POIR 613: Measurement Models and Statistical Computing

Pablo Barberá

School of International Relations
University of Southern California
pablobarbera.com

Course website:
pablobarbera.com/POIR613/

Today

1. Solutions for last week's challenge
2. Deadline: YESTERDAY for descriptive statistics
3. Next: first full draft on November 17
4. Other announcements:
   - Guest lecture November 14: **Franziska Keller** (Hong Kong University of Science and Technology, UCSD), social network analysis of Chinese elites
   - Talk, November 29: **Dean Knox** (MSR/Princeton) & **Chris Lucas** (Harvard), audio as data
   - No class on November 21st
   - Office hours 4-5.30pm only tomorrow
5. Today:
   - Latent variable models
   - Collecting social media data

Collecting Facebook data

Facebook only allows access to public pages' data through the Graph API:

1. Posts on public pages and groups
2. Likes, reactions, comments, replies...

Some public user data (gender, location) was available through previous versions of the API (not anymore)

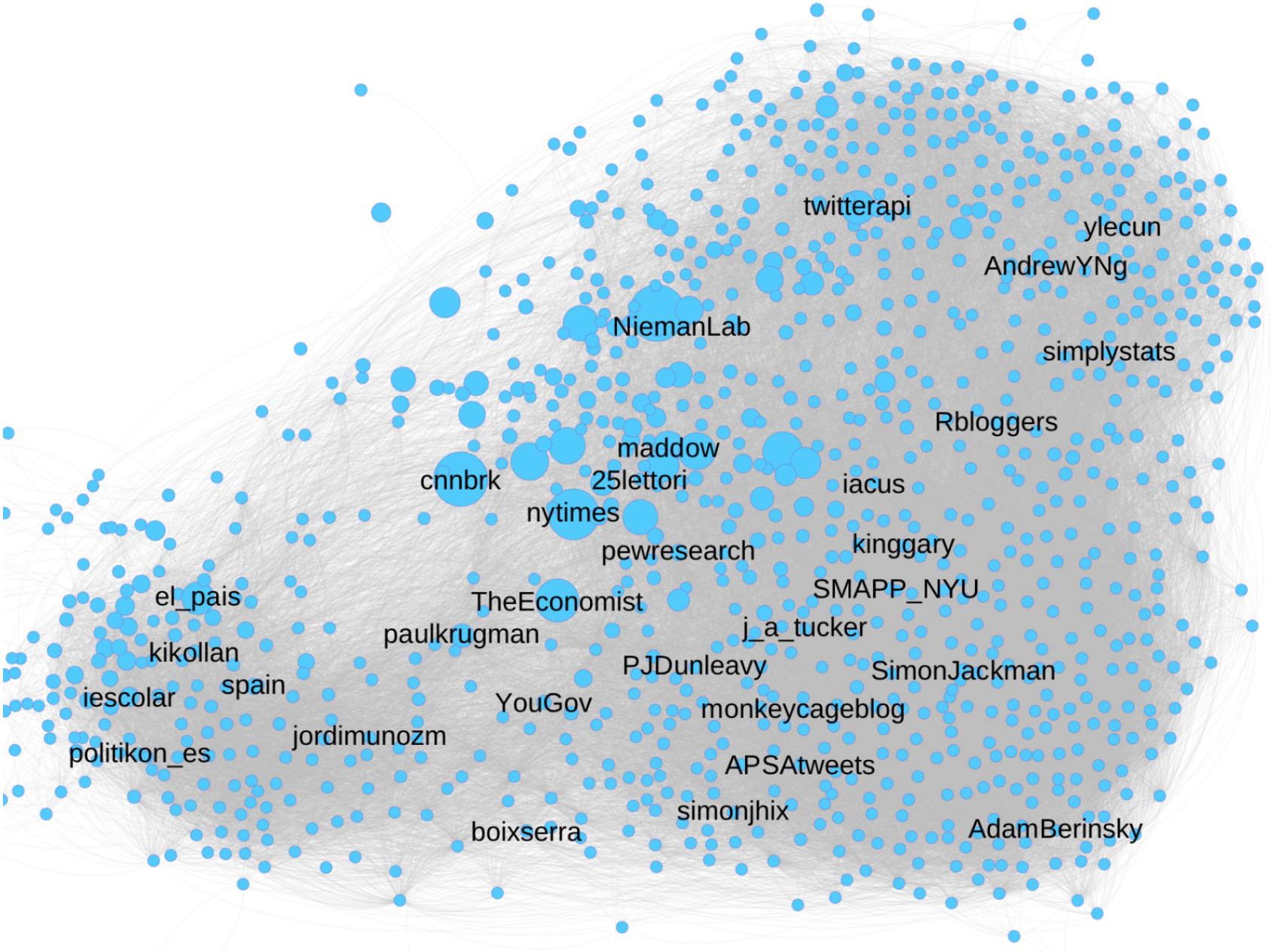
Aggregate-level statistics available through the FB Marketing API. See the code by Connor Gilroy (UW)

Access to other (anonymized) data used in published studies requires permission from Facebook or from users

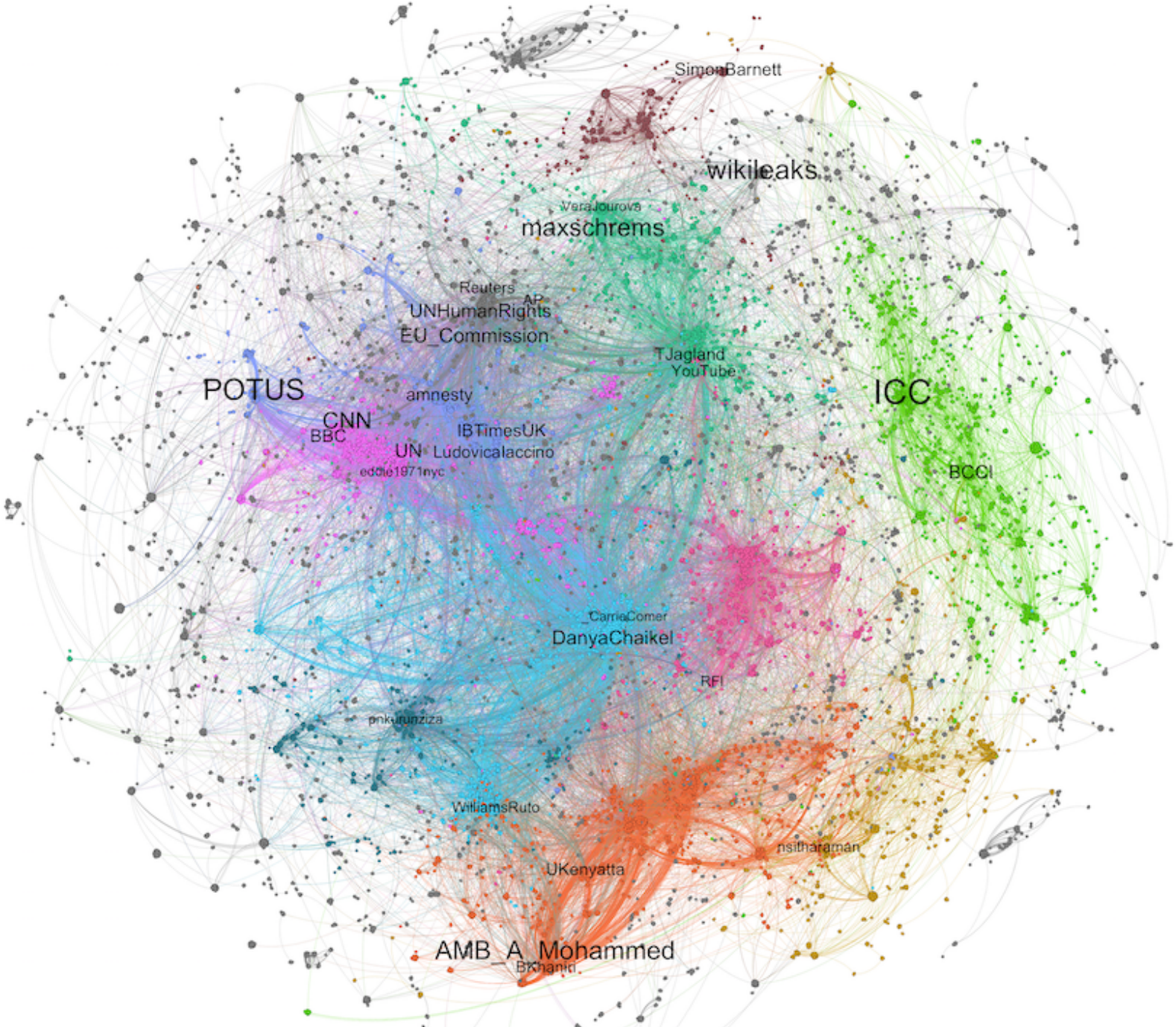
R library: Rfacebook

Discovery in large-scale networks

Latent structure of social networks



The dreaded *hairball*



Discovery in large-scale networks

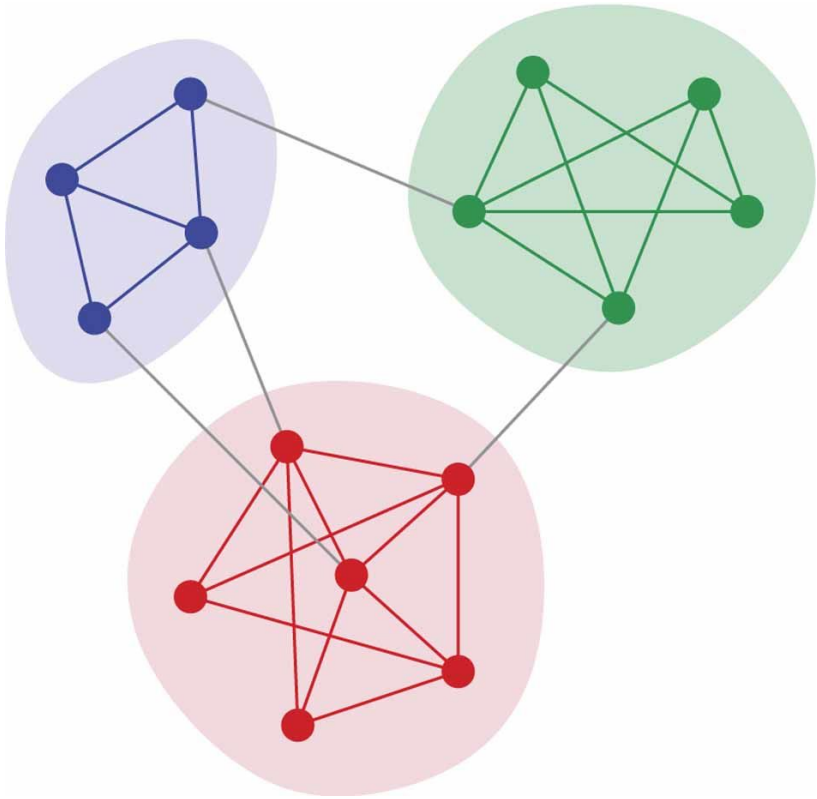
How to understand the structure of large-scale networks?

- 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

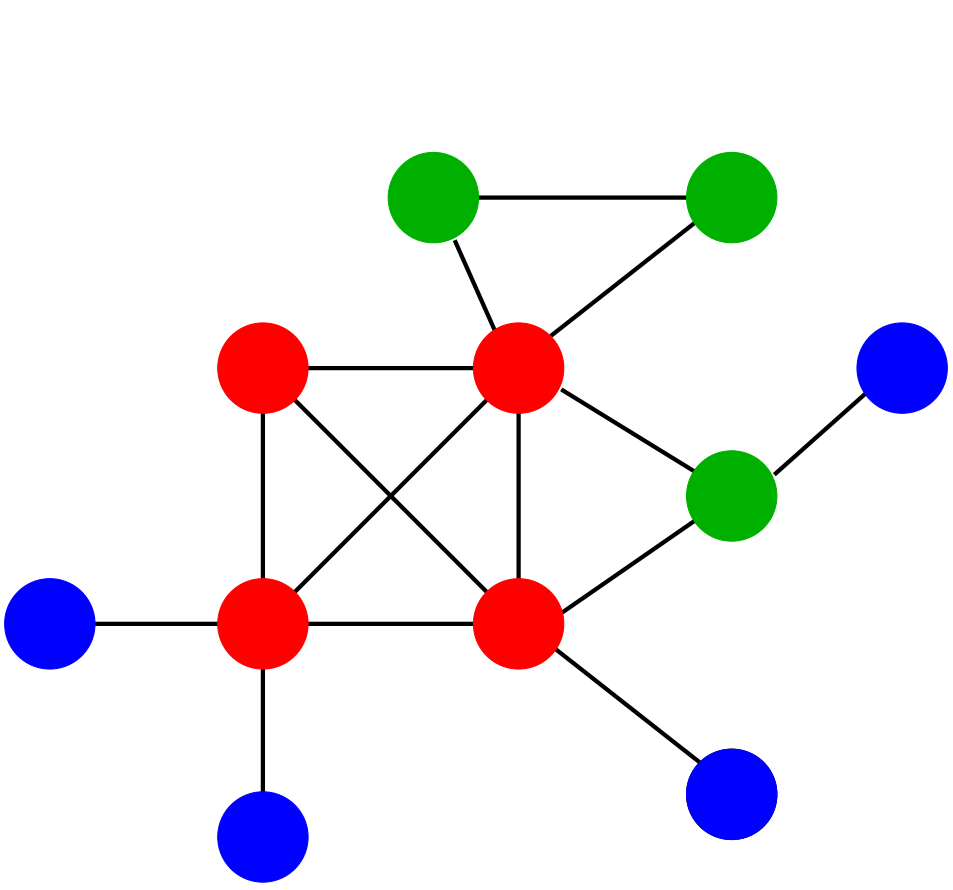


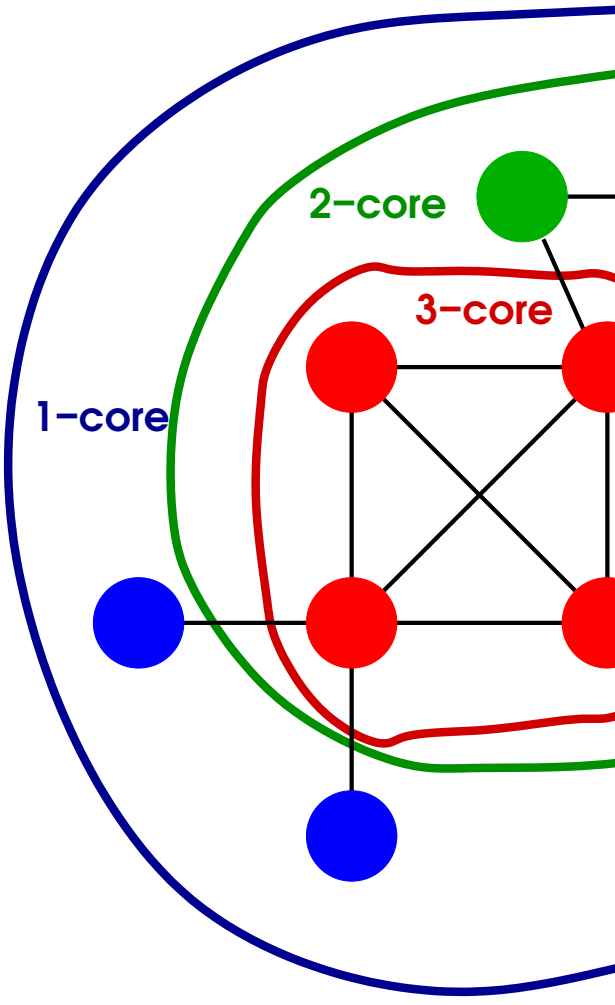
Source: Newman (2012)

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: $\mathcal{O}(n + e)$

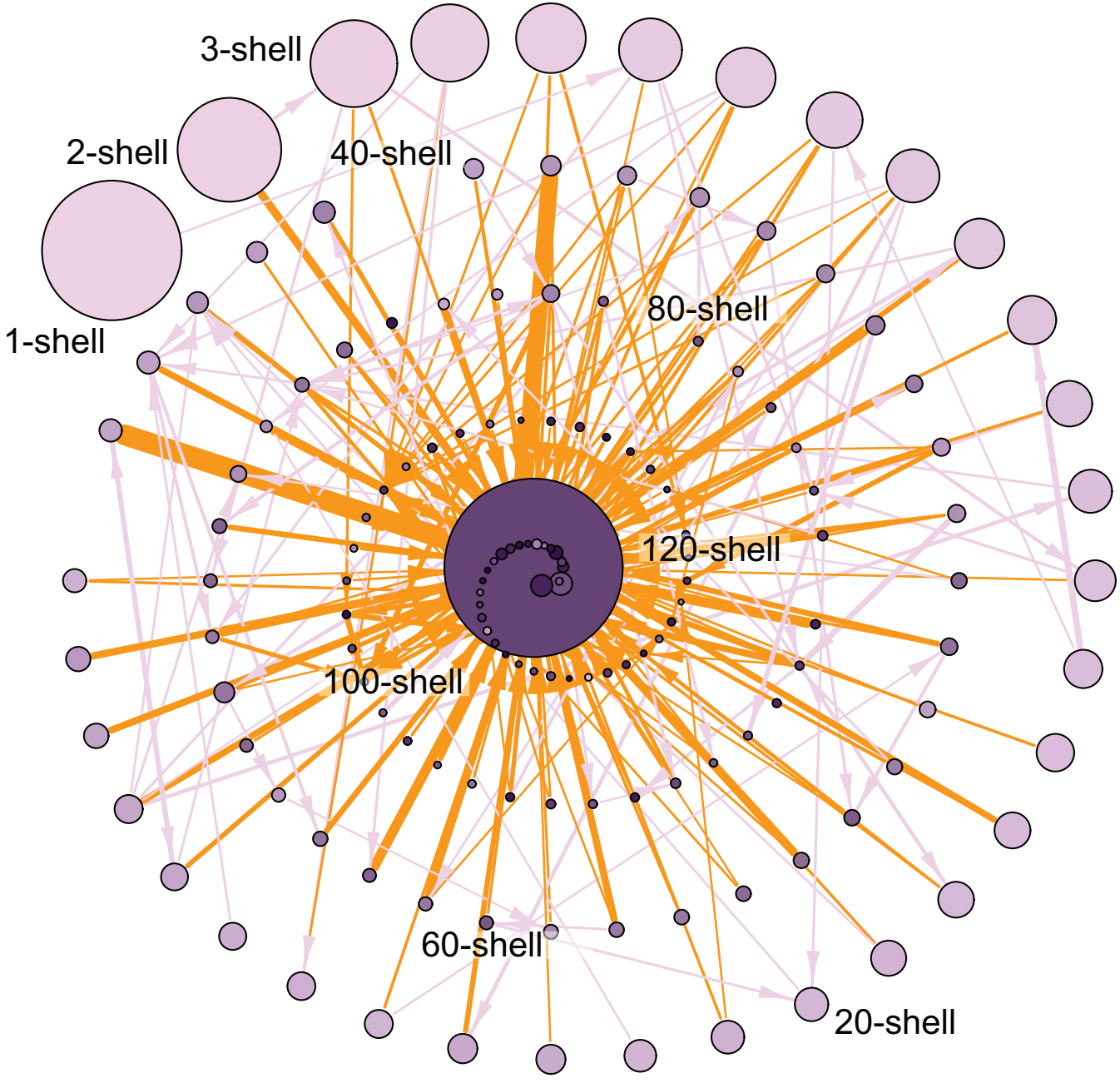
k-core decomposition





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

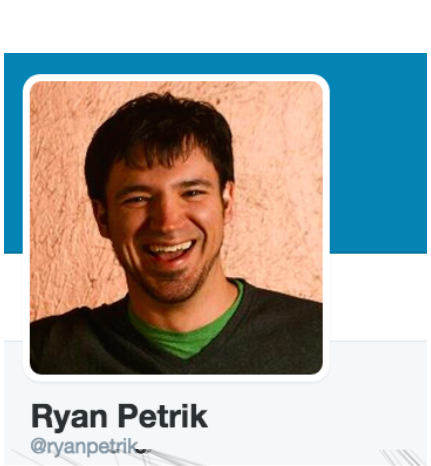
Latent space models

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



pol. account m	BarackObama	WhiteHouse	GOP	maddow	FoxNews	HRC	...
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