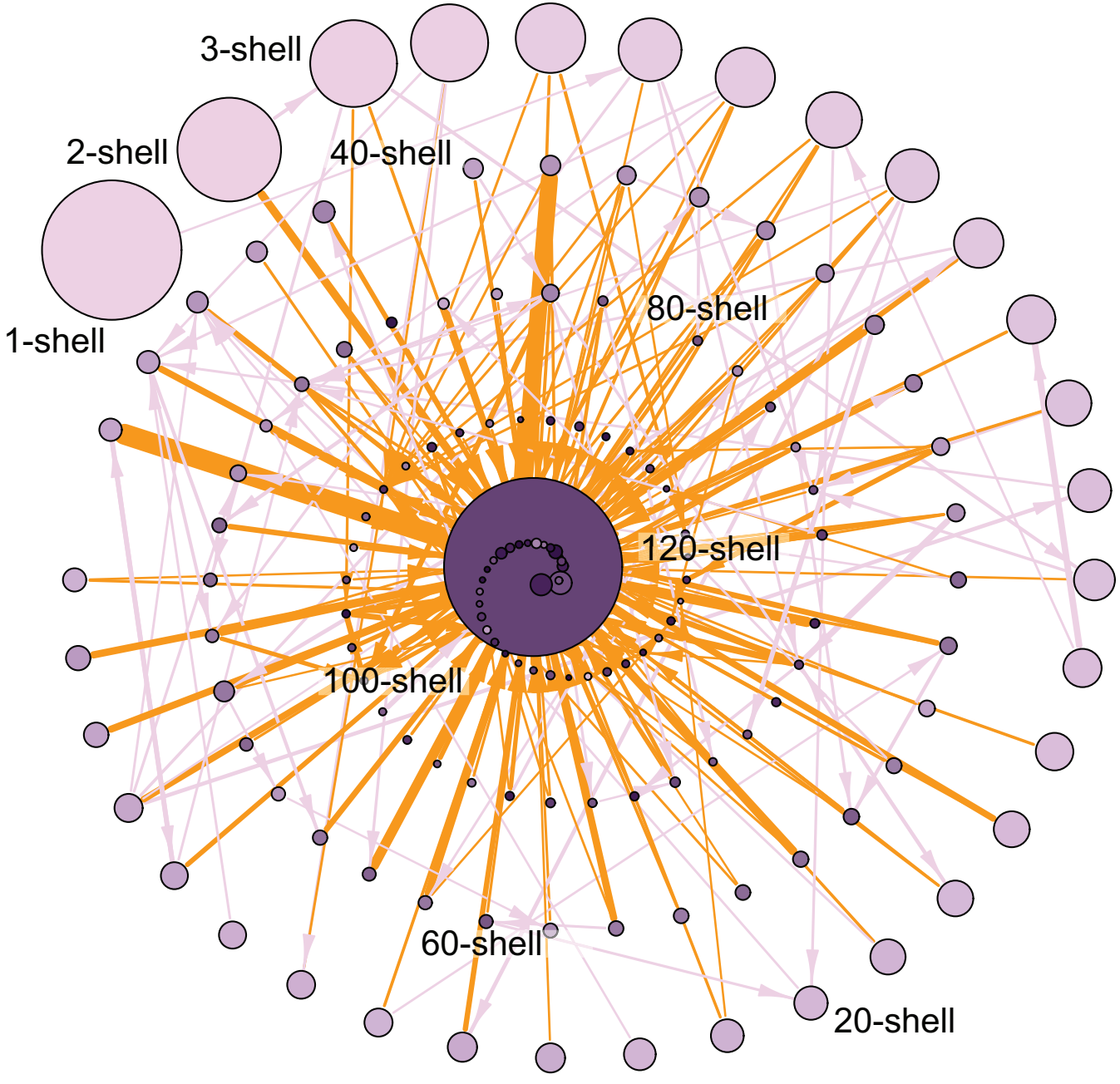
- **Intuition**
  - Large-scale networks have hierarchical properties
- **Network core:**
  1. *Centrality*: high relative importance in network
  2. *Connectivity*: many possible distinct paths between individuals
     (not captured by simple topological measures)
- **k-core decomposition**
  - Algorithm to partition a network in nested shells of connectivity
  - The k-core of a graph is the maximal subgraph in which every node has at least degree k
  - Many applications; scales well to large networks: $\mathcal{O}(n + e)$

k-core decomposition



Source: Alvarez-Hamelin et al, 2005

k-core decomposition of #OccupyGezi network



periphery

core

activity
(no. of tweets)

in Taksim

max 18%

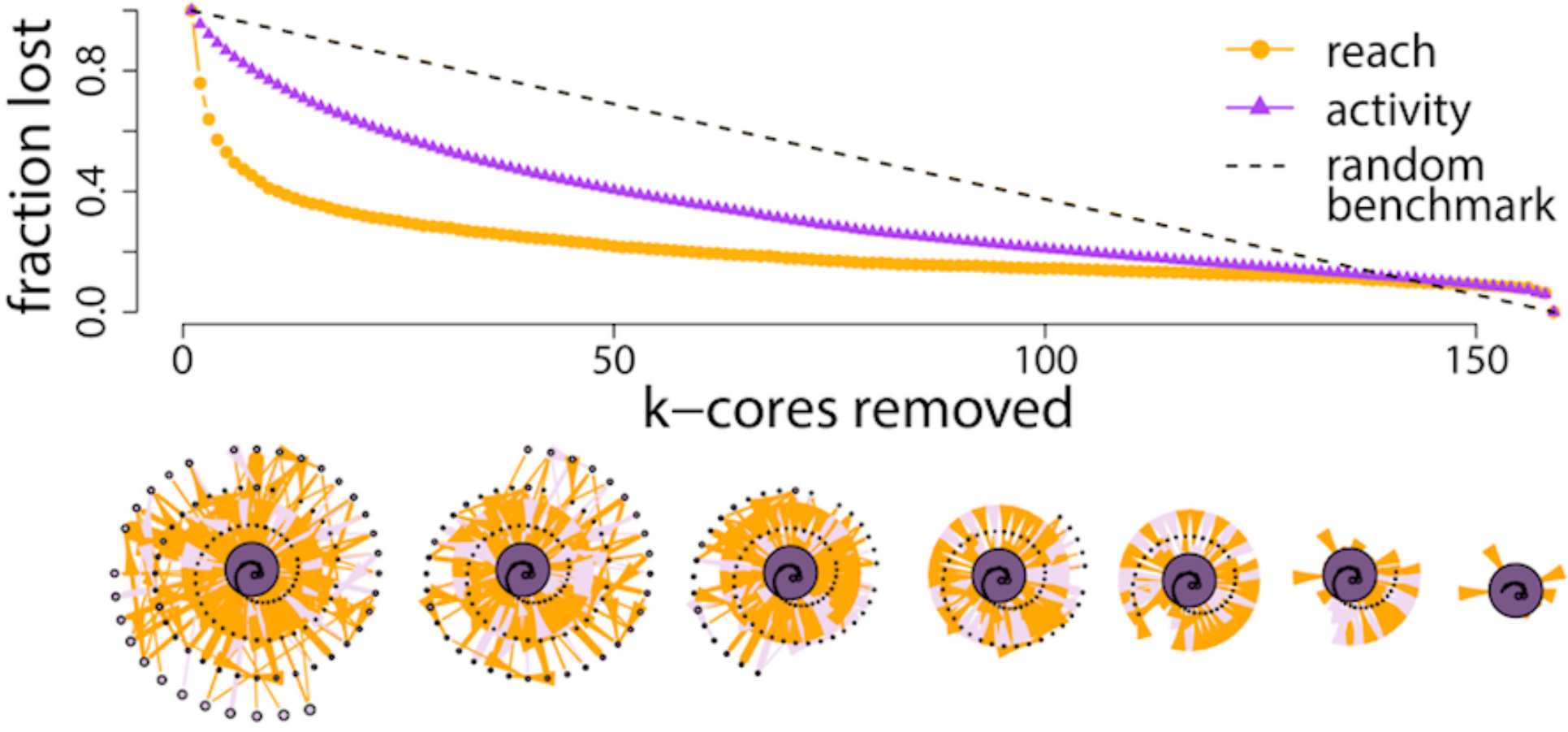
min .25%

RTs

periphery to core

periphery to periphery

Relative importance of core and periphery



reach: aggregate size of participants' audience

activity: total number of protest messages published (not only RTs)

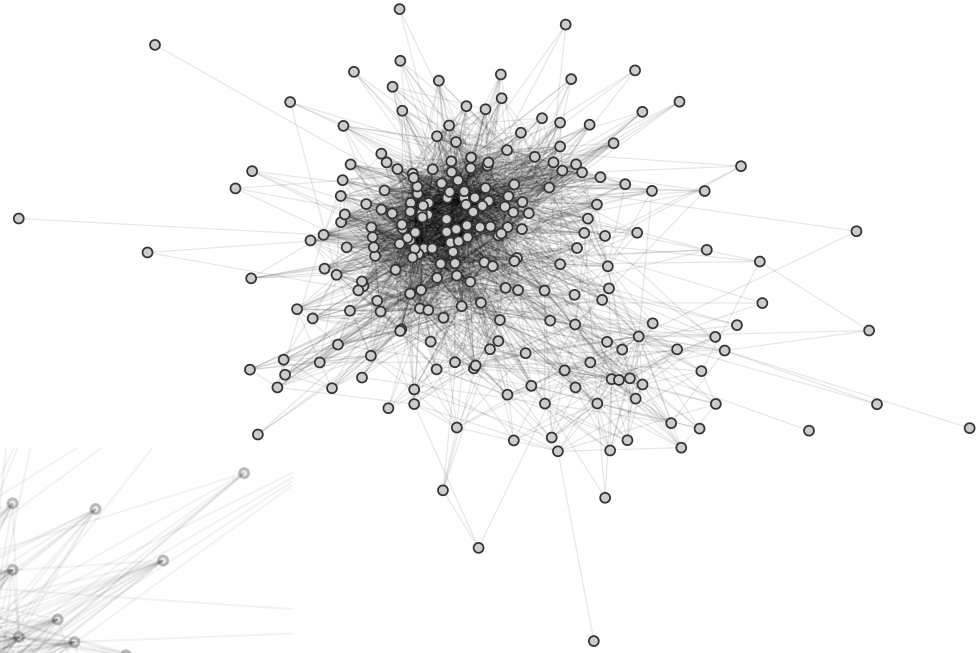
Latent space models

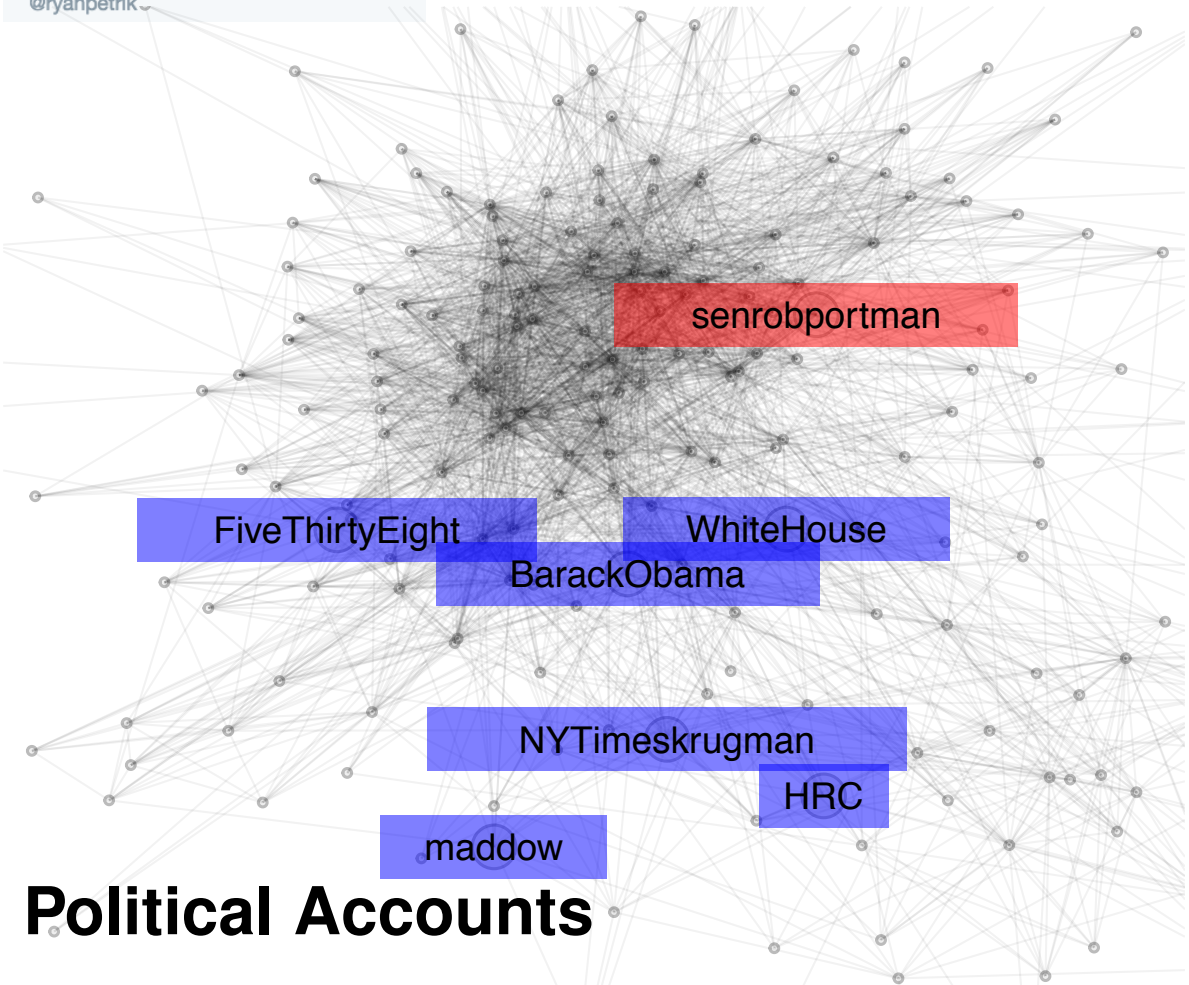
Spatial models of social ties (Enelow and Hinich, 1984; Hoff *et al*, 2012):

- Actors have unobserved positions on latent scale
- Observed edges are costly signal driven by similarity

Spatial *following* model:

- **Assumption**: users prefer to *follow* political accounts they perceive to be ideologically close to their own position.
- Following decisions contain information about allocation of scarce resource: attention
- Selective exposure: preference for information that reinforces current views
- Statistical model that builds on assumption to estimate positions of both individuals and political accounts





	BarackObama	WhiteHouse	GOP	maddow	FoxNews	HRC	...	pol. account m
ryanpetrik	1	1	0	1	0	1	...	
user 2	0	0	1	0	1	0	...	
user 3	0	0	1	0	1	0	...	
user 4	1	1	0	0	0	1	...	
user 5	0	1	0	0	0	1	...	
...								
user n	0	1	1	0	0	0	...	

Estimated ideology: $\theta_i = -1.05$

Political Accounts

Spatial following model

- Users' and political accounts' ideology (θ_i and ϕ_j) are defined as latent variables to be estimated.
- Data: "following" decisions, a matrix of binary choices (**Y**).
- Probability that user *i* follows political account *j* is

$$P(y_{ij} = 1) = \text{logit}^{-1}\left(\alpha_j + \beta_i - \gamma(\theta_i - \phi_j)^2\right) \quad ,$$

- with latent variables:

  θ_i measures *ideology* of user *i*

  ϕ_j measures *ideology* of political account *j*

- and:

  α_j measures *popularity* of political account *j*

  β_i measures *political interest* of user *i*

  γ is a normalizing constant

Estimation

- Likelihood function:

$$p(\mathbf{y}|\theta, \phi, \alpha, \beta, \gamma) = \prod_{i=1}^{n} \prod_{j=1}^{m} \text{logit}^{-1}(\pi_{ij})^{y_{ij}} (1 - \text{logit}^{-1}(\pi_{ij}))^{1-y_{ij}}$$

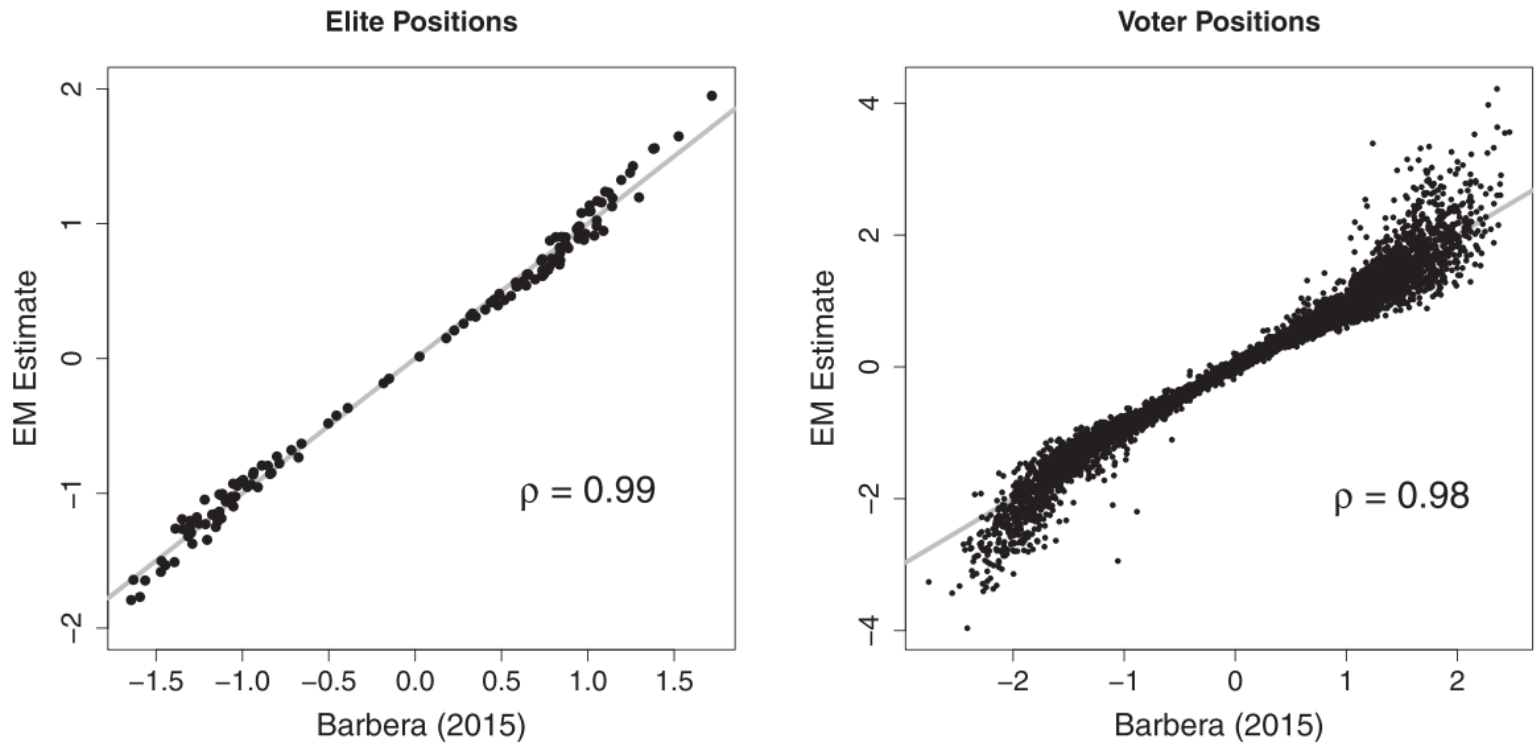
$$\text{where } \pi_{ij} = \alpha_j + \beta_i - \gamma(\theta_i - \phi_j)^2$$

- Intractable with maximum likelihood methods $\rightarrow$ MCMC.
- Two-stage estimation (*Political Analysis*, 2015):
  - First stage: HMC in *Stan* with random (dense) sample of **Y** to compute posterior distribution of *j*-indexed parameters.
  - Second stage: parallelized MH in *R* for rest of *i*-indexed parameters (assuming independence), on HPC.
- Identification:
  - Unit variance restriction on θ: $\theta_i \sim N(0, 1)$
  - Fix hyperparameters $\mu_\alpha = 0$ and $\sigma_\alpha = 1$

Estimation

Variational inference: EM algorithm (Imai *et al*, APSR, 2016)

FIGURE 14. Comparison of Our Ideal Point Estimates with Those of Barberá (2015)



Notes: Our ideal point estimates are based on the variational EM algorithm whereas those of Barberá (2015) are based on the Markov chain Monte Carlo algorithm as implemented in RStan. The left panel compares the two sets of ideal points for $J = 176$ political elites whereas the right panel conducts the same comparison for $N = 10{,}000$ voters. In both cases, the correlation between the two sets of estimates is very high. In addition, the variational EM algorithm only took 35 minutes to complete the estimation whereas RStan took 6.5 days to obtain 500 posterior draws on the same computer.

Estimation

Correspondence analysis (Greenacre, 1984; 2010)

1. Compute matrix of standardized residuals, $\mathbf{S}$:

$$\mathbf{S} = \mathbf{D}_r^{1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{D}_c^{1/2}$$

where $\mathbf{P} = \mathbf{Y} / \sum_{ij} y_{ij}$

$\mathbf{r}$, $\mathbf{c}$ are row/column masses: e.g. $r_i = \sum_j p_{ij}$

$\mathbf{D}_r = \text{diag}(\mathbf{r})$, $\mathbf{D}_c = \text{diag}(\mathbf{c})$

2. Calculate SVD of $\mathbf{S}$:

$$\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathbf{T} \text{ where } \mathbf{U}^\mathbf{T}\mathbf{U} = \mathbf{V}^\mathbf{T}\mathbf{V} = \mathbf{I}$$

3. Project rows and columns onto low-dimensional space:

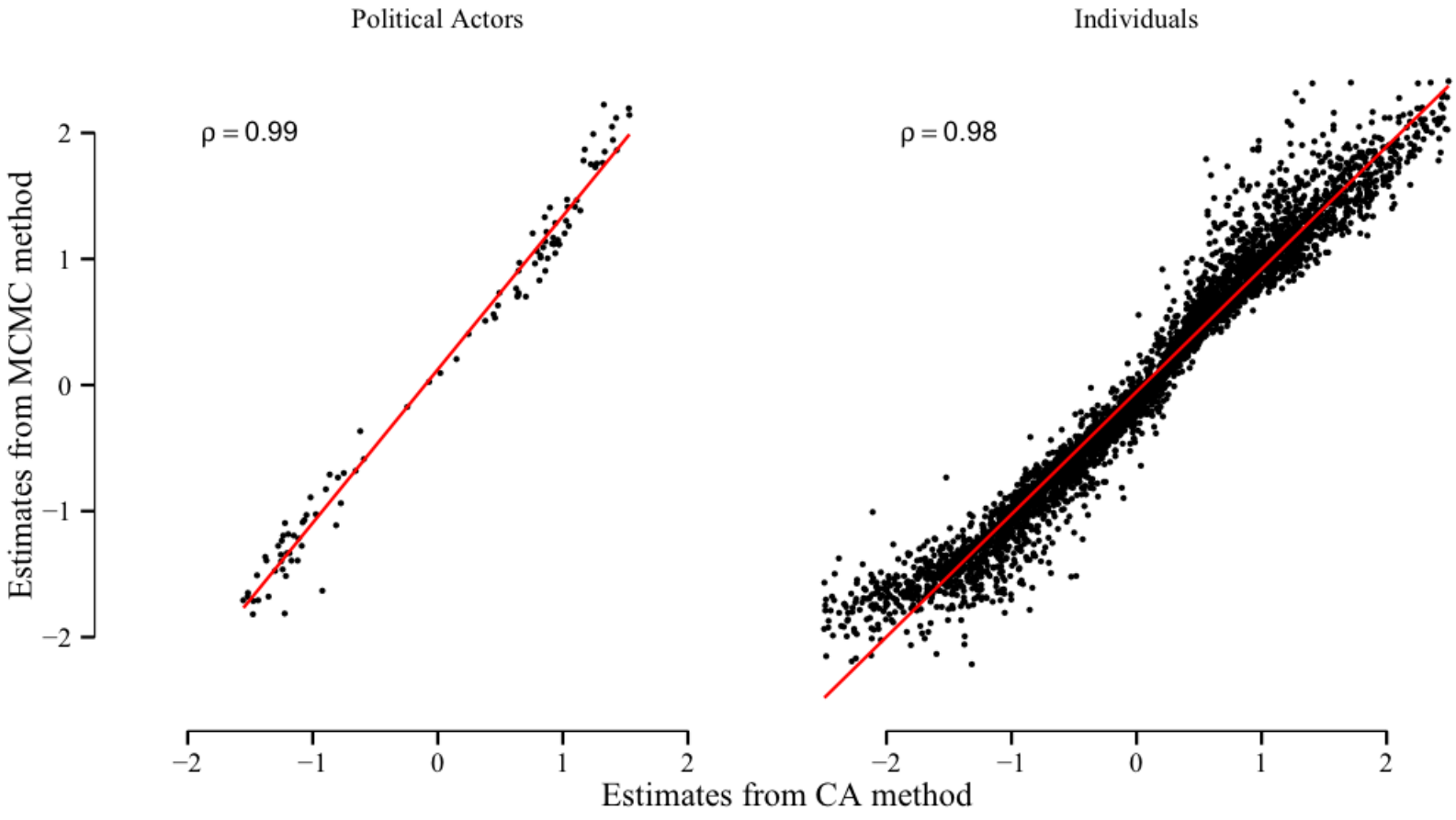
$$\theta = \mathbf{D}_r^{1/2}\mathbf{U} \text{ for rows (ordinary users)}$$

$$\phi = \mathbf{D}_c^{1/2}\mathbf{V} \text{ for columns (political accounts)}$$

Mathematically close to log-linear latent space model (Lowe, 2008) and computationally efficient, even with full matrix.

Estimation

Correspondence analysis (Greenacre, 1984; 2010)

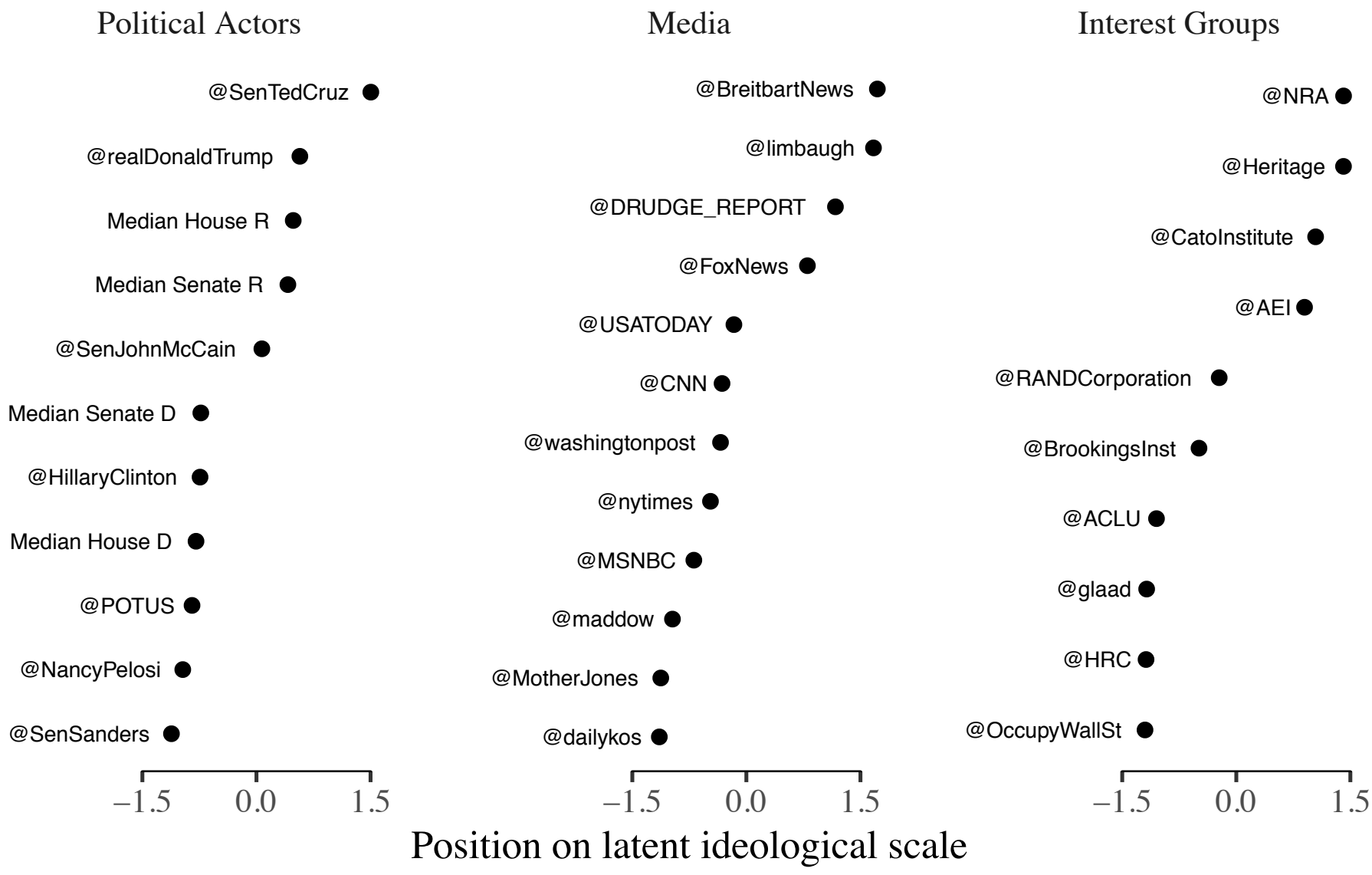


Runtime for N=10,000 users and J=173 political accounts:
MCMC = 6.5 days; EM = 35 minutes; CA = 1.88 seconds

Model validation

- m = list of 571 popular political accounts in U.S.
  - $\rightarrow$ Legislators, president, candidates, other political figures, media outlets, journalists, interest groups...
- n = followers of at least five of these accounts
  - $\rightarrow$ 12.6M users ($\sim$30% of U.S. users)

Face validity: political accounts

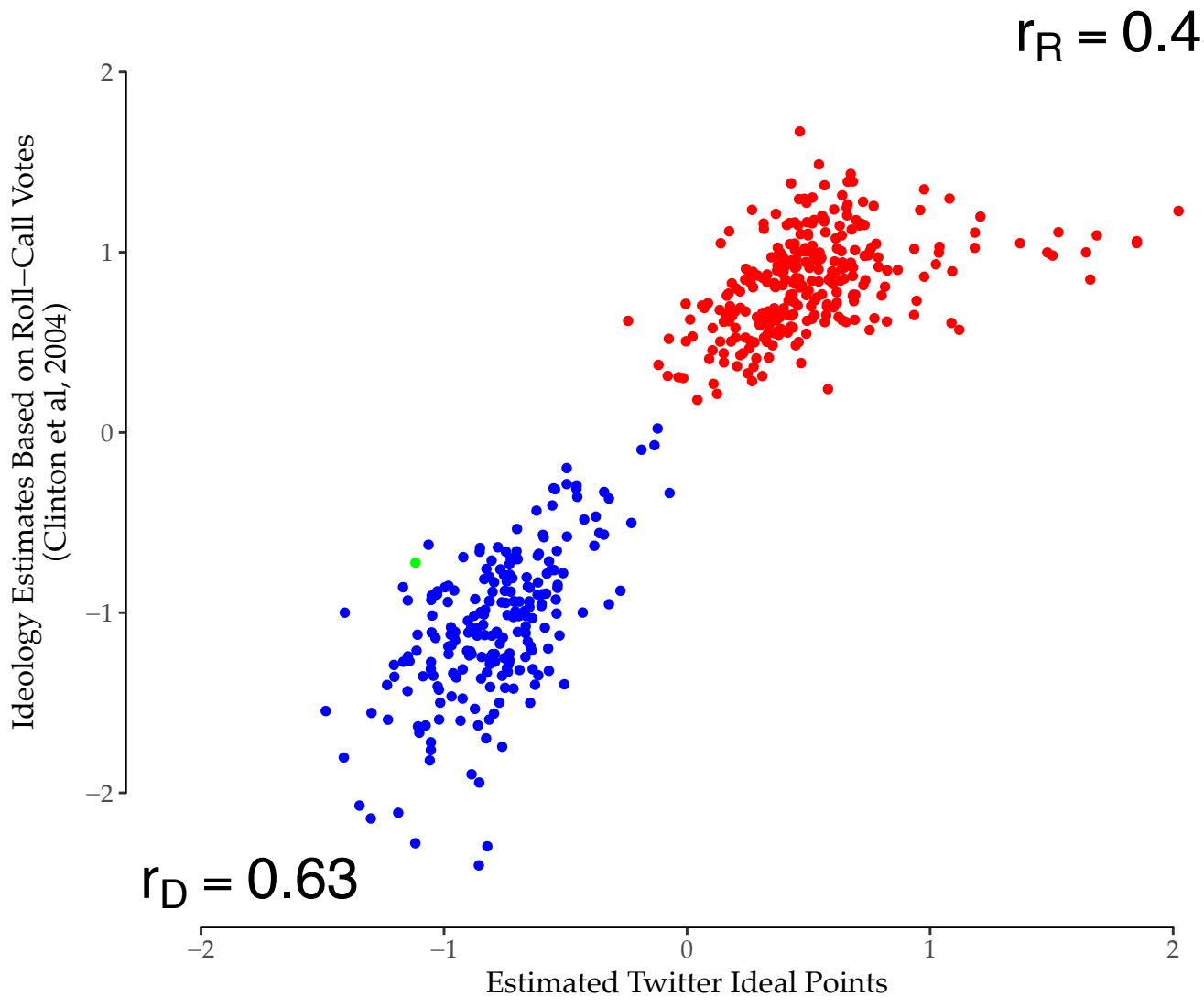


Validation

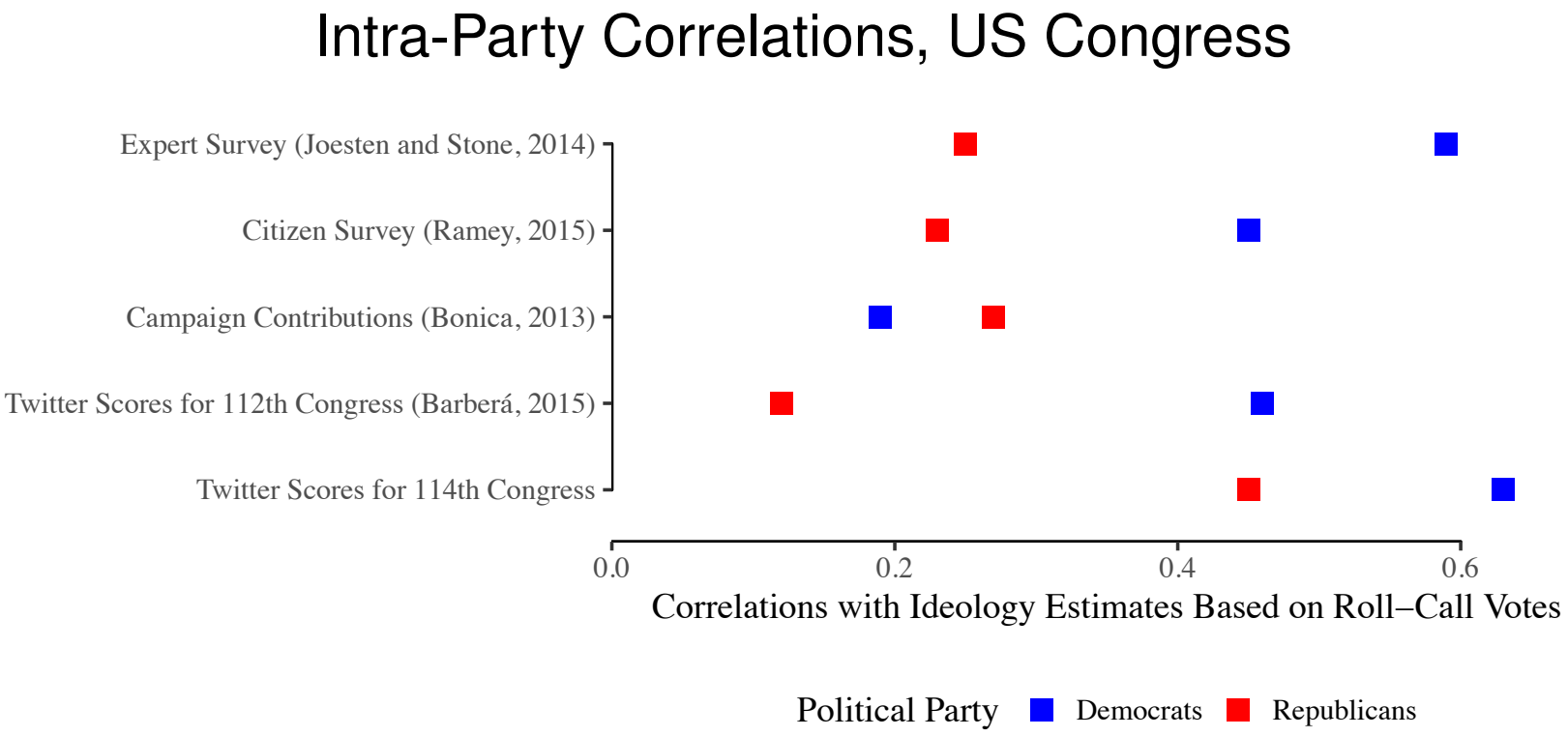
This method is able to correctly classify and scale Twitter users on the left-right dimension:

1. Political elites
   - Correlated with measures based on roll-call votes.
   - Predicts votes in Congress beyond just party ID
2. Citizens
   - Correlated with self-reported measures of ideology
   - Estimates at city/state level match survey aggregates
   - Accurately predicts party registration in voter files

Political elites



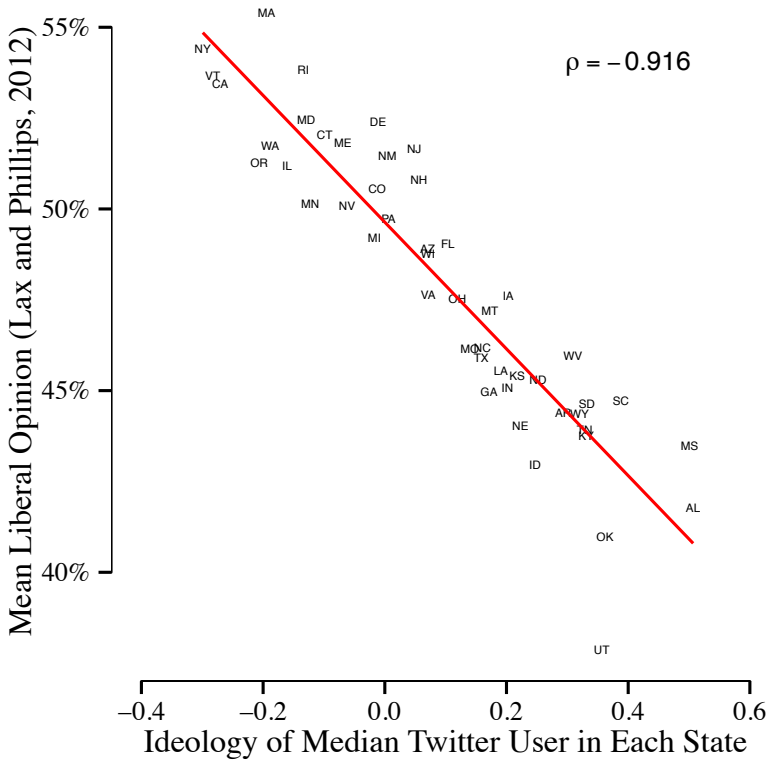
Political elites

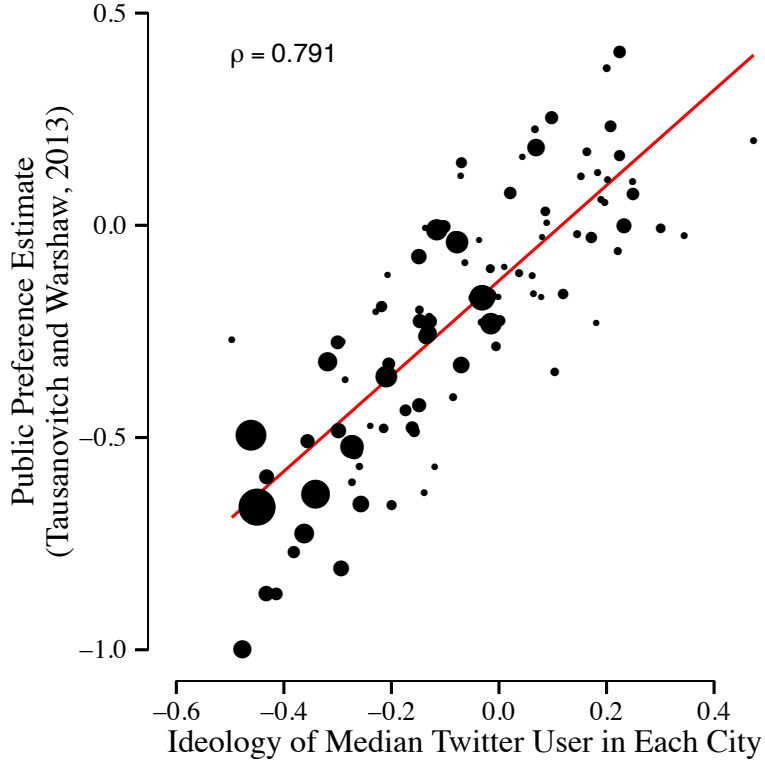


Source: Tausanovitch and Warshaw, *Political Analysis*, forthcoming

Citizens

State- and city-level average ideology is correlated with aggregated survey responses

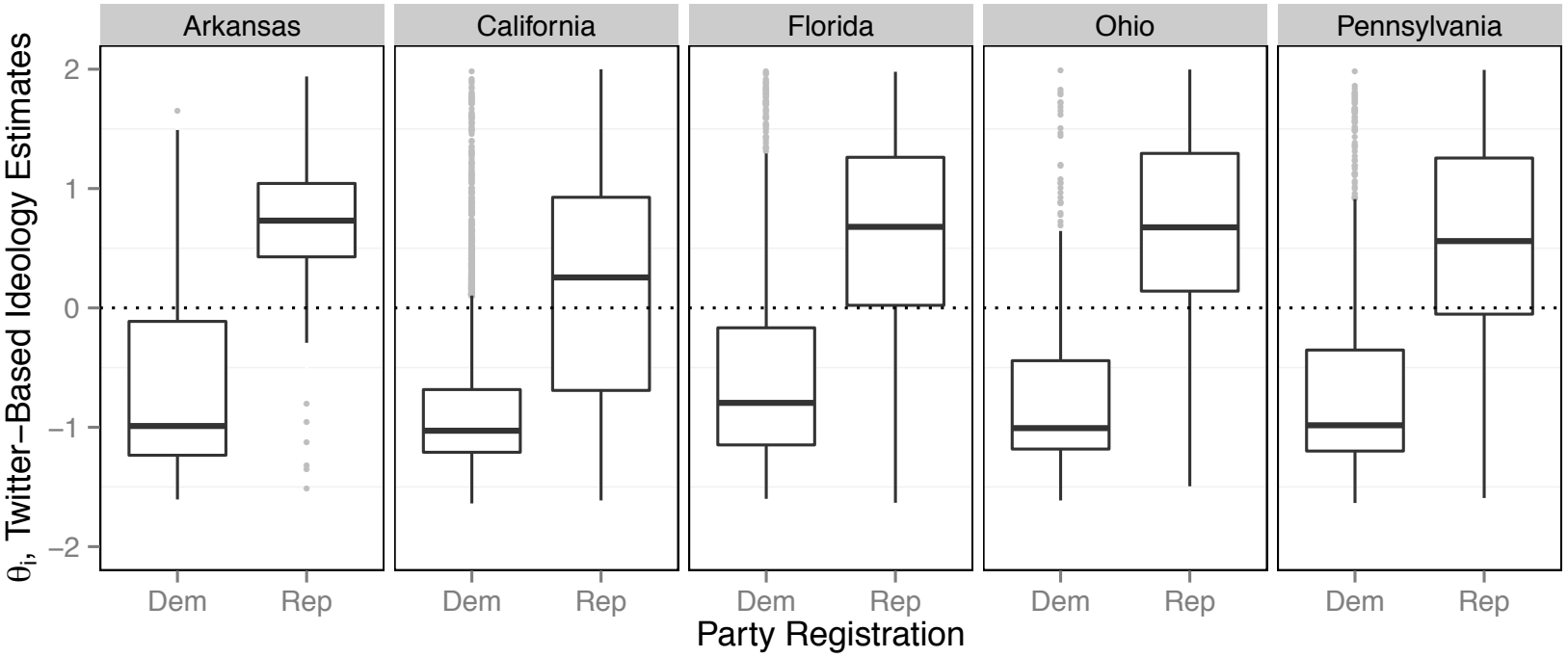




Data: Lax and Phillips (2012), Tausanovitch and Warshaw (2013)

Citizens

Estimated Twitter ideology predicts party identification in voting registration records, matched to geolocated Twitter accounts



Predictive accuracy for party affiliation is 83%

Matching Twitter Accounts with Offline Voting Records

Geographic location for Twitter users:

- 1.2 billion geolocated tweets ($\sim$8TB) from July 2013 to June 2014 $\rightarrow$ 250M in the U.S. (4.4M unique users)
- Use shape files to identify county and zipcode in U.S.

Voting registration records:

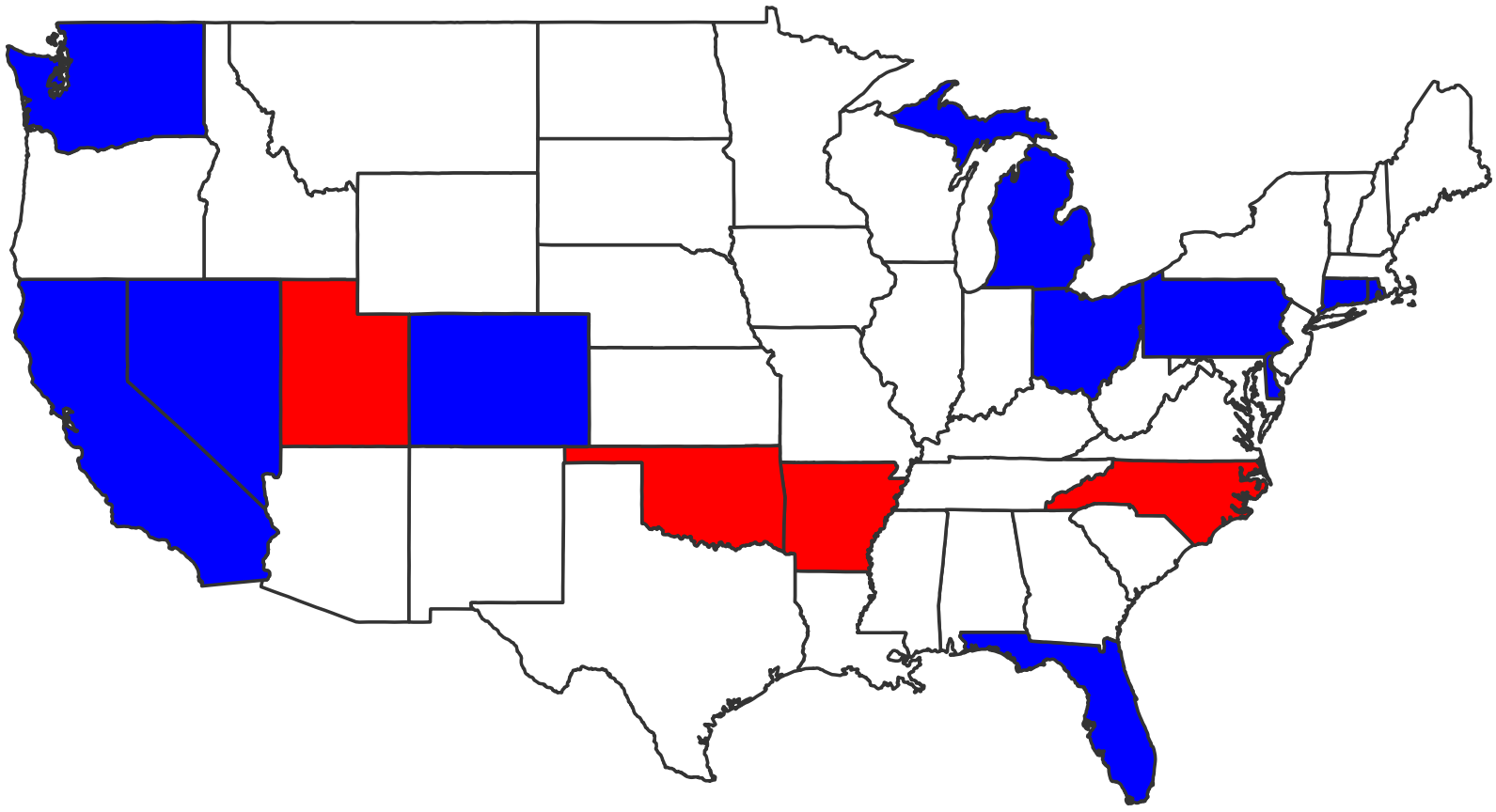
FIRST	LAST	VOTERID	COUNTY	PARTY	2012	GENDER	...
angela	myers	610901468	franklin	REP	X	F	...
ryan	petrik	610901998	franklin	DEM	X	M	...
...							

	RESIDENTIAL ADDRESS	ZIP	RACE	...
...	123 Main St, Columbus Oh	08001	W	...
...	77 Canal St, Columbus Oh	08009	W	...

Matching process:

- Perfect *and* unique matches of first/last name at county level
- If duplicated, match at zipcode level.

Matching Twitter Accounts with Offline Voting Records



Code: `github.com/pablobarbera/voter-files`

15 states, 77M registered voters (35-50% of U.S. total)

Matched Twitter accounts: 250,000 (12.3% match rate)

Citizens

Twitter ideology is correlated with self-reported ideology by YouGov panelists who gave access to their Twitter accounts (Rivero, 2016)