Community Structure in "Multislice" Networks

(Community Structure in Time-Dependent, Multiscale, and Multiplex Networks)

Mason A. Porter

Mathematical Institute, University of Oxford

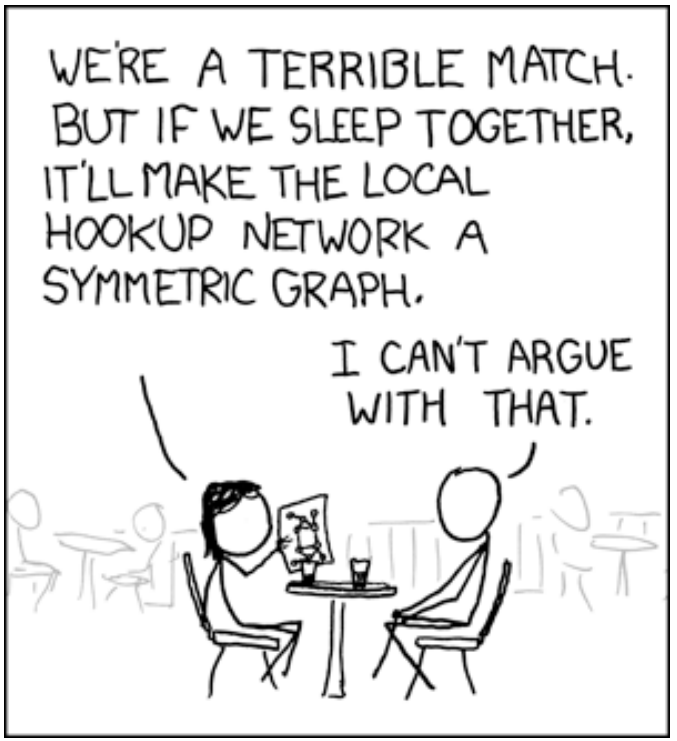
P. J. Mucha, T. Richardson, Kevin Macon, M. A. Porter, & J.-P. Onnela, *arXiv:0911.1824*
(and an additional paper in 2010 Nonlinear Science Gallery, to appear in *Chaos*)

Goal

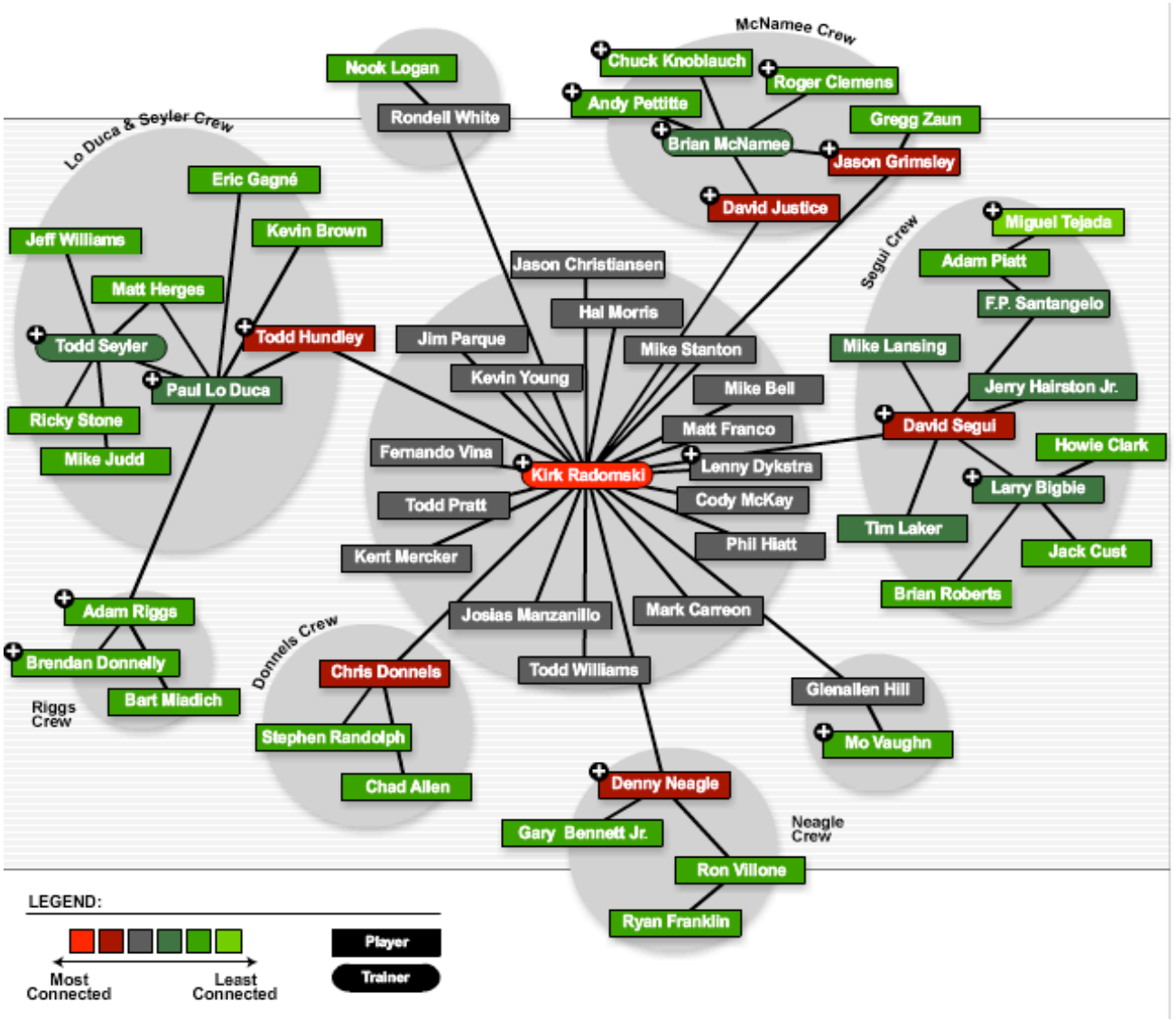
- Extend quality-optimization methods of community detection to networks with the following features:
  - *Multiscale:* Consider multiple resolution parameters at once (without sweeping)
  - *Time-Dependent:* Nodes and edges can change in time
  - *Multiplex:* Multiple types of edges

Outline

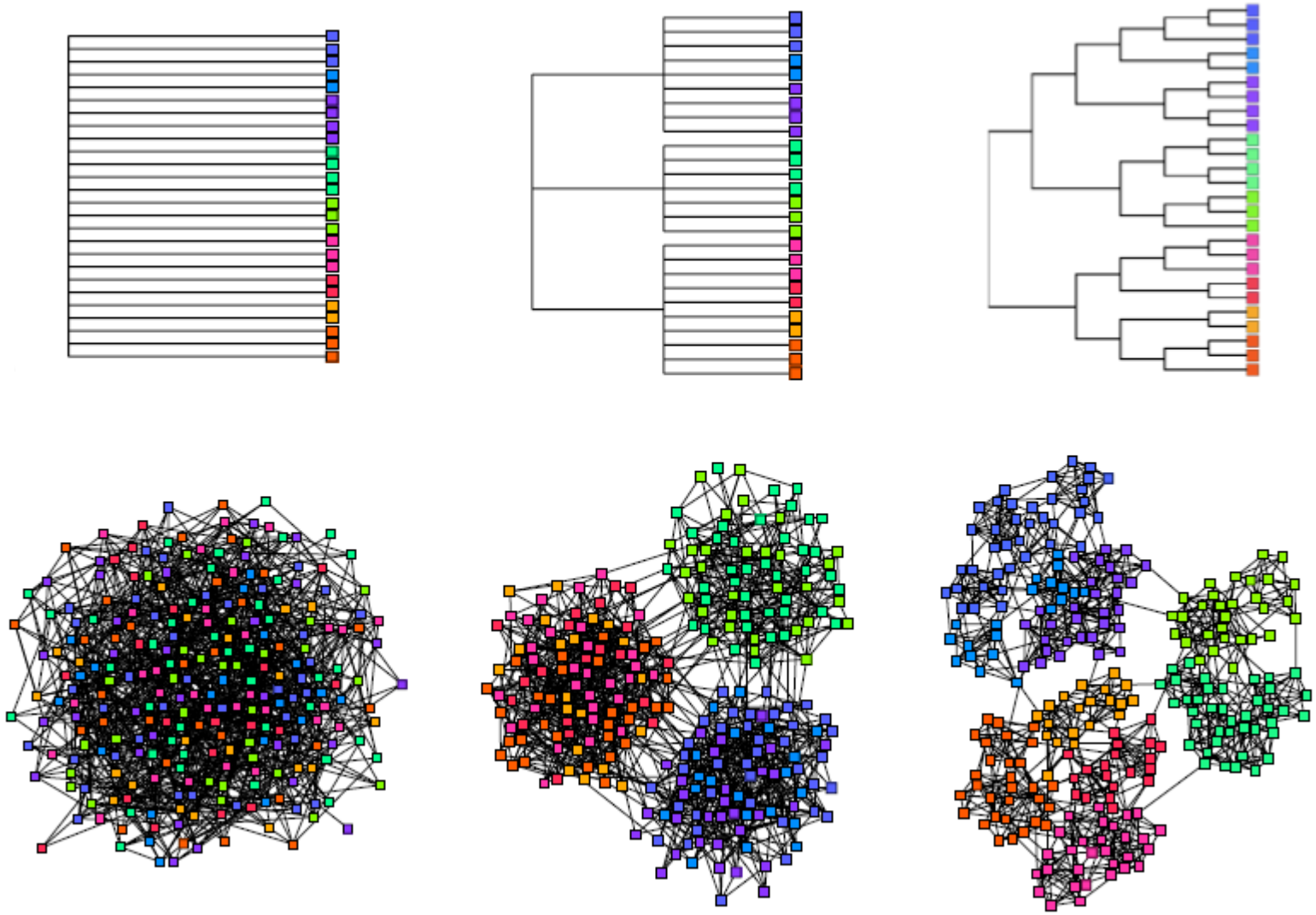
- Background: Community structure and community detection
- Multislice networks
- Examples
- Conclusions



Community Structure by hand?: Baseball Steroids Networks

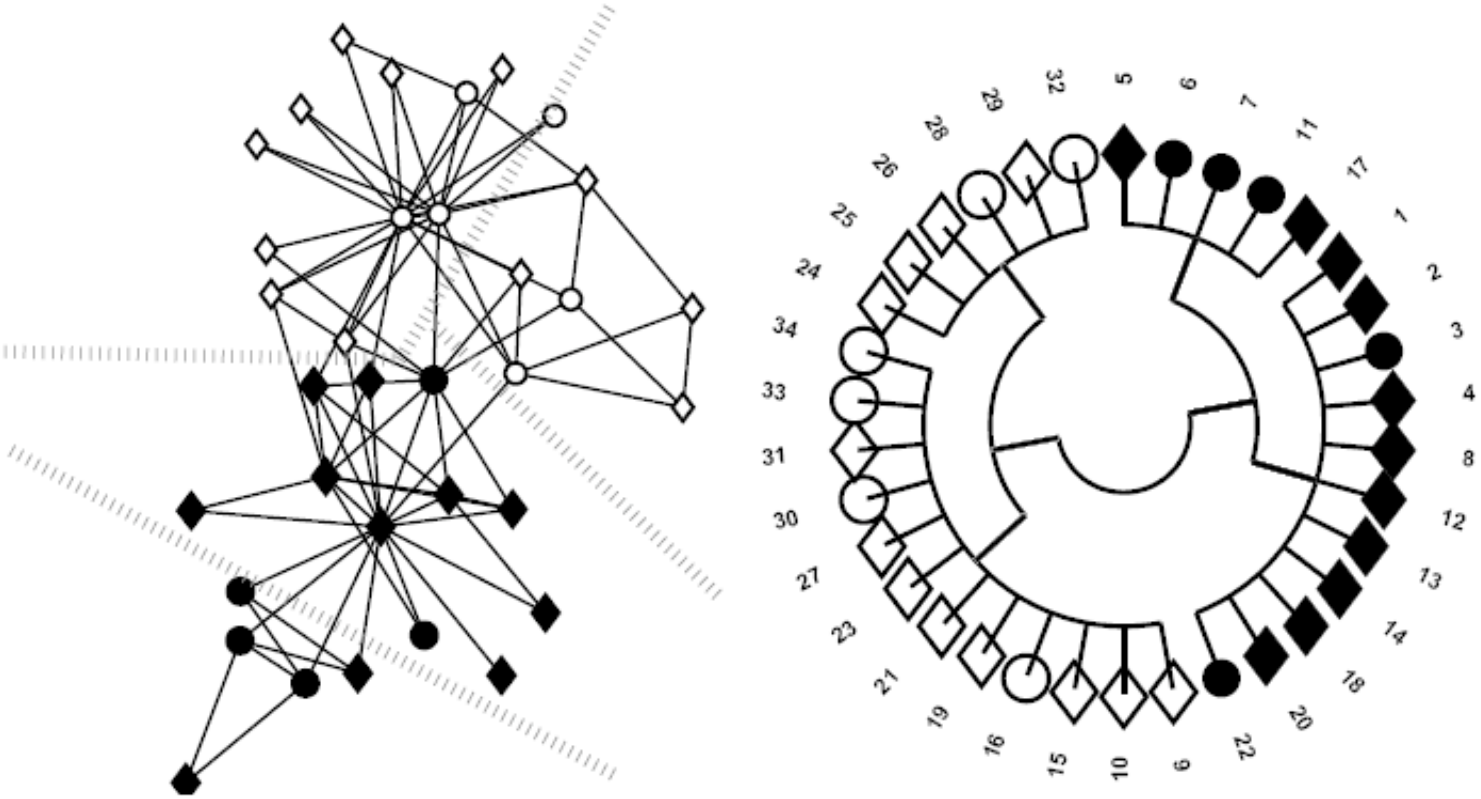


Identifying Communities Algorithmically



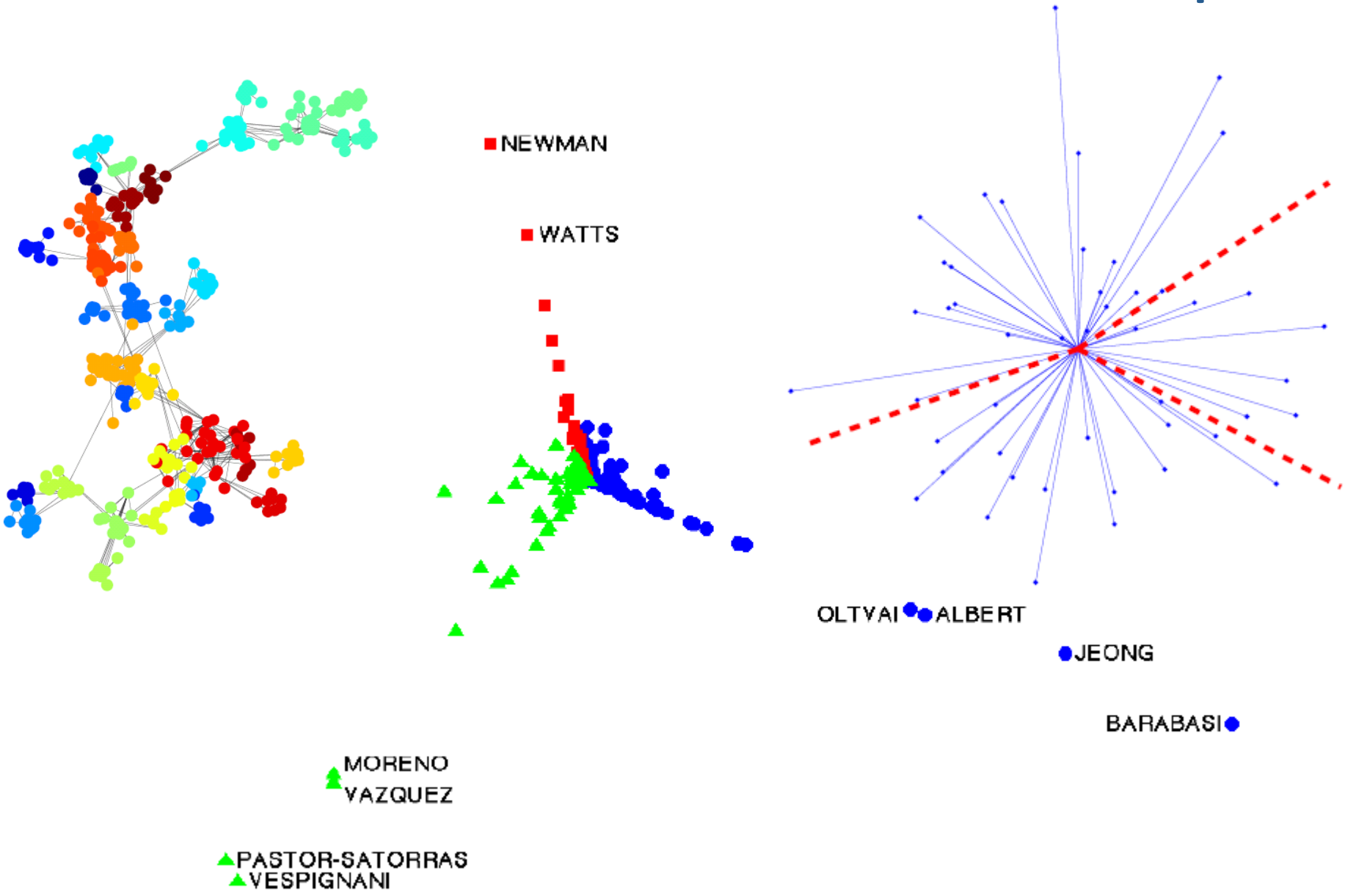
Images from A. Clauset, C. Moore, & M. E. J. Newman (*Nature*, 2008)

“This wouldn’t be a community detection talk without the Zachary Karate Club.”

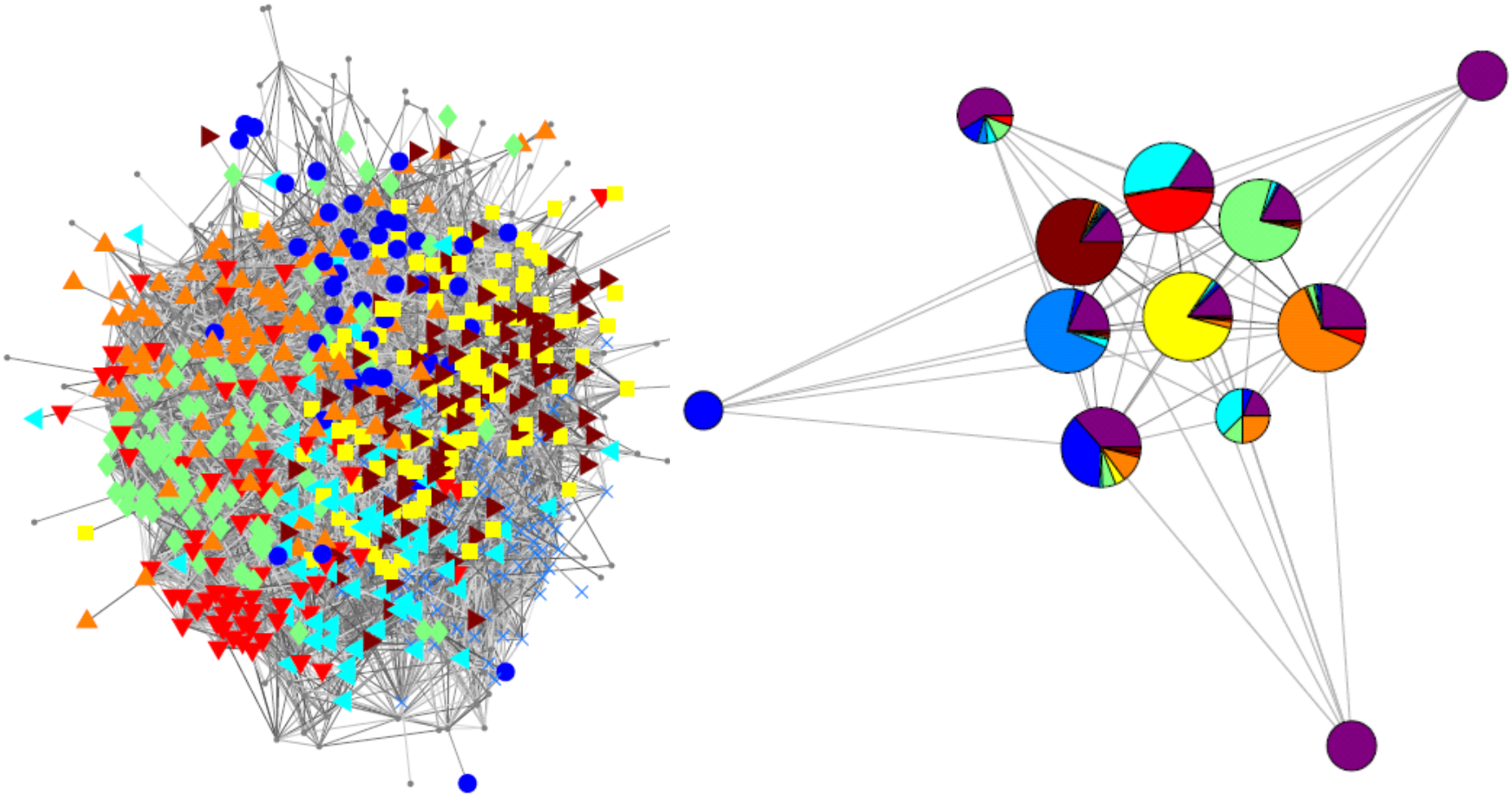


This partition optimizes modularity, which measures the number of intra-community ties (relative to randomness)

“Network Science” Coauthorship



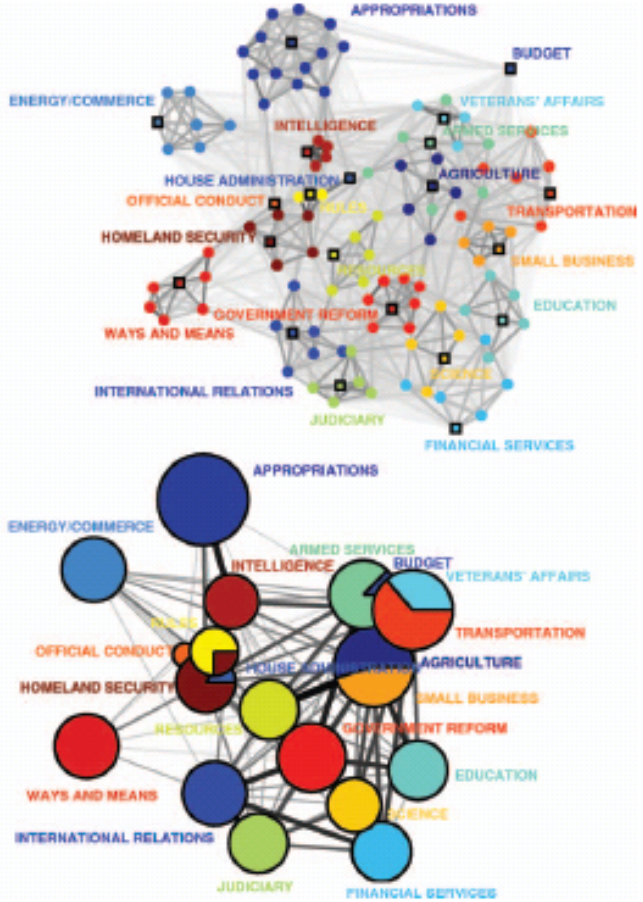
Facebook Friendship Networks



A. L. Traud, E. D. Kelsic, P. J. Mucha, & M. A. Porter, *arXiv:0809.0960*

Congressional Committees

(MAP, PJM, M. E. J. Newman, C. M. Warmbrand, & A. J. Friend)



Preliminaries

- "Hard/rigid" versus "soft/fuzzy/overlapping" clustering
- A *community* should describe a "cohesive group" of nodes
  - Tons of algorithms available
- Unifying notion: more intra-community edges than one would expect at random
  - *But what does "at random" mean?*
- Review articles
  - "Communities in Networks," M. A. Porter, J.-P. Onnela & P. J. Mucha, *Notices of the American Mathematical Society* **56**, 1082-97 & 1164-6 (2009).
  - "Community Detection in Graphs," S. Fortunato, *Physics Reports* **486**, 75-174 (2010).

Quality / Modularity

- Popular approach: Use a "modularity" quality function

$$Q = \frac{1}{2W} \sum_{i,j} B_{ij} \delta(C_i, C_j) \,, \quad B_{ij} = A_{ij} - P_{ij}$$

where $\delta(C_i, C_j)$ indicates that the B_{ij} components are only summed over cases in which nodes i and j are classified in the same community. The factor $W = \frac{1}{2} \sum_{ij} A_{ij}$ is the total edge strength in the network (equal to the total number of edges for unweighted networks), where k_i again denotes the strength of node i. In (3.2), P_{ij} denotes the components of a *null model* matrix, which specifies the relative value of intra-community edges in assessing when communities are closely connected [8, 77].

- GOAL: Assign nodes to communities to maximize modularity.

Community Detection: Null Models

- Erdös-Rényi (Bernoulli)

$$P_{ij} = p$$

- Newman-Girvan*

$$P_{ij} = \gamma \frac{k_i k_j}{2W}$$

- Leicht-Newman* (directed)

$$P_{ij} = \gamma \frac{k_i^{in} k_j^{out}}{W}$$

- Barber* (bipartite)

$$P_{ij} = \begin{cases} \gamma \frac{k_i d_j}{W} \\ 0 \end{cases}$$

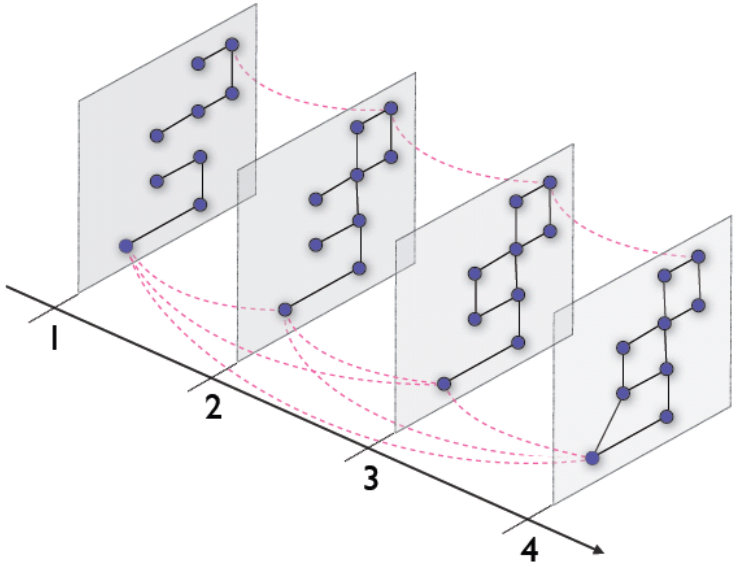
** with resolution parameters γ*

Community Detection: Computational Heuristics

$$Q = \frac{1}{2W} \sum_{i,j} B_{ij} \delta(C_i, C_j) \,, \;\; B_{ij} = A_{ij} - P_{ij}$$

- Cannot guarantee optimal quality without full enumeration of possible partitions
  - *NP-complete problem*
  - Many algorithms available (simulated annealing, etc.)
  - Need to pick null model appropriate to problem
  - Extreme degeneracies in good local optima of Q
    - (B. H. Good, Y.-A. de Montjoye, & A. Clauset, to appear in *PRE*; see Aaron's talk on Wednesday)

Multislice Networks



- Typical formulation for studying networks: Static networks, with a single kind of tie, partitioned at a single spatial resolution
  - Also potentially sweep over multiple resolutions (or occasionally multiple times) but in an ad hoc fashion
- Multislice framework: dynamic, multiplex, and with communities at multiple scales
- Simple idea: Glue common individuals across “slices”

What is the appropriate null model?

$$Q = \frac{1}{2W} \sum_{i,j} B_{ij} \delta(C_i, C_j) \,, \; B_{ij} = A_{ij} - P_{ij}$$

- Each slice is a network (static, single type) with a specified spatial resolution of interest
- Different slices can mean: different value of resolution parameter, different time snapshot, different type of connection
- Have both intra-slice edges & inter-slice edges
- How to choose a null model?

Quality of Partition via "Stability"

- Idea: use a dynamical process on a network to learn about network structure
  - We build on work of R. Lambiotte, J.-C. Delvenne, & M. Barahona [*arXiv:0812.1770*]
- Quality of a network partition expressed in terms of its "stability" (autocovariance function of an ergodic Markov process on the network):

$$R_{\mathcal{M}}(t) = \sum_{C \in \mathcal{P}} P(C,t) - P(C,\infty)$$

  - P(C,t) = probability, for a given community C, for a random walker to be in that community both initially and at time t
- Stability measures the quality of a partition in terms of the persistence of the dynamics by giving a positive contribution to communities from which a random walker is unlikely to escape with a given time t

Laplacian Dynamics (i.e., random walks)

- Lambiotte, Delvenne, & Barahona [*arXiv: 0812.1770*] *derived* modularity from normalized Laplacian dynamics

$$\dot{p}_i = \sum_j \frac{A_{ij}}{k_j} p_j - p_i, \qquad p_i^* = k_i/2m.$$

$$R_{\mathrm{NL}}(t) = \sum_C \sum_{i,j \in C} \left[\left(e^{t(B-I)}\right)_{ij} \frac{k_j}{2m} - \frac{k_i}{2m}\frac{k_j}{2m} \right]. \qquad B_{ij} = A_{ij}/k_j$$

Expansion of matrix exponential to first-order in t recovers Newman-Girvan modularity with resolution $\gamma = 1/t$.

Question: How do we apply this idea to multislice networks?

Generalized Laplacian Dynamics

a) Calculate (to first order in t) the probability of observing a tie between nodes i and j, *conditional on the type of connection necessary to move j → i*

b) Generalize dynamics to include motion along different types of edges

c) Different spreading weights on different types of edges

Multislice Networks

$$k_{js} = \sum_i A_{ijs}, c_{js} = \sum_r C_{jsr}, \ \kappa_{js} = k_{js} + c_{js}.$$

$$\dot{p}_{is} = \sum_{jr} (A_{ijs}\delta_{sr} + \delta_{ij} C_{jsr}) p_{jr}/\kappa_{jr} - p_{is}$$

$$\sum_{ij} \left[(\delta_{ij} + tL_{ij})\, p_j^* - \rho_{i|j} p_j^* \right] \delta(c_i, c_j). \quad \gamma = 1/t$$

$$p_{jr}^* = \kappa_{jr}/(2\mu), \text{ where } 2\mu = \sum_{jr} \kappa_{jr}.$$

$$\rho_{is|jr} p_{jr}^* = \left[\frac{k_{is}}{2m_s} \frac{k_{jr}}{\kappa_{jr}} \delta_{sr} + \frac{C_{jsr}}{c_{jr}} \frac{c_{jr}}{\kappa_{jr}} \delta_{ij} \right] \frac{\kappa_{jr}}{2\mu}$$

$$Q_{\text{multislice}} = \frac{1}{2\mu} \sum_{ijsr} \left\{ \left(A_{ijs} - \gamma_s \frac{k_{is} k_{js}}{2m_s} \delta_{sr} \right) + \delta_{ij} C_{jsr} \right\} \delta(c_{is}, c_{jr})$$

Special Case: Bipartite Networks

- Recover Barber null model with resolution parameter:

$$\sum_{ij} \left[(\delta_{ij} + tL_{ij})\, p_j^* - \rho_{i|j} p_j^* \right] \delta(c_i, c_j).$$

$$L_{ij} = A_{ij}/k_j - \delta_{ij} \qquad\qquad p_j^* = k_j/(2m)$$

$$\rho_{i|j} = b_{ij} k_i / m, \qquad\qquad \gamma = 1/t$$

$$Q_{\text{bipartite}} = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \gamma b_{ij} \frac{k_i k_j}{m} \right] \delta(c_i, c_j)\,,$$

Special Case: Directed Networks

- Recover Leicht-Newman null model with a resolution parameter:

$$\sum_{ij} \left[(\delta_{ij} + tL_{ij})\, p_j^* - \rho_{i|j} p_j^* \right] \delta(c_i, c_j).$$

$$\dot{p}_i = \sum_j L_{ij} p_j = \sum_j \frac{1}{k_j} (A_{ij} + A_{ji}) p_j - p_i.$$

$$p_j^* = k_j/(2m) \qquad k_j = k_j^{\text{in}} + k_j^{\text{out}}$$

$$\rho_{i|j} p_j^* = \left(\frac{k_i^{\text{in}}}{m} \frac{k_j^{\text{out}}}{k_j} + \frac{k_i^{\text{out}}}{m} \frac{k_j^{\text{in}}}{k_j} \right) \frac{k_j}{2m} = \frac{k_i^{\text{in}} k_j^{\text{out}} + k_i^{\text{out}} k_j^{\text{in}}}{2m^2},$$

$\gamma = 1/t$

$$Q_{\text{directed}} = \frac{1}{m} \sum_{ij} \left[A_{ij} - \gamma \frac{k_i^{\text{in}} k_j^{\text{out}}}{m} \right] \delta(c_i, c_j),$$

Special Case: Signed Networks

- Recover null models of Traag *et al*. & Gomez *et al*.:

$$\sum_{ij} \left[\left(\delta_{ij} + tL_{ij} \right) p_j^* - \rho_{i|j} p_j^* \right] \delta(c_i, c_j).$$

$$L_{ij} = (A_{ij}^+ + A_{ij}^-)/k_j - \delta_{ij} \text{ (with } k_j = k_j^+ + k_j^- \text{)}$$

$$\rho_{i|j} p_j^* = \left(\frac{k_i^+}{2m^+} \frac{k_j^+}{k_j} - \frac{k_i^-}{2m^-} \frac{k_j^-}{k_j} \right) \frac{k_j}{2m} = \frac{1}{2m} \left(\frac{k_i^+ k_j^+}{2m^+} - \frac{k_i^- k_j^-}{2m^-} \right)$$

$\gamma = 1/t$

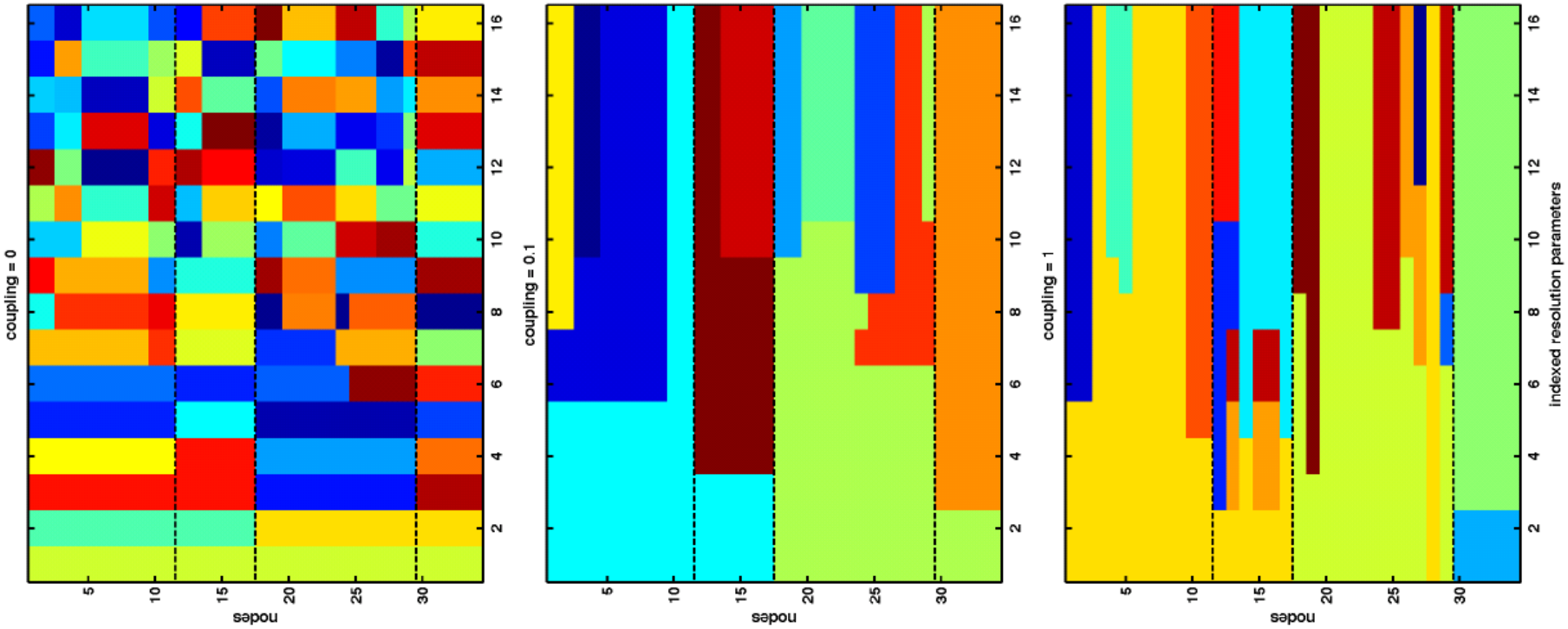
$$Q_{\text{signed}} = \frac{1}{2m} \sum_{ij} \left[A_{ij}^+ - A_{ij}^- - \left(\gamma^+ \frac{k_i^+ k_j^+}{2m^+} - \gamma^- \frac{k_i^- k_j^-}{2m^-} \right) \right] \delta(c_i, c_j)$$

Examples

- (Revenge of the) Zachary Karate Club
  - Multiple resolution parameter values at once ("multiscale")
- Tastes, Ties, & Time
  - Multiplex (multiple edge types)
- 200 years of roll call votes in U.S. Senate
  - Time-dependent

$$Q_{\text{multislice}} = \frac{1}{2\mu} \sum_{ijsr} \left\{ \left(A_{ijs} - \gamma_s \frac{k_{is} k_{js}}{2m_s} \delta_{sr} \right) + \delta_{ij} C_{jsr} \right\} \delta(c_{is}, c_{jr})$$

Zachary Karate Club



$$C_{jsr} = \{0, \omega\}$$

Tastes, Ties, & Time

- Data from Lewis *et al*. 2008
  - "not-Harvard data set"
- First wave of private northeastern school
- Edge types:
  - Facebook friends
  - Picture friends
  - Roommates
  - Housing Groups

ω	#Communities	#Communities per Individual			
		1	2	3	4
0	1036	0	0	0	1640
0.1	122	230	664	611	135
0.2	66	326	805	415	94
0.3	49	430	792	354	64
0.4	36	522	770	302	46
0.5	31	645	695	276	24
1	16	1640	0	0	0

Roll Call Voting Networks

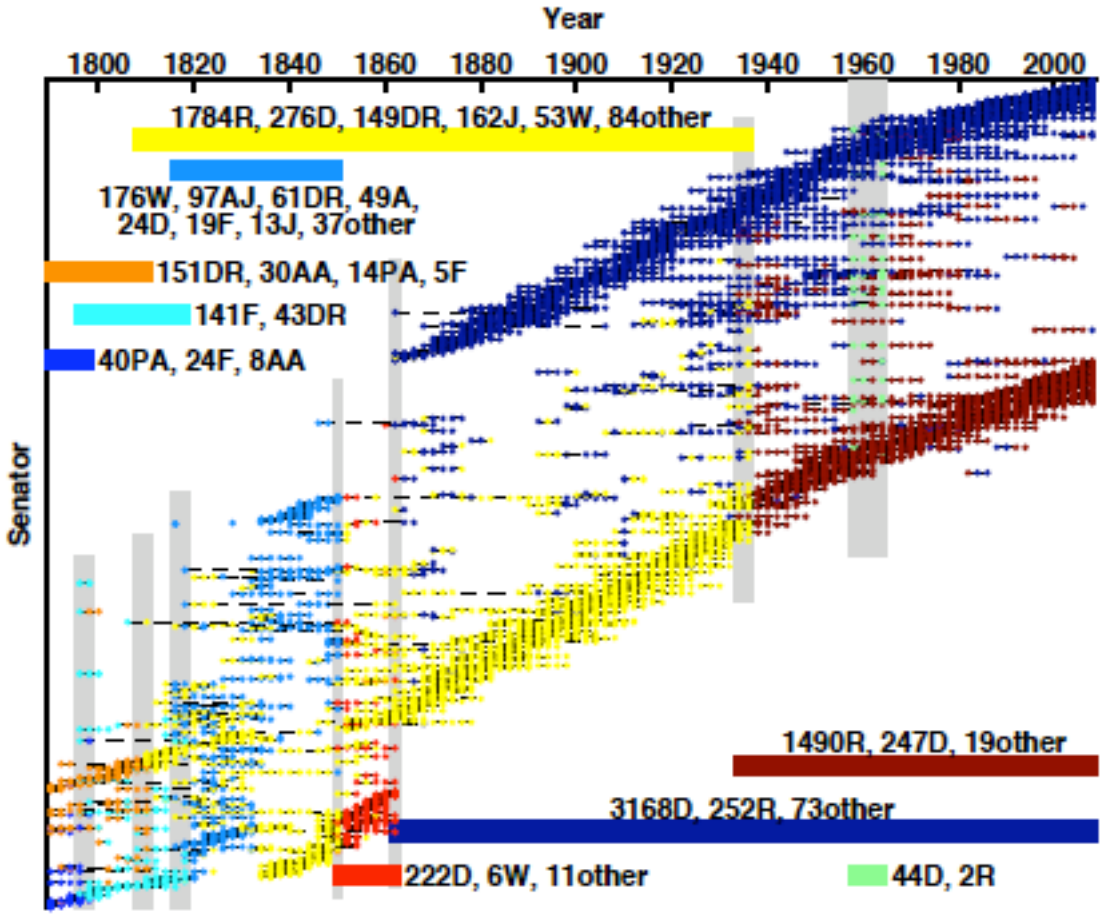
into an $n \times n$ *adjacency matrix* A, with elements $A_{ij} \in [0, 1]$ representing the extent of voting agreement between legislators i and j, with elements defined here by

$$A_{ij} = \frac{1}{b_{ij}} \sum_k \alpha_{ijk} \,, \tag{1}$$

where α_{ijk} equals 1 if legislators i and j voted the same on bill k and 0 otherwise and b_{ij} is the total number of bills on which both legislators voted. The matrix A encodes a network of weighted affiliations between legislators, with weights determined by the similarity of their roll-call records

- A. S. Waugh, L. Pei, J. H. Fowler, P. J. Mucha, & M. A. Porter, *arXiv:0907.3509* (without multislice formulation)
- Modularity as a measure of polarization
- Modularity as a predictor of majority turnover (the "partial polarization hypothesis")
- One network slice for each two-year Congress

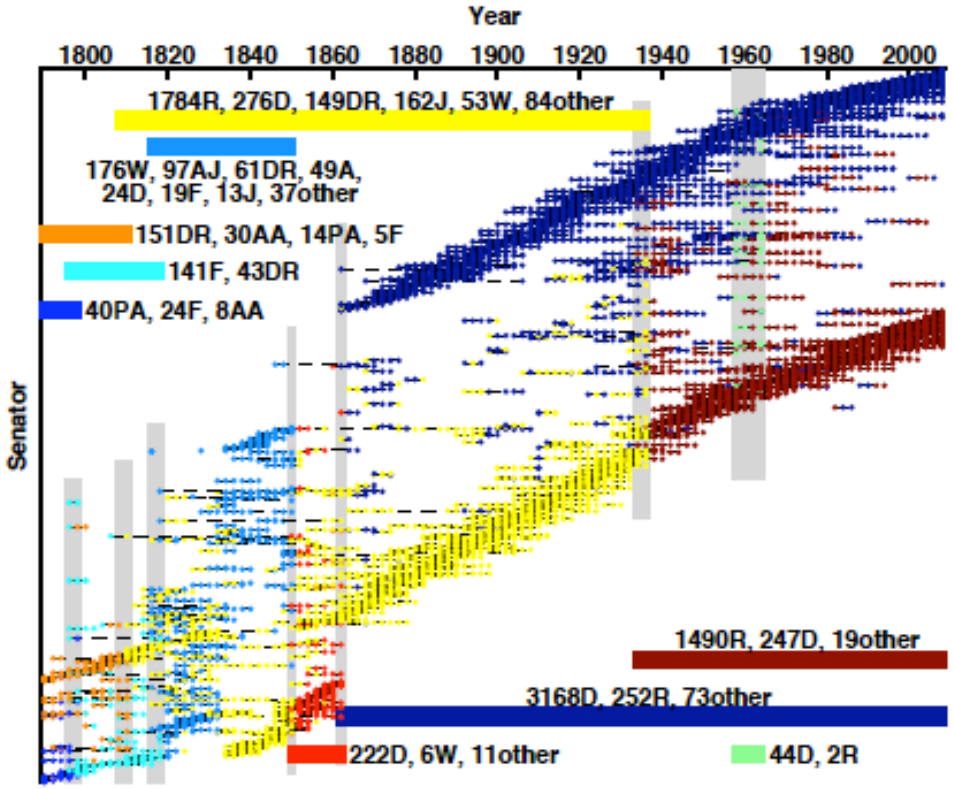
110 Senates (220 years)



Nominal party affiliations:

- Pro-Administration (PA)
- Anti-Administration (AA)
- Federalist (F)
- Democratic-Republican (DR)
- Whig (W)
- Anti-Jackson (AJ)
- Adams (A)
- Jackson (J)
- Democratic (D)
- Republican (R)

110 Senates

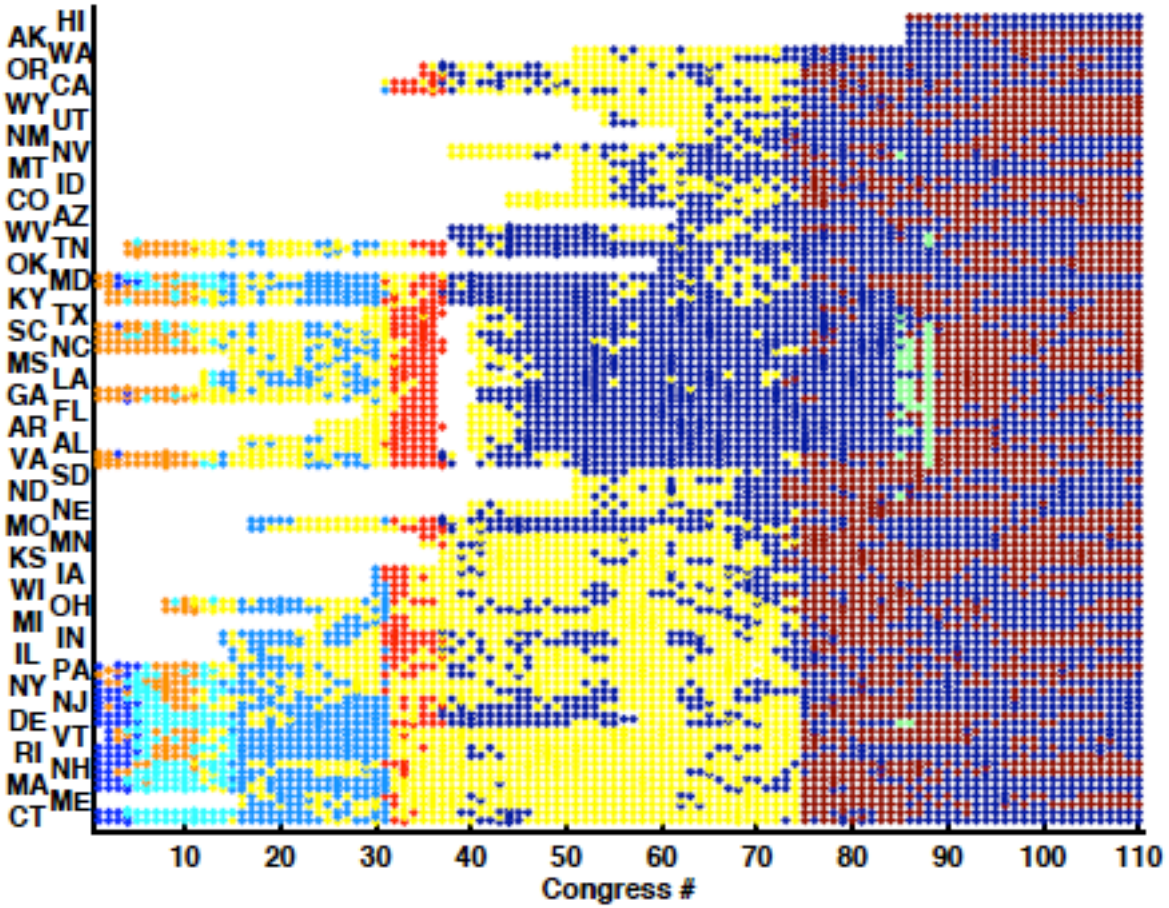


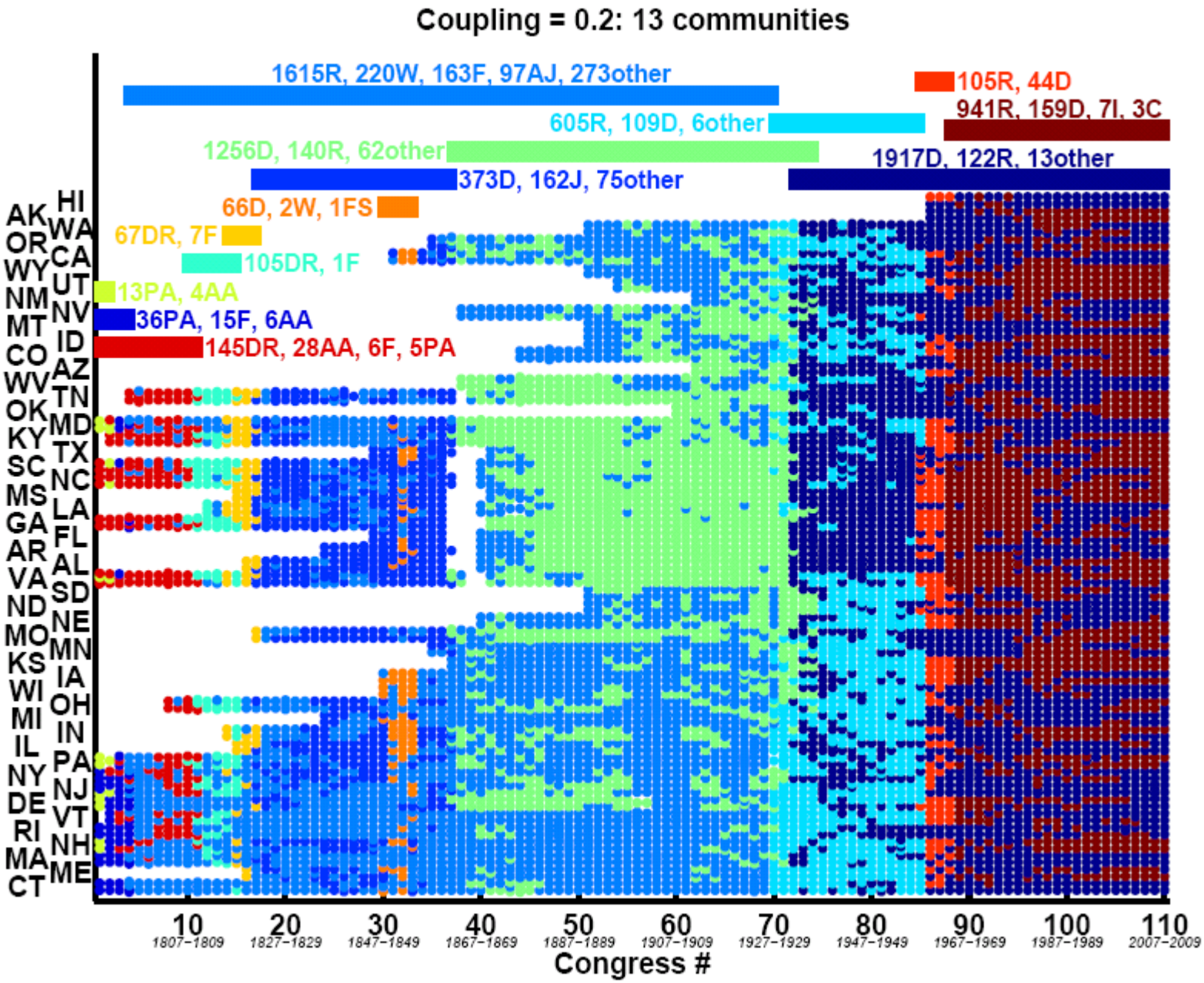
Gray areas:

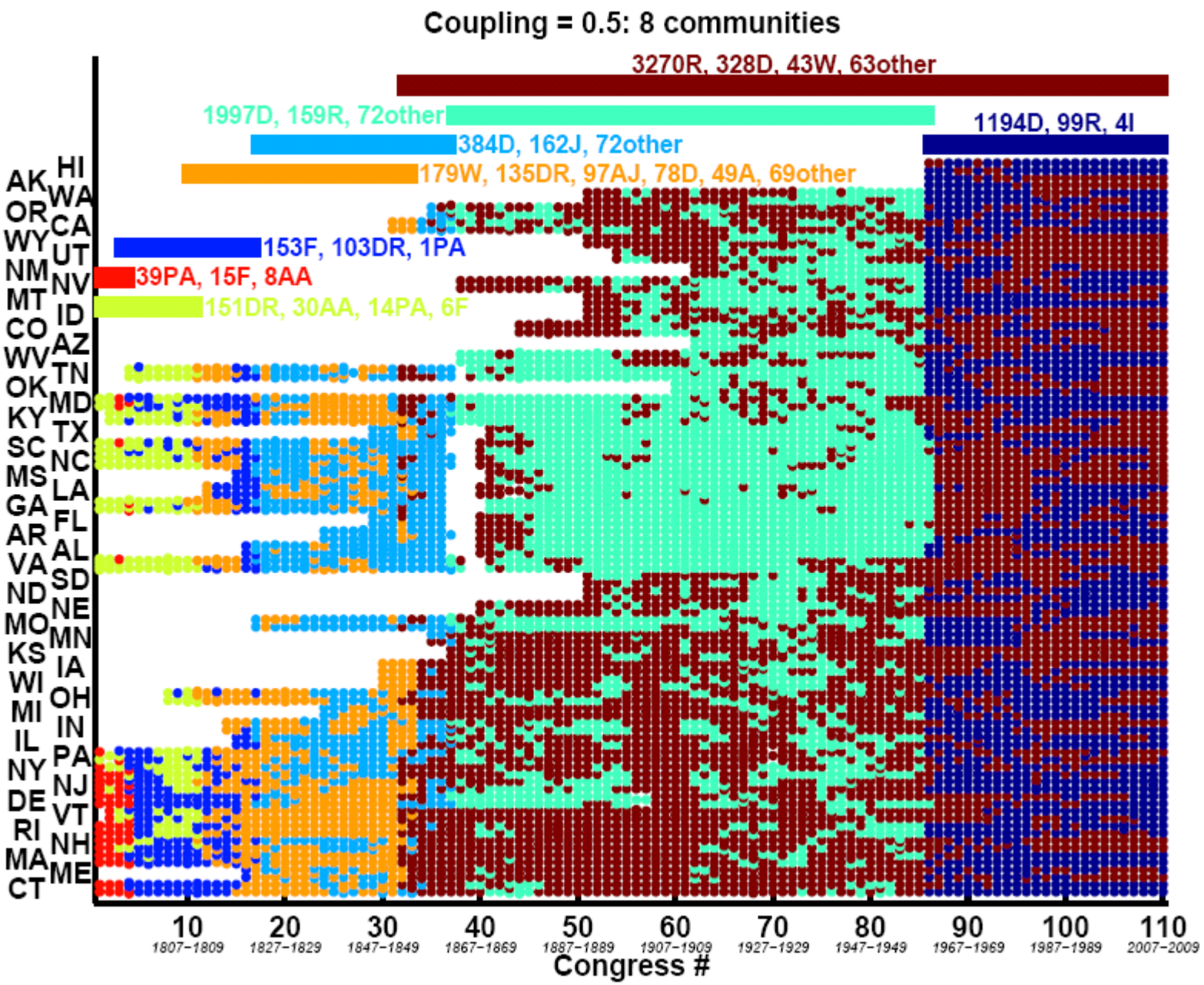
- 4th and 5th: First with political parties
- 10th and 11th: Vice President Aaron Burr's indictment for treason
- 14th and 15th: Changing structures in Democratic-Republican party
- 31st: Compromise of 1850
- 37th: Beginning of the American Civil War
- 73rd and 74th: Landslide 1932 election amidst the Great Depression
- 85th to 88th: Brought the major American civil rights acts

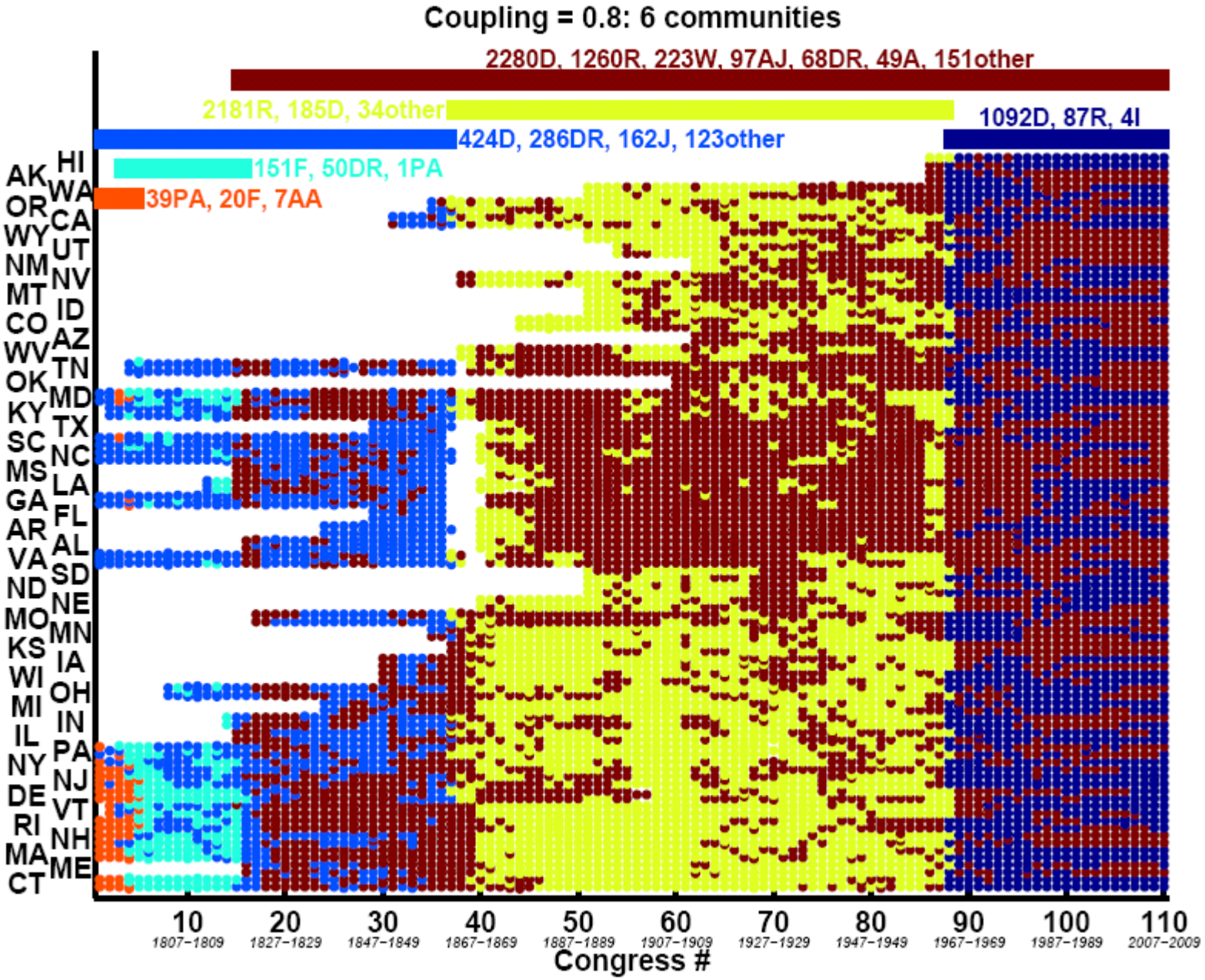
Gray areas: 3 communities exist at the same time (9 communities in total; $\omega = 0.5$)

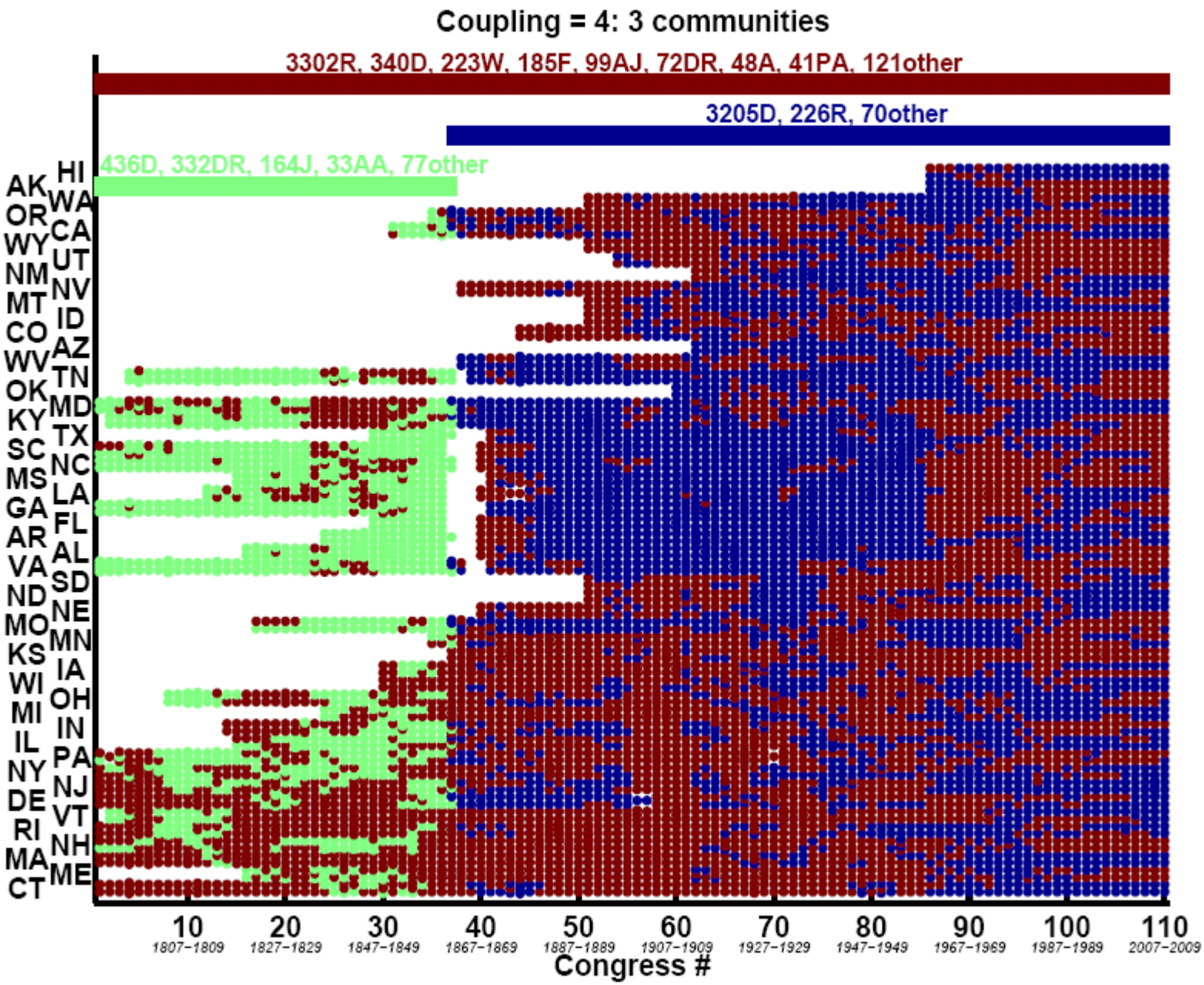
Arranged by state…




Coupling = 0.2: 13 communities
1615R, 220W, 163F, 97AJ, 273other
105R, 44D
941R, 159D, 7I, 3C
605R, 109D, 6other
1256D, 140R, 62other
1917D, 122R, 13other
373D, 162J, 75other
66D, 2W, 1FS
67DR, 7F
105DR, 1F
13PA, 4AA
36PA, 15F, 6AA
145DR, 28AA, 6F, 5PA
HI
AK
WA
OR
CA
WY
UT
NM
NV
MT
ID
CO
AZ
WV
TN
OK
MD
KY
TX
SC
NC
MS
LA
GA
FL
AR
AL
VA
SD
ND
NE
MO
MN
KS
IA
WI
OH
MI
IN
IL
PA
NY
NJ
DE
VT
RI
NH
MA
ME
CT
10
1807–1809
20
1827–1829
30
1847–1849
40
1867–1869
50
1887–1889
60
1907–1909
70
1927–1929
80
1947–1949
90
1967–1969
100
1987–1989
110
2007–2009
Congress #


Coupling = 0.5: 8 communities
3270R, 328D, 43W, 63other
1997D, 159R, 72other
1194D, 99R, 4I
384D, 162J, 72other
179W, 135DR, 97AJ, 78D, 49A, 69other
153F, 103DR, 1PA
39PA, 15F, 8AA
151DR, 30AA, 14PA, 6F
HI
AK WA
OR CA
WY UT
NM NV
MT ID
CO AZ
WV TN
OK MD
KY TX
SC NC
MS LA
GA FL
AR AL
VA SD
ND NE
MO MN
KS IA
WI OH
MI IN
IL PA
NY NJ
DE VT
RI NH
MA ME
CT
10 1807–1809
20 1827–1829
30 1847–1849
40 1867–1869
50 1887–1889
60 1907–1909
70 1927–1929
80 1947–1949
90 1967–1969
100 1987–1989
110 2007–2009
Congress #


Coupling = 0.8: 6 communities
2280D, 1260R, 223W, 97AJ, 68DR, 49A, 151other
2181R, 185D, 34other
1092D, 87R, 4I
424D, 286DR, 162J, 123other
151F, 50DR, 1PA
39PA, 20F, 7AA
HI
AK WA
OR CA
WY UT
NM NV
MT ID
CO AZ
WV TN
OK MD
KY TX
SC NC
MS LA
GA FL
AR AL
VA SD
ND NE
MO MN
KS IA
WI OH
MI IN
IL PA
NY NJ
DE VT
RI NH
MA ME
CT
10 1807–1809
20 1827–1829
30 1847–1849
40 1867–1869
50 1887–1889
60 1907–1909
70 1927–1929
80 1947–1949
90 1967–1969
100 1987–1989
110 2007–2009
Congress #


Coupling = 4: 3 communities
3302R, 340D, 223W, 185F, 99AJ, 72DR, 48A, 41PA, 121other
3205D, 226R, 70other
436D, 332DR, 164J, 33AA, 77other
HI
AK
WA
OR
CA
WY
UT
NM
NV
MT
ID
CO
AZ
WV
TN
OK
MD
KY
TX
SC
NC
MS
LA
GA
FL
AR
AL
VA
SD
ND
NE
MO
MN
KS
IA
WI
OH
MI
IN
IL
PA
NY
NJ
DE
VT
RI
NH
MA
ME
CT
10
1807–1809
20
1827–1829
30
1847–1849
40
1867–1869
50
1887–1889
60
1907–1909
70
1927–1929
80
1947–1949
90
1967–1969
100
1987–1989
110
2007–2009
Congress #

Conclusions

- "Multislice" framework generalizes the investigation of commmunity structure to more complicated, more realistic, and much more interesting situations: ***dynamic/longitudinal*** data, ***multiplex*** ties, and communities across ***multiple scales***.
- Visualization tools for graphs that incorporate community structure: http://netwiki.amath.unc.edu/VisComms/VisComms
- Current efforts: Apply multislice community detection to epidemics, politicians, and brains
  - Other ideas: new null models, choice of inter-slice coupling, generalizations of clustering coefficients and other quantities to multislice networks, …

Binary Trees