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

**Pablo Barberá**

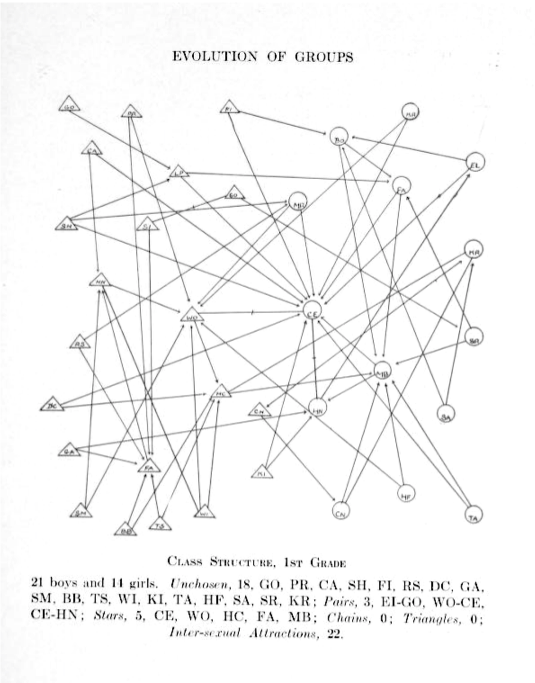
London School of Economics
pablobarbera.com

Course website:
pablobarbera.com/ECPR-SC105

# Discovery in large-scale networks

Human behaviour is characterized by connections to others

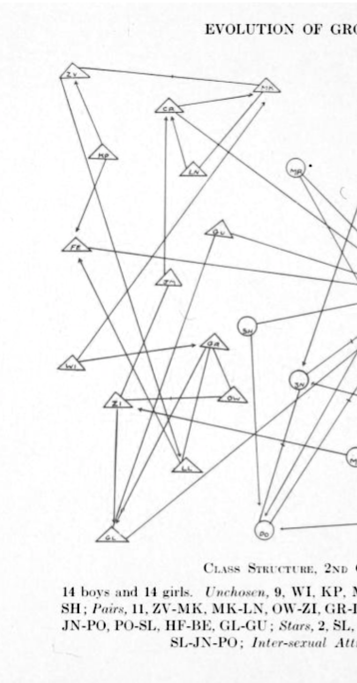
Digital technologies have led to an explosion in the availability of networked data



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.

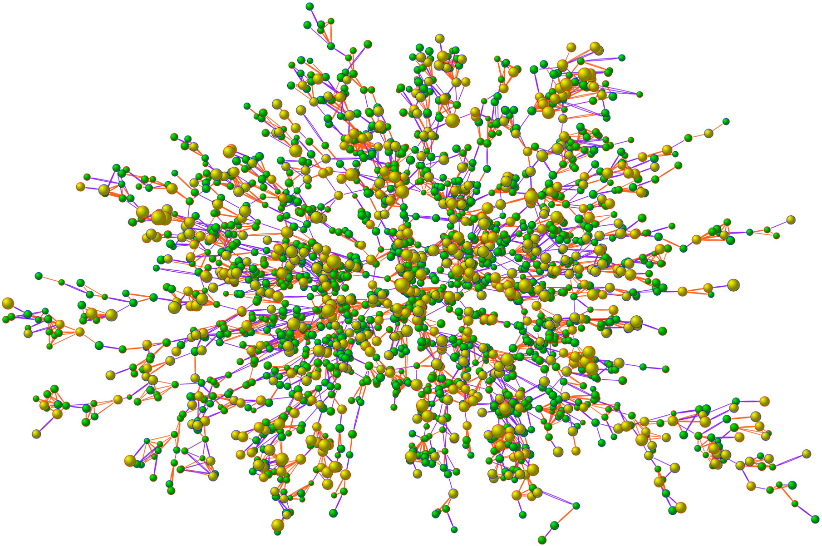


EVOLUTION OF GRO

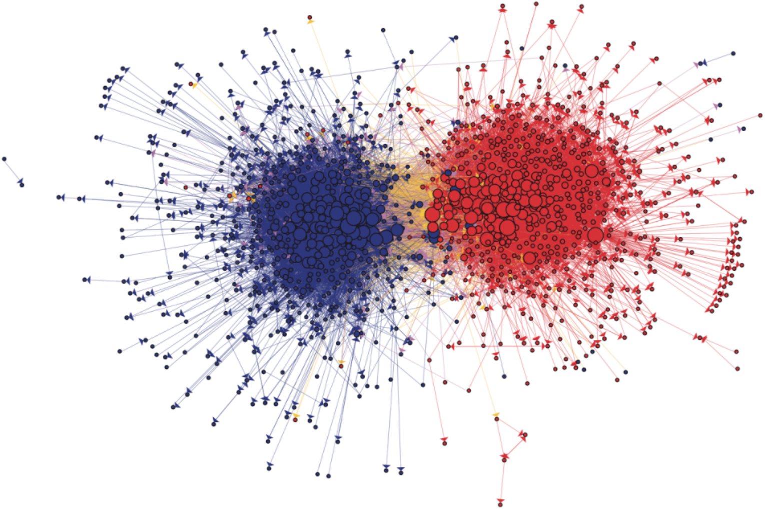
CLASS STRUCTURE, 2ND (

14 boys and 14 girls. *Unchosen*, 9, WI, KP, M
SH; *Pairs*, 11, ZV-MK, MK-LN, OW-ZI, GR-I
JN-PO, PO-SL, HF-BE, GL-GU; *Stars*, 2, SL,
SL-JN-PO; *Inter-sexual Att*

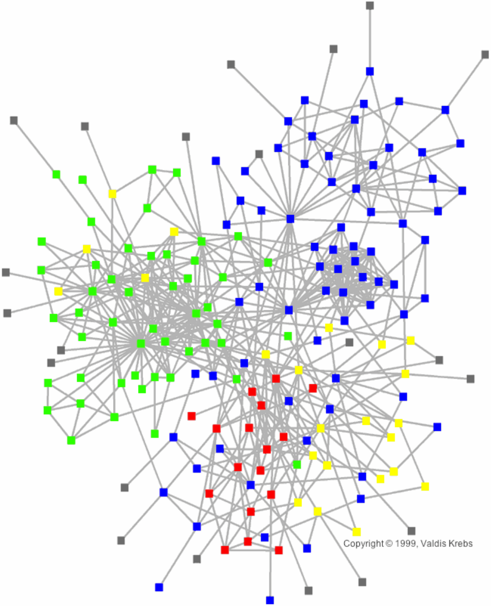
Moreno, “Who Shall Survive?” (1934)



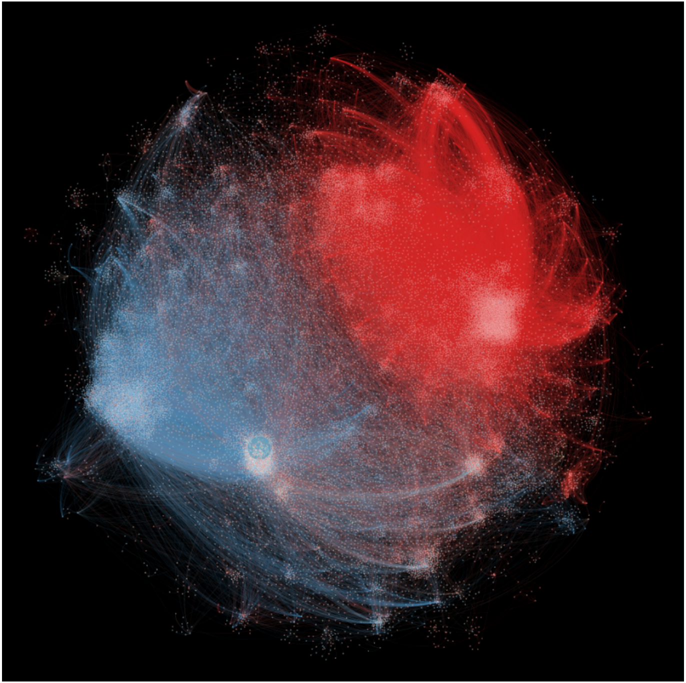
Christakis & Fowler, NEJM, 2007



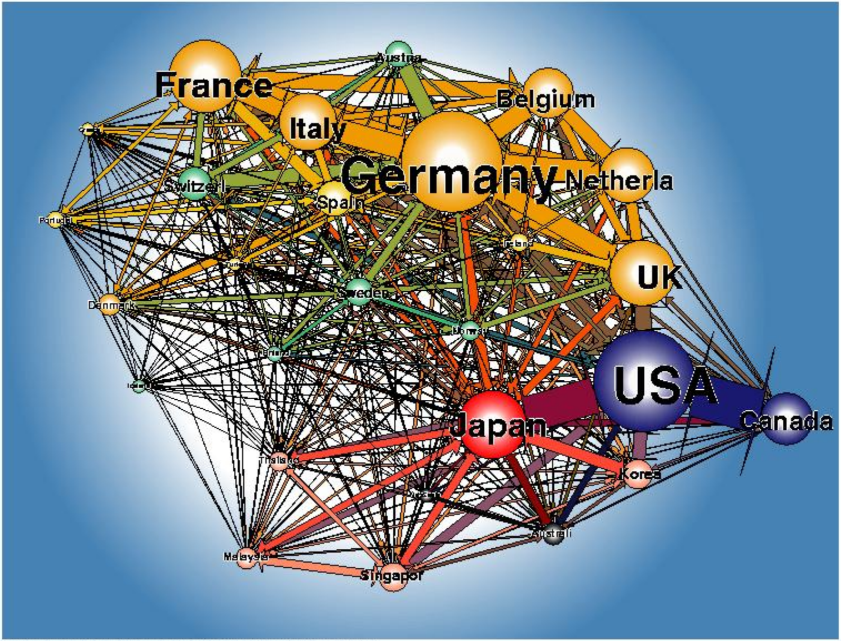
Adamic & Glance, 2004, IWLD



Email network of a company



Barbera et al, 2015, Psychological Science



(c) Lothar Krempel, Max Planck Institut f. Gesellschaftsforschung, Cologne

9/16/98 11:57:22 > OECDT92.EPS World Trade 1992 (OECD) enhanced 1981 Ausschnitt: 0.0,0.0,1.0,1.0

# Basic concepts

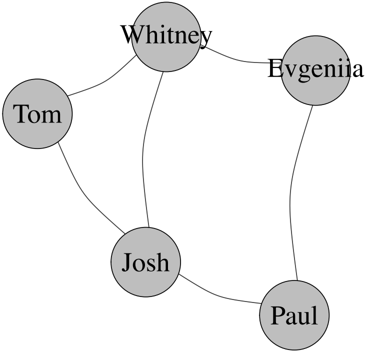
- Node (vertex): each of the units in the network
- Edge (tie): connection between nodes
  - Undirected: symmetric connection, represented by lines
  - Directed: imply direction, represented by arrows

  - Unweighted: all edges have same strength
  - Weighted: some edges have more strength than others
- A network consists of a set of nodes and edges

  i.e. a set of actors and their relationships

# Basic concepts

## Network Visualization

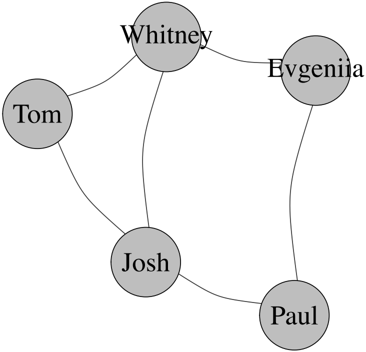


## Adjacency Matrix

|   | P | J | E | W | T |
|---|---|---|---|---|---|
| P | 0 | 1 | 1 | 0 | 0 |
| J | 1 | 0 | 0 | 1 | 1 |
| E | 1 | 0 | 0 | 1 | 0 |
| W | 0 | 1 | 1 | 0 | 1 |
| T | 0 | 1 | 0 | 1 | 0 |

# Basic concepts

## Network Visualization



## Edgelist

| | Node1 | Node2 |
|---|---|---|
| 1 | Paul | Josh |
| 2 | Paul | Evgeniia |
| 3 | Josh | Whitney |
| 4 | Josh | Tom |
| 5 | Whitney | Tom |
| 6 | Evgeniia | Whitney |

# The universality of networks

- Academic literature: papers / citations
- Internet: websites / hyperlinks
- Twitter: users / retweets
- Power grid: plants, transformers / cables
- Biology: neurons / connections
- Text: documents / cosine similarity
- Hollywood: actors / playing in same movie

# Intellectual and societal impact of networks

Why do networks matter?

- **Economic impact**: most successful companies in 21st century base their technology and business model on networks
- **Health**: importance of networks in molecular biology, spread of human diseases, pharmacology
- **Fighting terrorism**: network-centric warfare can disrupt the financial networks of terrorist organizations and map adversarial networks
- **Epidemics**: role of transportation networks in the spread of viruses
- **Neuroscience**: the human brain, one of the least-understood networks
- **Political behavior**: voting is contagious; opinion formation as a social process

# Social network analysis

Three levels of analysis:

1. **Micro**: who are the most *influential* nodes? (centrality measures)
2. **Meso**: what type of communities or clusters emerge in the network? (community detection, latent space models...)
3. **Macro**: what are the mechanisms that explain how nodes are connected? (hierarchy, homophily, diffusion..)

Three types of tools

1. **Visualization**: layout algorithms
2. **Quantification**: measures of centrality, modularity, etc.
3. **Experimentation**: at node and network level

# Node centrality

How to measure actor influence or importance in a network?

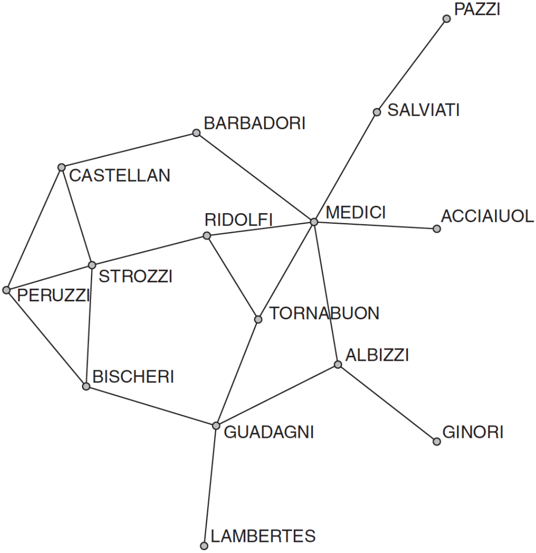
Two main conceptual definition of centrality:

1. **Degree centrality**: number of connections for each node (potential for direct reach)
   - Indegree: incoming connections
   - Outdegree: outgoing connections
2. **Betweenness centrality**: gatekeeping potential
   - How well a node connects different parts of the network
   - Fraction of shortest paths between any two nodes on which a particular node lies

→ Other measures:

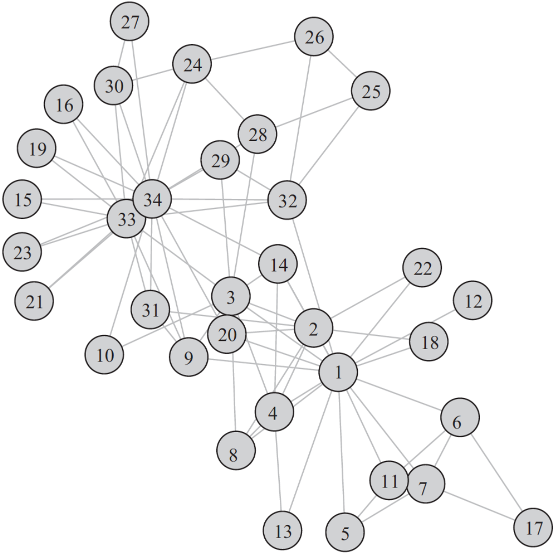
- **Closeness centrality**: broadcasting potential
- **Eigenvector centrality and coreness**: centrality measured as being connected to other central neighbors

# Florentine family marriages in the 15th century



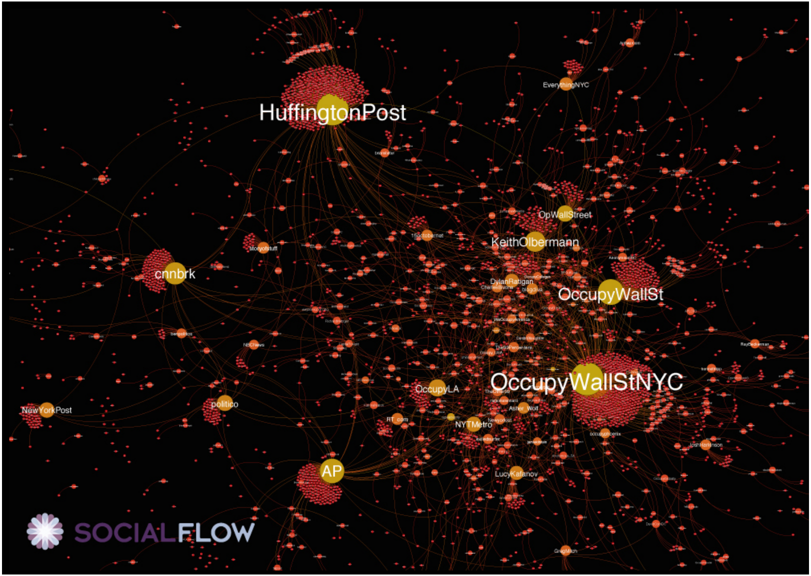
**Source**: Padgett (1993) and Sinclair (2016)

# Social activities in a Karate club



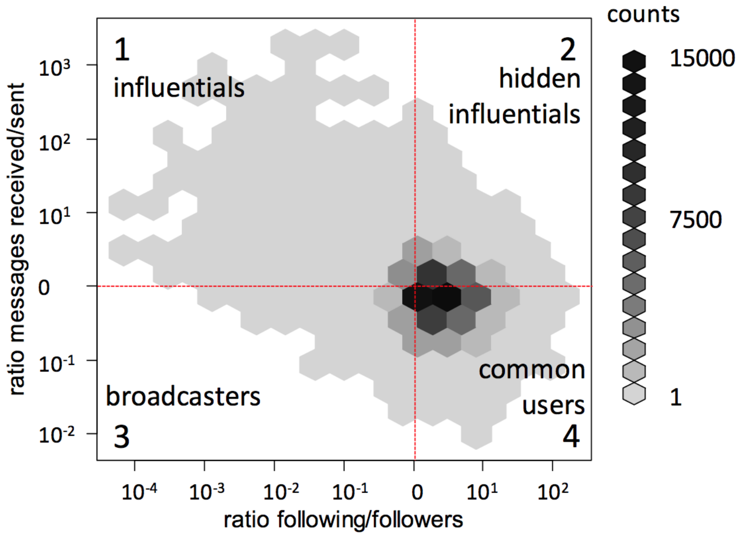
**Source**: Zachary (1997) and Sinclair (2016)

# Occupy Wall Street Twitter networks



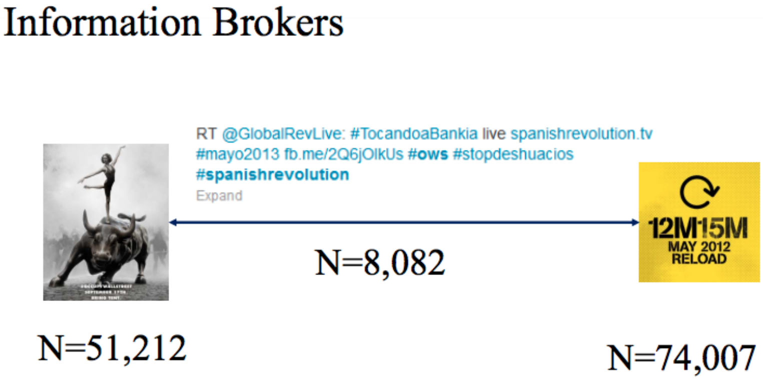
**Source**: Lotan (2011)

# Protest networks on Twitter



**Source**: González-Bailón et al (2013)

# Occupy Wall Street Twitter networks



## Information Brokers

RT @GlobalRevLive: #TocandoaBankia live spanishrevolution.tv #mayo2013 fb.me/2Q6jOlkUs **#ows** #stopdeshuacios **#spanishrevolution**

Expand

N=8,082

N=51,212

N=74,007

**Source**: González-Bailón and Wang (2016)

# Discovery in large-scale networks

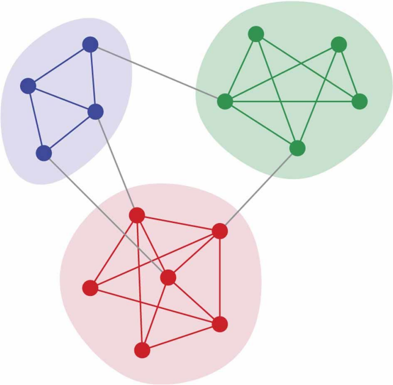
How to understand the structure of large-scale networks?

- 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

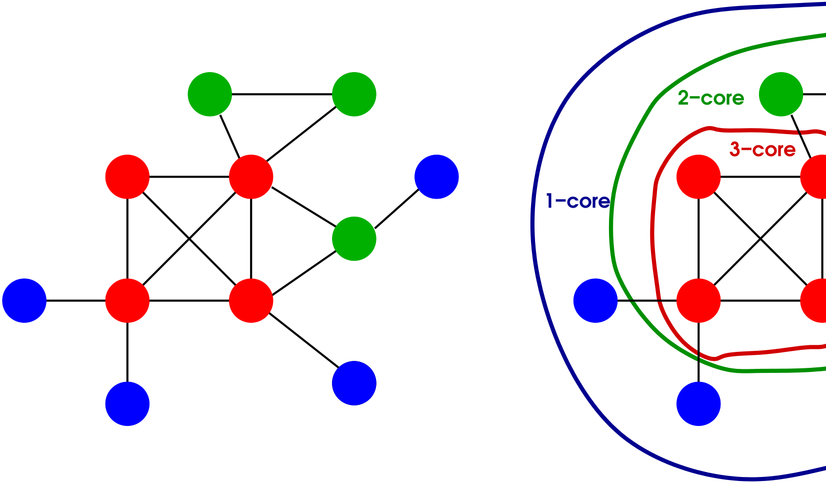


**Source**: Newman (2012)

# Network hierarchy

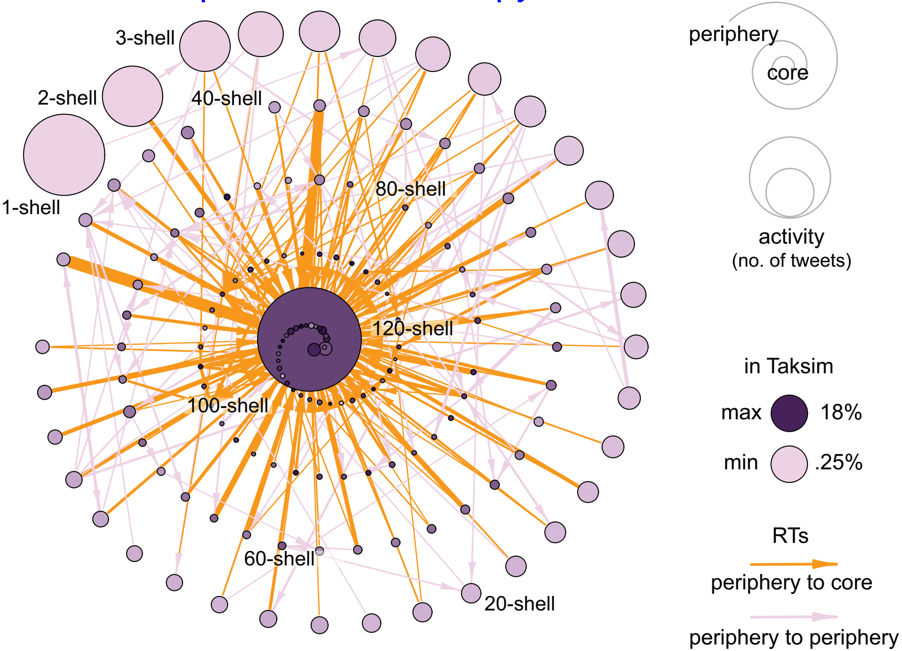
- **Intuition**
  - Large-scale networks have hierarchical properties
- **Network core:**
  1. *Centrality*: high relative importance in network
  2. *Connectivity*: many possible distinct paths between individuals
     (not captured by simple topological measures)
- **k-core decomposition**
  - Algorithm 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.

# k-core decomposition



Source: Alvarez-Hamelin et al, 2005

# k-core decomposition of #OccupyGezi network



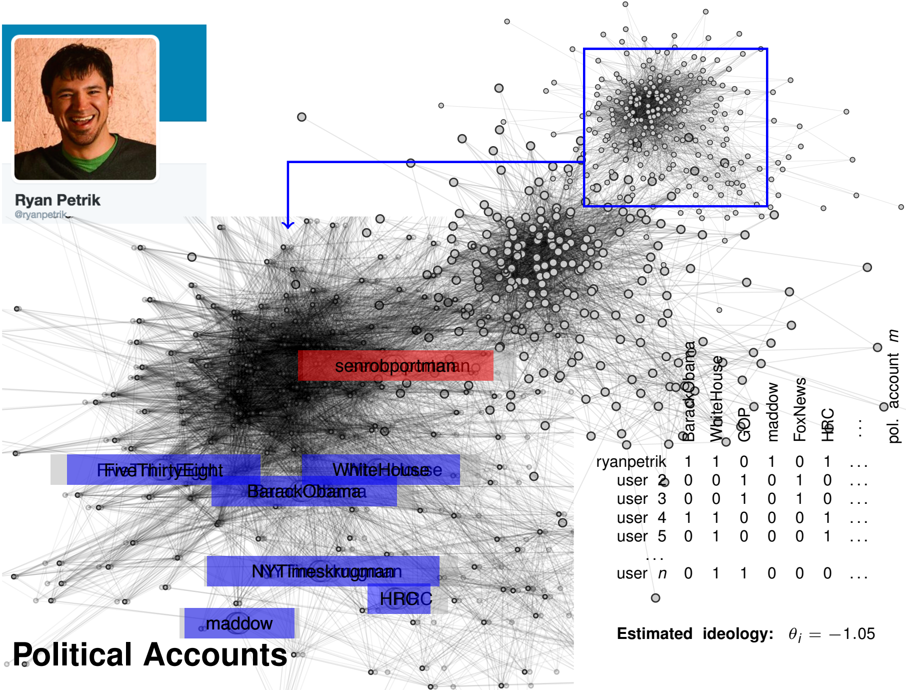
# Latent space models

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



# Political Accounts

$$
\begin{array}{rccccccc}
 & \text{BarackObama} & \text{WhiteHouse} & \text{GOP} & \text{maddow} & \text{FoxNews} & \text{HRC} & \ldots & \text{pol. account } m \\
\text{ryanpetrik} & 1 & 1 & 0 & 1 & 0 & 1 & \ldots \\
\text{user } 2 & 0 & 0 & 1 & 0 & 1 & 0 & \ldots \\
\text{user } 3 & 0 & 0 & 1 & 0 & 1 & 0 & \ldots \\
\text{user } 4 & 1 & 1 & 0 & 0 & 0 & 1 & \ldots \\
\text{user } 5 & 0 & 1 & 0 & 0 & 0 & 1 & \ldots \\
\ldots & & & & & & & \\
\text{user } n & 0 & 1 & 1 & 0 & 0 & 0 & \ldots
\end{array}
$$

**Estimated ideology:** $\theta_i = -1.05$

# Spatial following model

- Users' and political accounts' ideology ($\theta_i$ and $\phi_j$) 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

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

- with latent variables:

  $\theta_i$ measures *ideology* of user *i*

  $\phi_j$ measures *ideology* of political account *j*

- and:

  $\alpha_j$ measures *popularity* of political account *j*

  $\beta_i$ measures *political interest* of user *i*

  $\gamma$ is a normalizing constant

# Estimation

- Likelihood function:

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

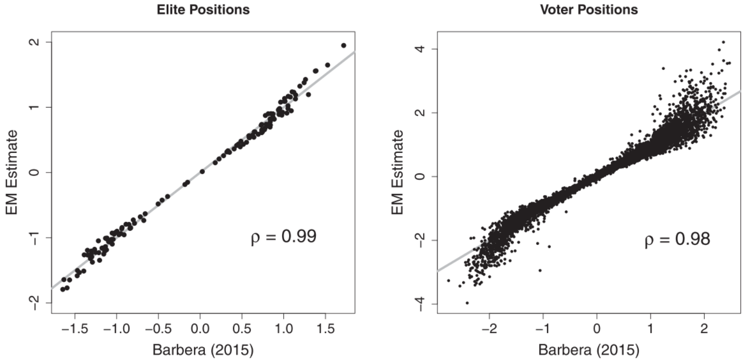
$$\text{where } \pi_{ij} = \alpha_j + \beta_i - \gamma(\theta_i - \phi_j)^2$$

- Intractable with maximum likelihood methods → 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 $\theta$: $\theta_i \sim N(0, 1)$
  - 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, **S**:

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

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

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

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

2. Calculate SVD of **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:

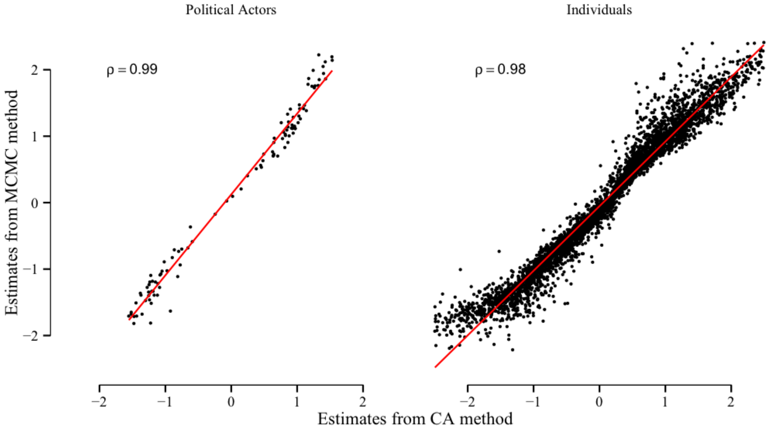
   $$\theta = \mathbf{D}_r^{1/2}\mathbf{U} \text{ for rows (ordinary users)}$$

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

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

# Estimation

## Correspondence analysis (Greenacre, 1984; 2010)

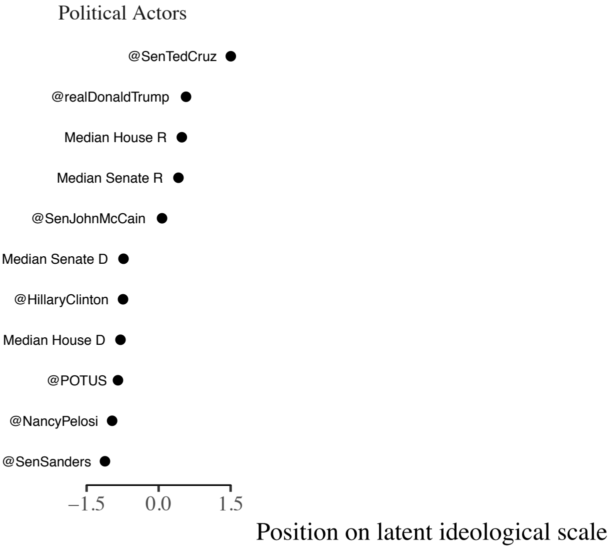


Runtime for N=10,000 users and J=173 political accounts:
MCMC = 6.5 days; EM = 35 minutes; CA = 1.88 seconds

# Model validation

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

# Face validity: political accounts



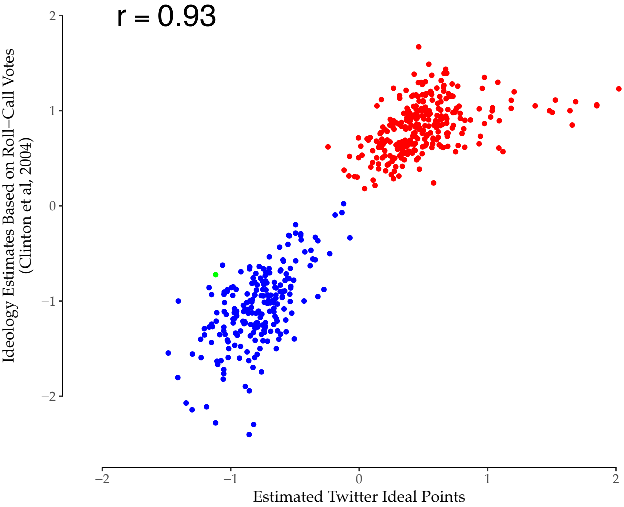
# Validation

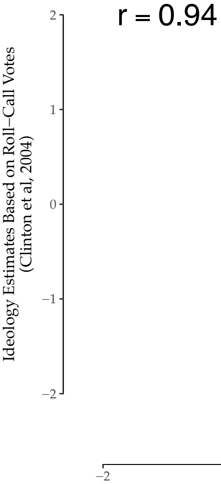
This method is able to correctly classify and scale Twitter users on the left-right dimension:

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

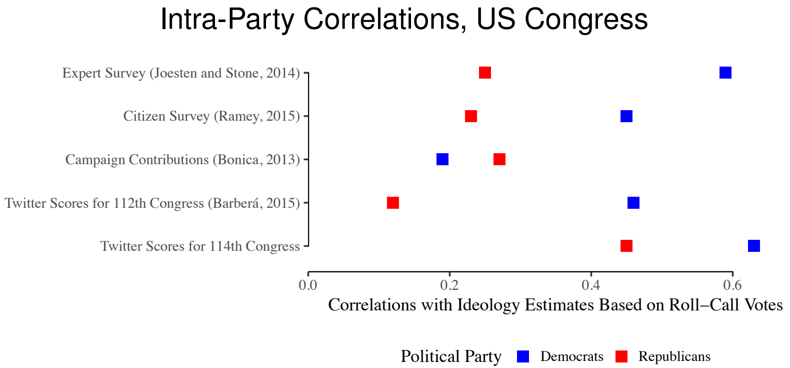
# Political elites

## Ideal Points of Members of the 114th U.S. Congress





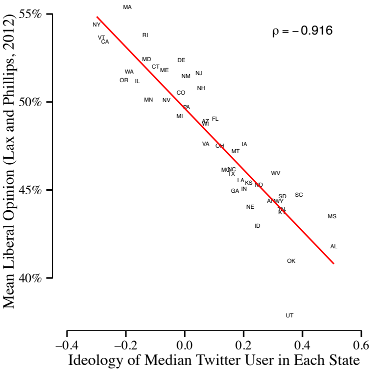
# Political elites

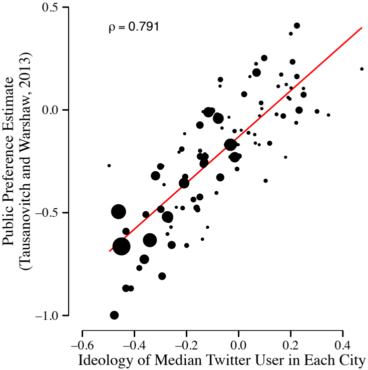


**Source:** Tausanovitch and Warshaw, *Political Analysis*, forthcoming

# Citizens

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

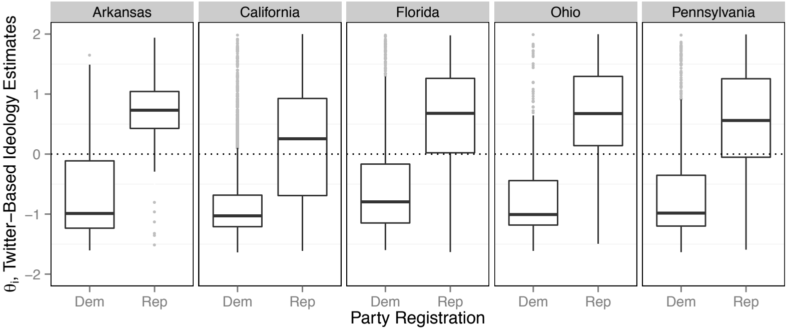




**Data:** Lax and Phillips (2012), Tausanovitch and Warshaw (2013)

# 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 (∼8TB) from July 2013 to June 2014 → 250M in the U.S. (4.4M unique users)
- Use shape files to identify county and zipcode in U.S.

## Voting registration records:

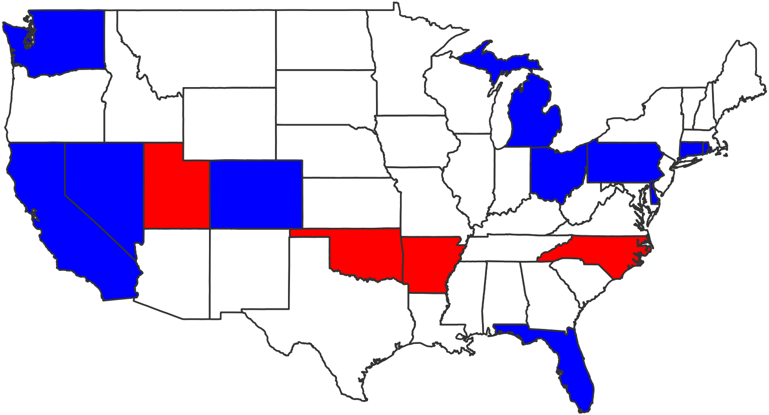
| FIRST | LAST | VOTERID | COUNTY | PARTY | 2012 | GENDER | ... |
|---|---|---|---|---|---|---|---|
| angela | myers | 610901468 | franklin | REP | X | F | ... |
| ryan | petrik | 610901998 | franklin | DEM | X | M | ... |
| ... | | | | | | | |

| | 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



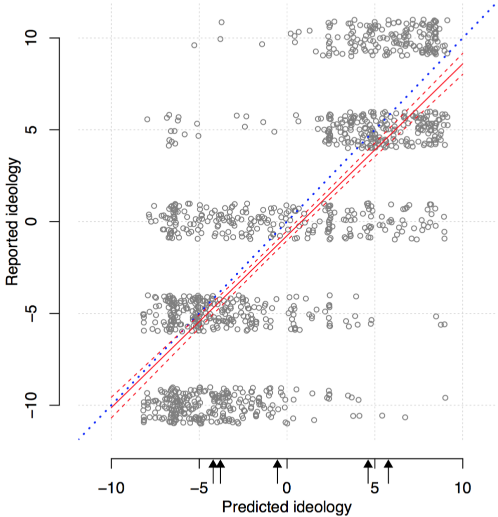
Code: `github.com/pablobarbera/voter-files`

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)


Reported ideology
Predicted ideology