Communities in Networks

MASON A. PORTER
MATHEMATICAL INSTITUTE
UNIVERSITY OF OXFORD

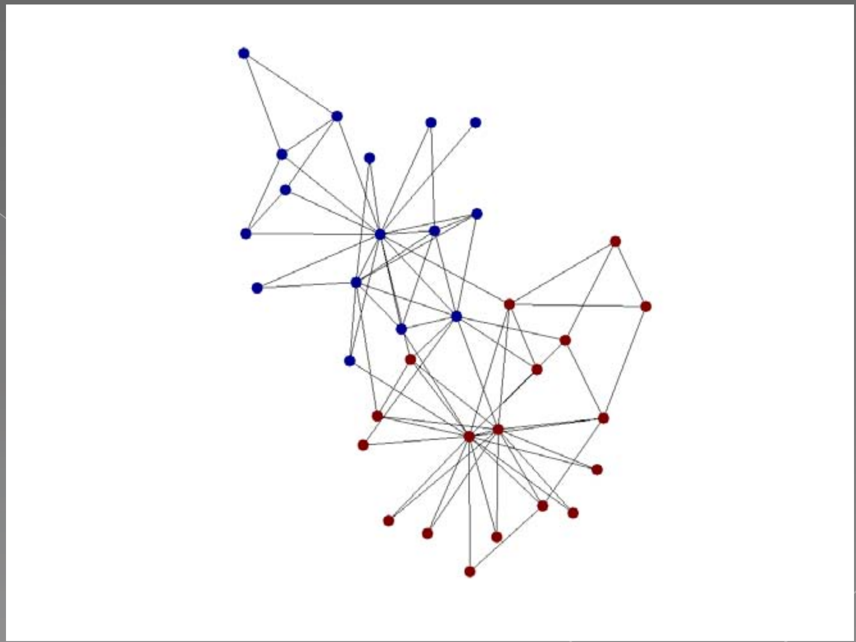
SOME COLLABORATORS: Dani Bassett, Jean Carlson, Karen Daniels, Charlotte Deane, Dan Fenn, James Fowler, James Gleeson, Scott Grafton, Lucas Jeub, Nick Jones, Eric Kelsic, Anna Lewis, Kevin Macon, Sergey Melnik, Peter Mucha, Sean Myers, JP Onnela, Eli Owens, Stephen Reid, Puck Rombach, Thomas Richardson, Puck Rombach, Amanda Traud, Brian Uzzi, Jonathan Ward, Nicholas Wymbs

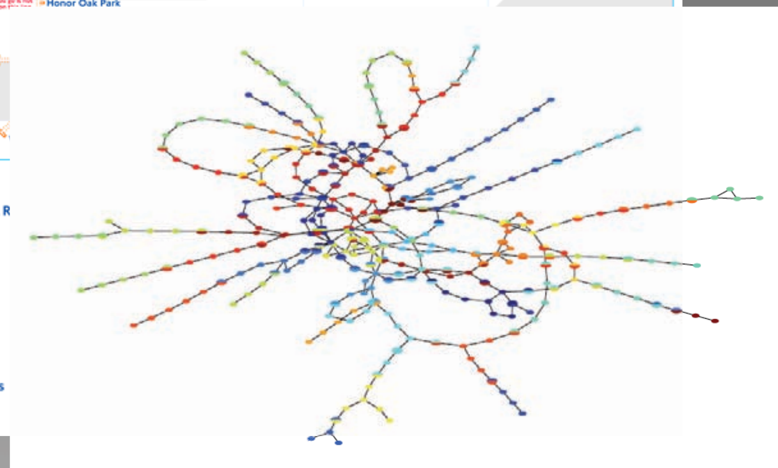
Outline

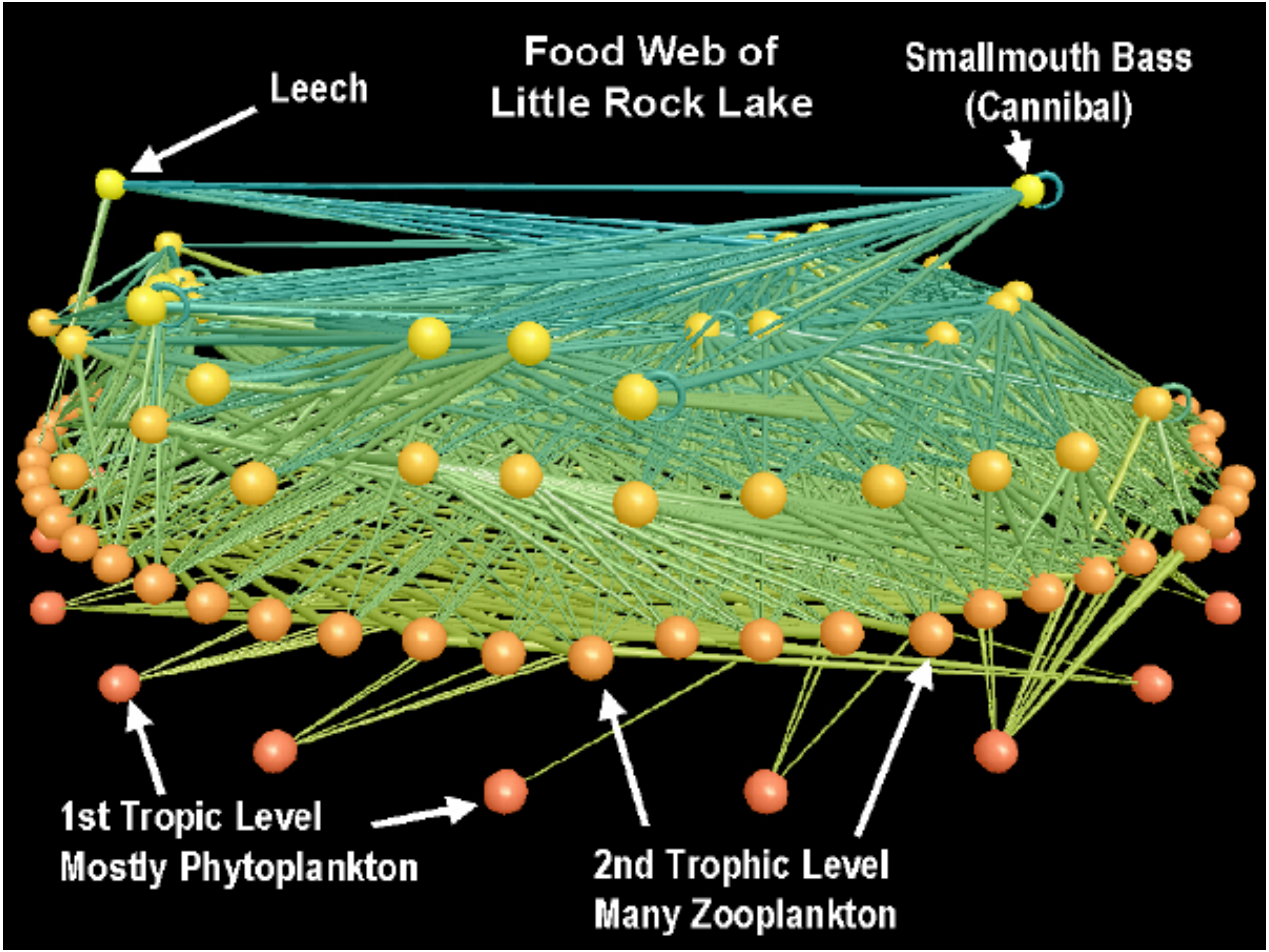
- Introduction
- Communities in Networks
- Example: Facebook Networks
- Communities, Dynamics, and Function: A Whirlwind Tour
- Conclusions

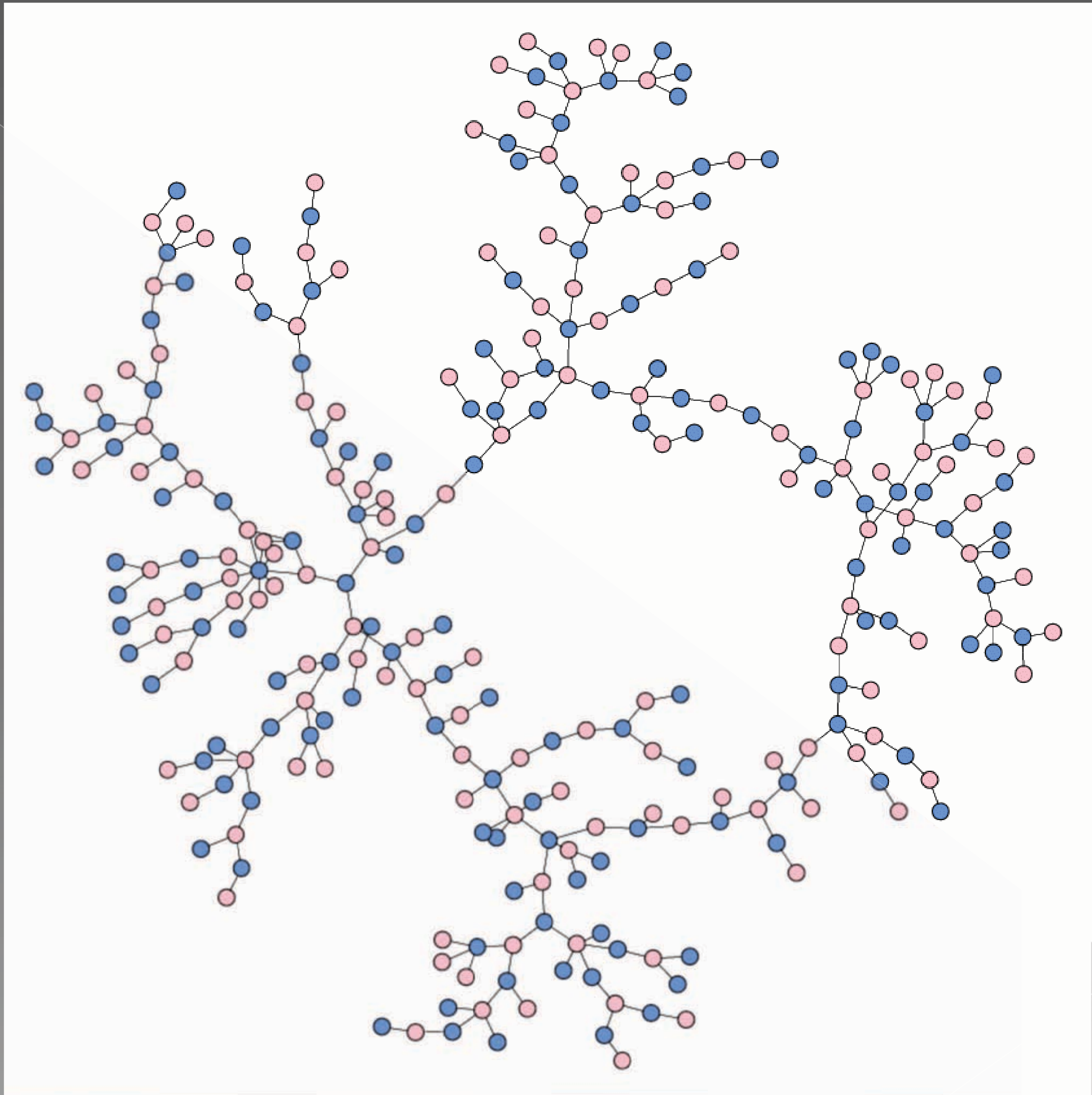
What is a Network?

- A network consists of *nodes* representing agents connected by *edges* representing ties
  - Binary edges: 0 or 1
  - Weighted edges
  - Directed edges
  - Bipartite networks
  - Time-dependence
  - Multiplexity
  - Hypergraphs
  - Spatially embedded networks



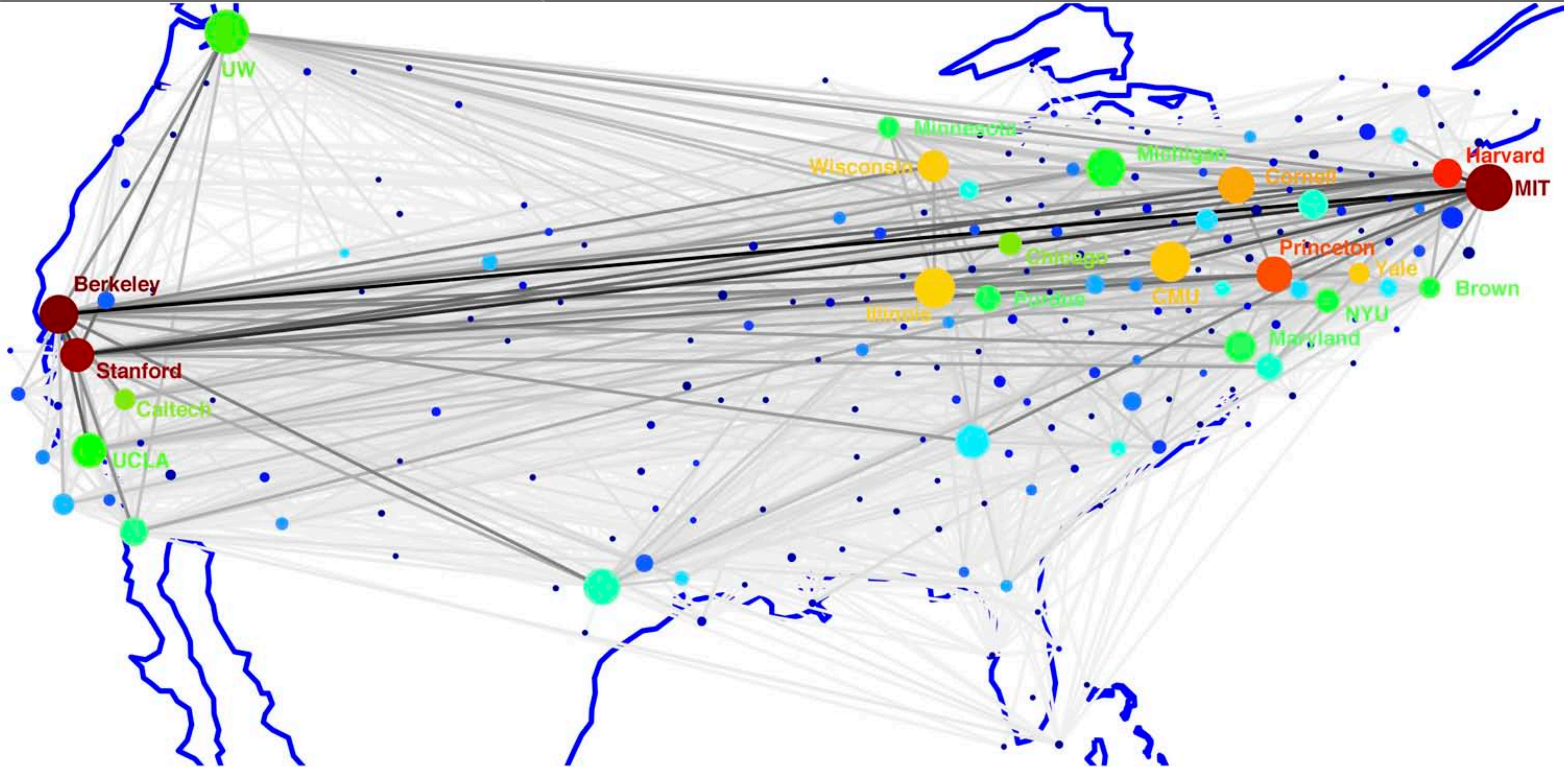

Victoria line closes from 2200 on Monday to Thursday evenings until 20 November 2008
East London line is closed for major line extension work to become part of the London Overground network.
Improvement works may affect your journey, particularly at weekends. Check before you travel; look for publicity at stations, visit tfl.gov.uk/check or call 020 7222 1234
MAYOR OF LONDON
River Thames
Tube map
Bakerloo
Central
Restricted service
Circle
District
Hammersmith & City
Jubilee
Metropolitan
Restricted service
Northern
Restricted service
Piccadilly
Victoria
Waterloo & City
London Overground
DLR
Replacement bus service
Interchange stations
Step-free access from the platform to the street
Connections with National Rail
Check before you travel. See index below
Connections with riverboat services
Connection with Tramlink
Location of Airport
Interchange with National R services to airport
Replacement bus services
Bicycle parking
Car parks
Toilets on site/nearby
Travel Information Centres



Food Web of
Little Rock Lake
Leech
Smallmouth Bass
(Cannibal)
1st Tropic Level
Mostly Phytoplankton
2nd Trophic Level
Many Zooplankton



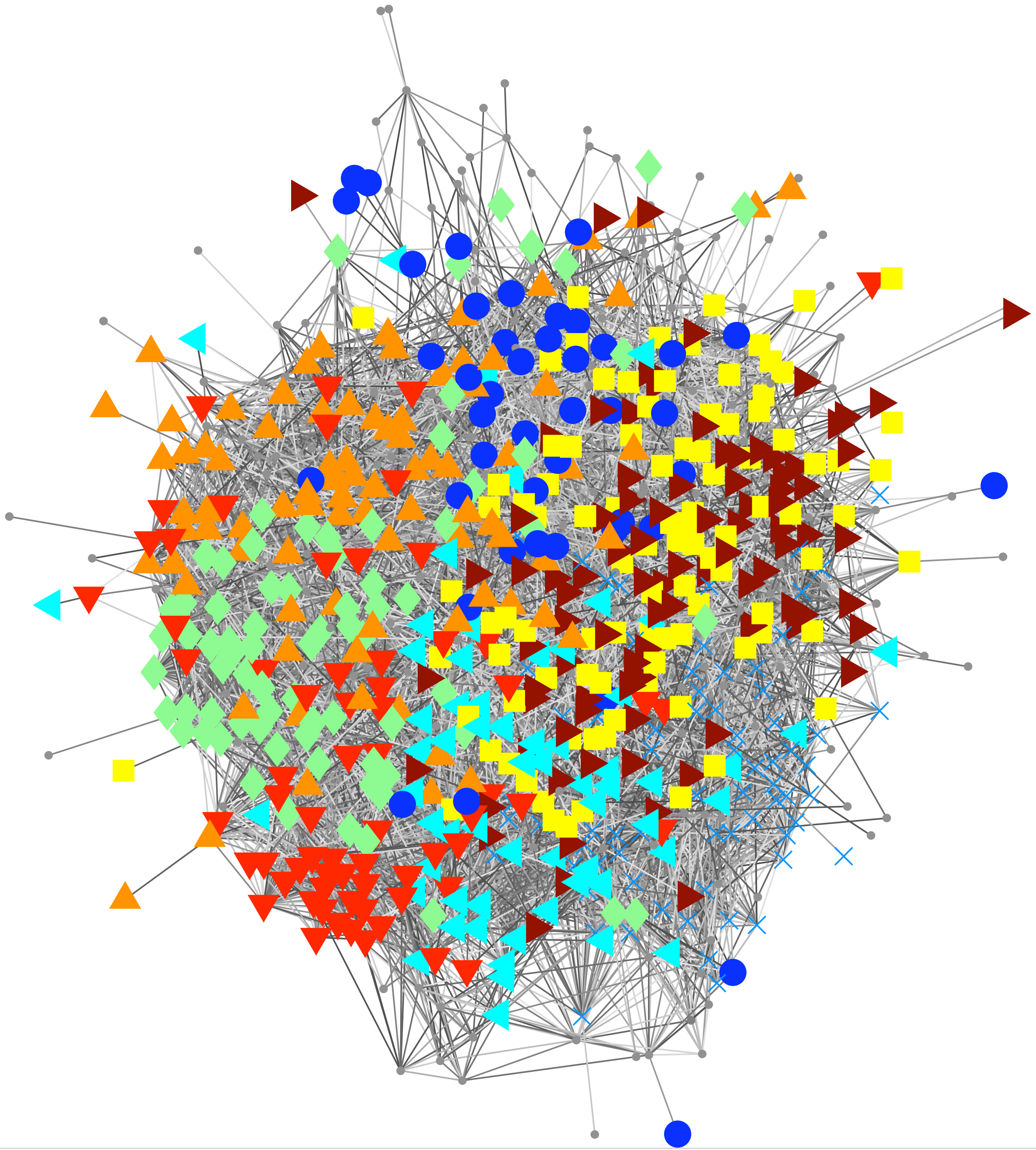
Mathematical Genealogy Network

S. Myers, P. J. Mucha, and MAP [2011]. "Mathematical Genealogy & Department Prestige", *Chaos* **21**(4): 041104 (Gallery of Nonlinear Images).



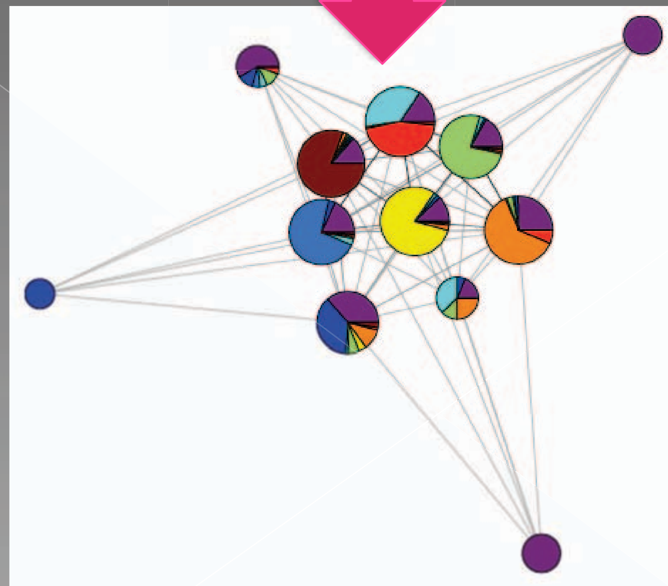
Suburban Bacon Network

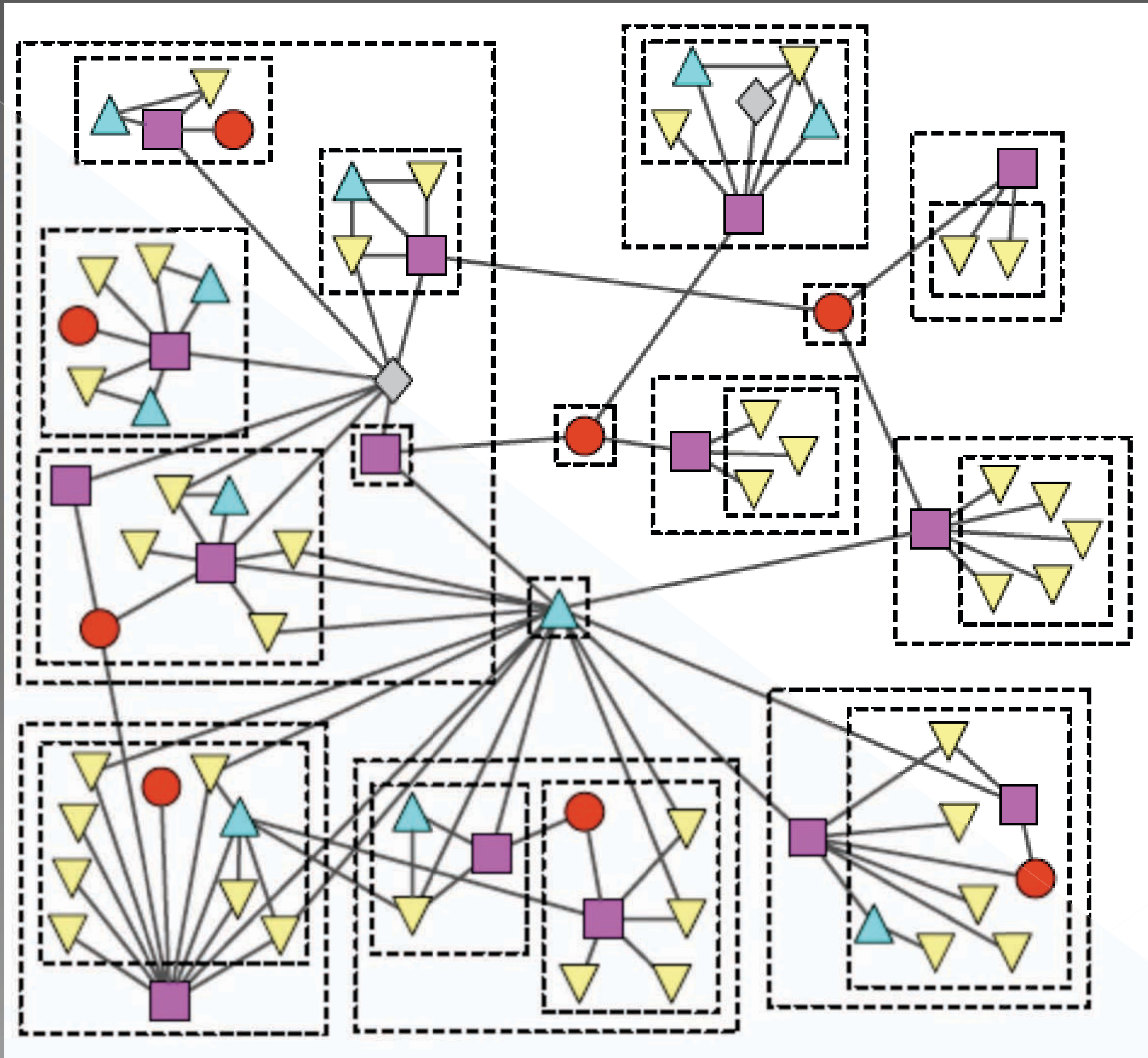
Consumers buying different brands of bacon (suburban dataset)
424 different consumers and 79 sorts of bacon

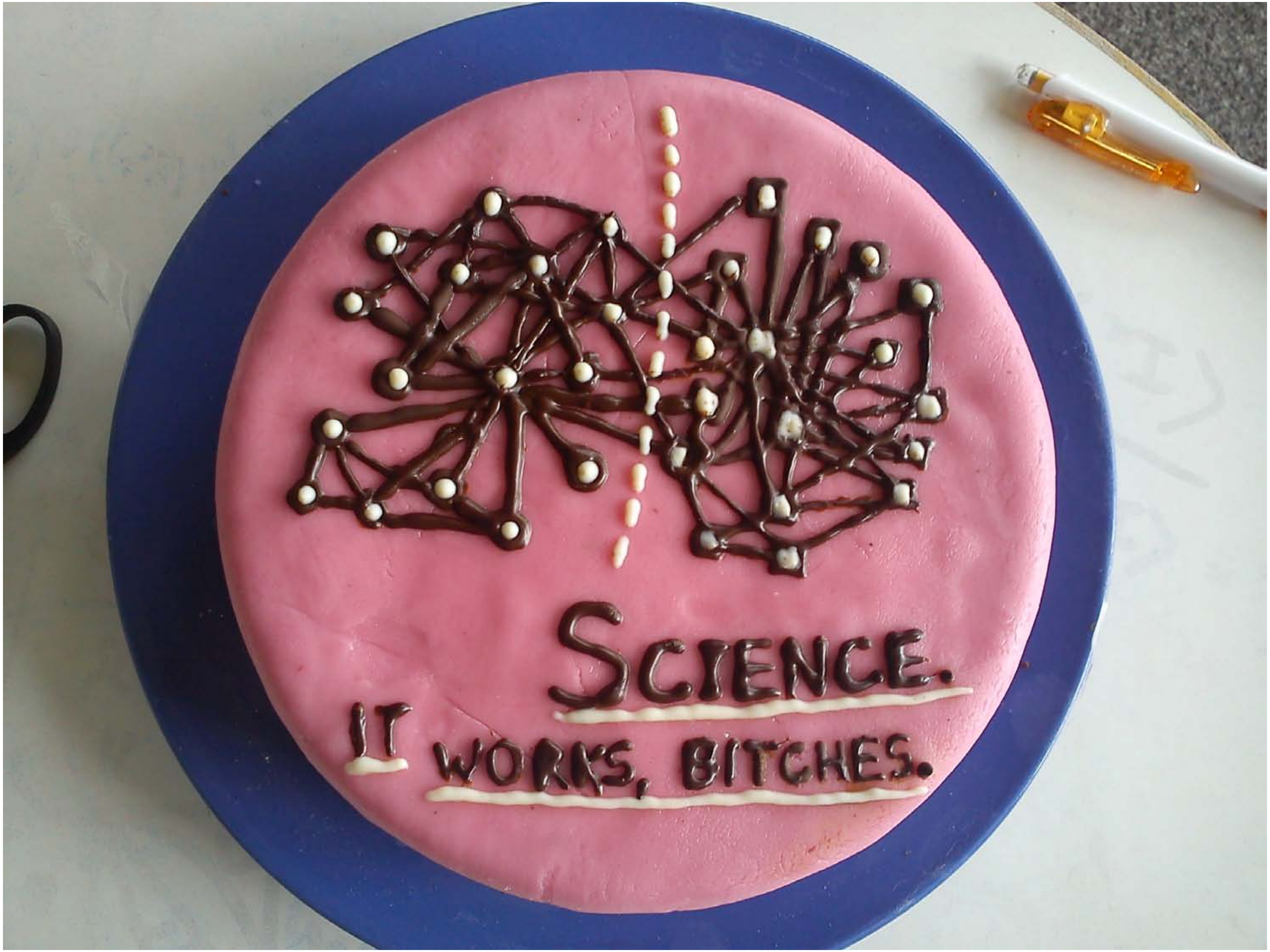


Network Communities

- Communities = cohesive groups/modules/ mesoscopic structures
  - In stat phys, one tries to derive macroscopic and mesoscopic insights from microscopic information
- Community structure is *both* modular and hierarchical
- Communities have larger density of internal ties relative to some null model for what ties are present at random
  - Modularity






SCIENCE.
IT WORKS, BITCHES.

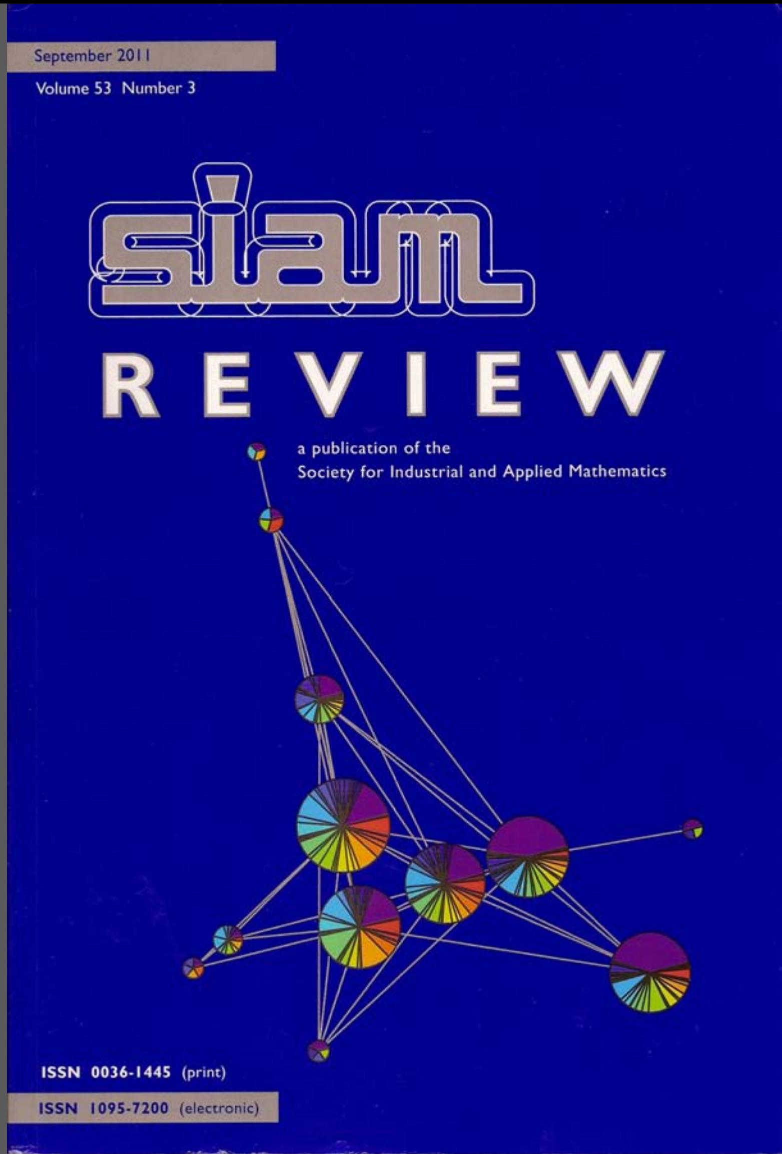
Detecting Communities

- Survey article
  - MAP, J.-P. Onnela, & P. J. Mucha [2009], *Notices of the American Mathematical Society* **56**(9): 1082-1097, 1164-1166
- Types of methods
  - Agglomerative
    - e.g., linkage clustering
  - Divisive
    - e.g., partitioning by optimizing modularity or using centrality-based methods (such as Girvan-Newman algorithm)
  - Local methods
    - e.g., k-clique percolation
  - Edge-based
    - e.g., Y.Y. Ahn's talk

Modularity and the Potts Method

- Minimize: $$H = -\sum_{ij} J_{ij}\delta(\sigma_i, \sigma_j)$$
  - Potts Hamiltonian
    - $\square_i$ = community assignment (spin state) of node i
    - $J_{ij} > 0$ → "ferromagnetic" interaction between i & j → nodes i and j try to be in the same state
    - $J_{ij} < 0$ → "antiferromagnetic" interaction between i & j → nodes i and j try to be in different states
- Modularity optimization: $$J_{ij} = \frac{A_{ij} - p_{ij}}{W}$$
  - A_{ij} = adjacency matrix
  - $W = (1/2)\Sigma_{ij}A_{ij}$ = sum of all edge weights
  - p_{ij} = prob(i connected to j) in null model
    - Newman-Girvan: $p_{ij} = k_i k_j/(2W)$, where $k_i = \Sigma_j A_{ij}$ = total edge weight of node i
    - "Resolution parameter": use $\lambda * p_{ij}$

EXAMPLE: FACEBOOK NETWORKS



A. L. Traud, E. D. Kelsic, PJM, & MAP [2011], SIAM Review, **53**(3): 526—543
ALT, C. Frost, PJM, & MAP [2009], Chaos 19(4): 041104 (Gallery of Nonlinear Images)
ALT, PJM, & MAP [2012], *Physica A* **391**(16): 4165—4180

Facebook Networks

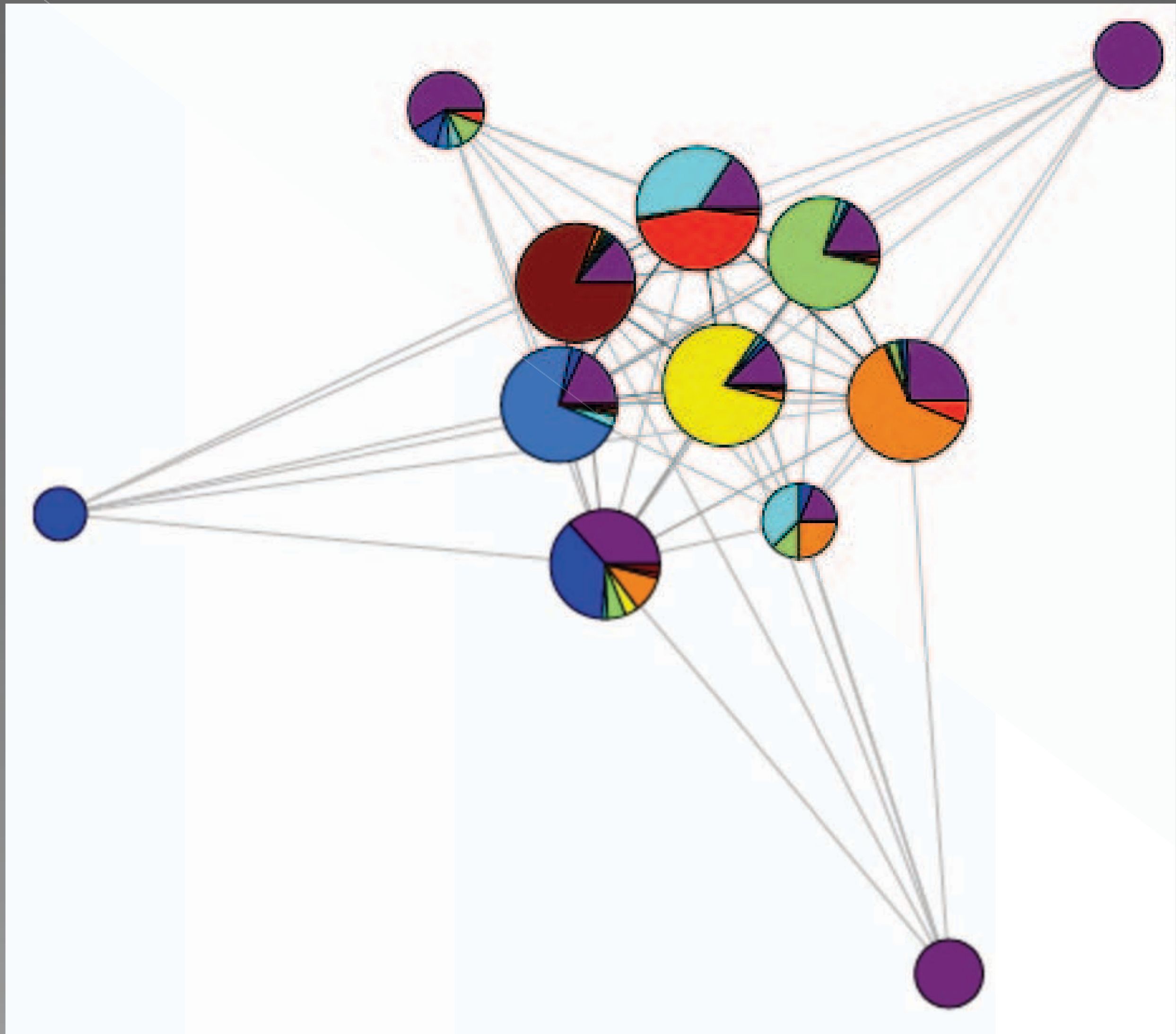
- Nodes = individuals
- Edges = self-identified friendships (1 or 0)
- Our data
  - 100 different universities (full networks)
  - Single-time snapshot: September 2005
    - Facebook was university-only
  - Self-reported demographics
    - Gender, class year, high school, major, dormitory/"house"
  - Provided by Adam D'Angelo & Facebook

Example Networks

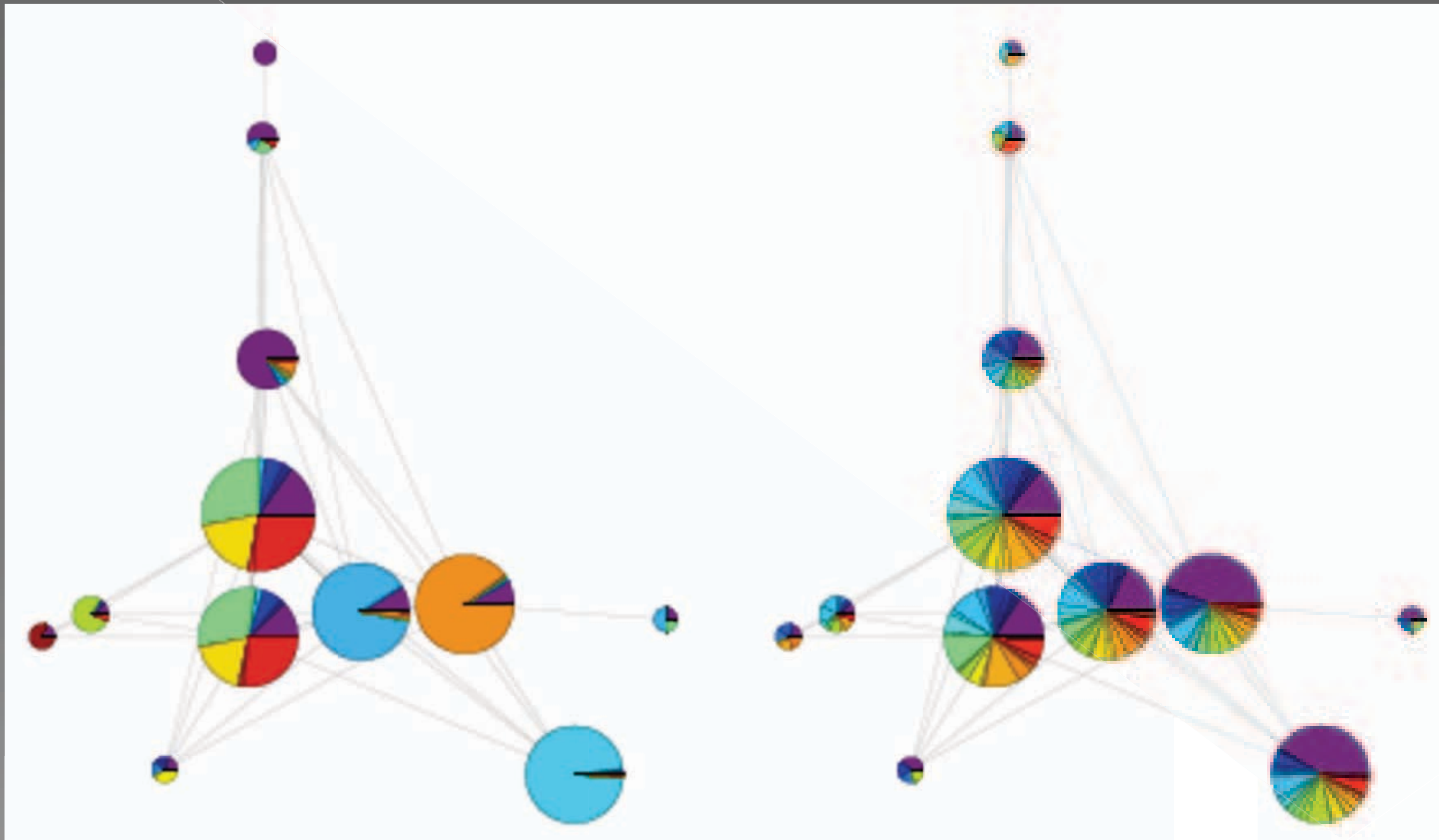
FULL NETWORKS (SINGLE UNIVERSITY, LARGEST CONNECTED COMPONENT)

Institution	Caltech	Georgetown	Oklahoma	Princeton	UNC
Nodes	1099	12195	24110	8555	24780
Connected Nodes	762	9388	17420	6575	18158
Connected Edges	16651	425619	892524	293307	766796
Mean Degree	43.7	90.7	102.5	89.2	84.5
# Communities	12	33	5	12	5
Modularity	0.4003	0.4801	0.3869	0.4527	0.4274

Visual Comparison: Caltech Houses



Princeton Class Year and Major



IS THIS RANDOM? IS IS CORRELATED? VISUALLY, IT'S NOT CLEAR!
WHAT QUANTITATIVE STATISTICAL TOOLS ARE AVAILABLE?

Quantitative Comparison

- Available methods: cluster matching, information theoretic methods, pair counting, generative models (e.g., ERGM)?
- Pair-counting indices: Rand, Jaccard, Minkowski, Fowlkes-Mallows, Gamma, Adjusted Rand, …
  - Simple to state, but have various problematic properties
  - We find a unified interpretation by recasting index values as z-scores relative to shuffled data (i.e., using permutation tests)

Pair-Counting Indices

- Related to other set distances, but applied to node pairs
- w_{11} = # node pairs put in the same group in 1st and also in the same group in 2nd partition
- w_{10} = # node pairs put in the same group in 1st partition but different groups in 2nd partition
- w_{01} and w_{00} defined analogously
- M = total node pairs = $\Sigma_{ij} w_{ij}$

$$S_{\mathrm{R}} = (w_{11} + w_{00})/M$$

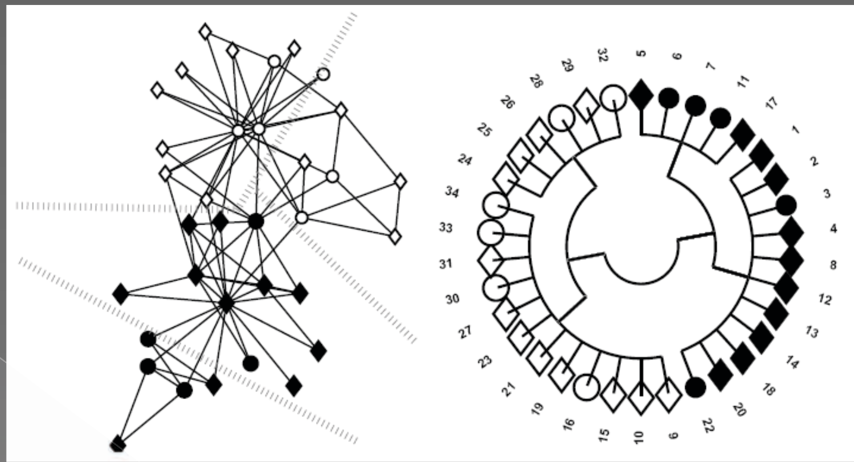
$$S_{\mathrm{J}} = w_{11}/(w_{11} + w_{10} + w_{01})$$

$$S_{\mathrm{FM}} = w_{11}/\sqrt{(w_{11} + w_{10})(w_{11} + w_{01})}$$

$$S_{\Gamma} = \frac{M w_{11} - (w_{11} + w_{10})(w_{11} + w_{01})}{\sqrt{(w_{11} + w_{10})(w_{11} + w_{01})(M - (w_{11} + w_{10}))(M - (w_{11} + w_{01}))}}$$

Similarity Values and Z-scores

1. Z-scores for Rand, Adjusted Rand, Fowlkes-Mallows, & Gamma indices are provably identical
2. Analytical formulas exist for the above indices (need permutation tests for Jaccard and Minkowski)



	S_{FM}	S_{Γ}	S_{J}	S_{M}	S_{R}	S_{AR}
"Observed"	0.7313	0.6092	0.5348	0.9327	0.7736	0.5414
"Random"	0.3867	0.0150	0.2204	1.4094	0.4831	0.0126

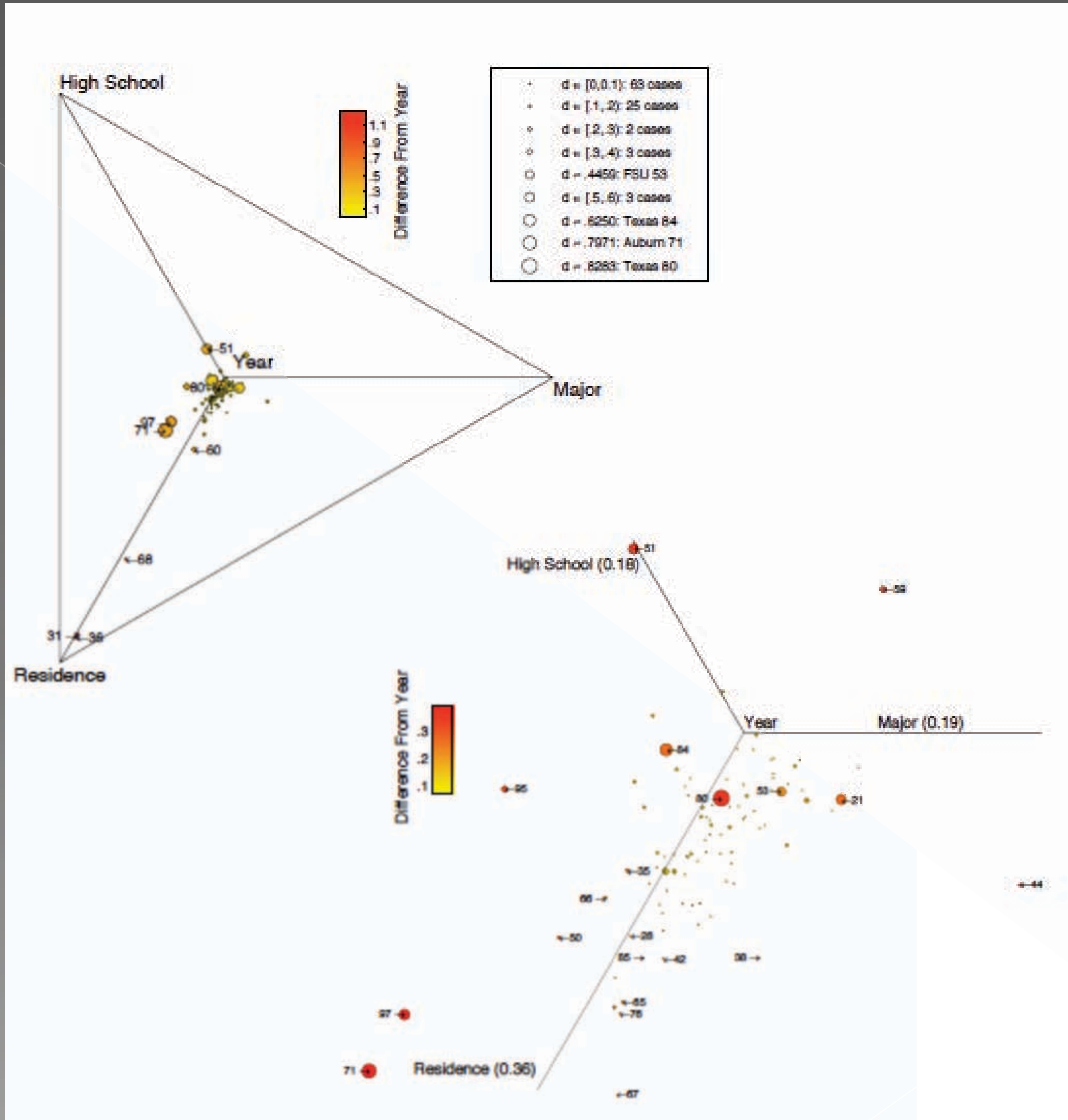
	z_{FM}	z_{Γ}	z_{J}	z_{M}	z_{R}	z_{AR}
"Observed"	14.6	14.6	18.0	17.1	14.6	14.6
"Random"	0.343	0.343	0.322	0.329	0.343	0.343

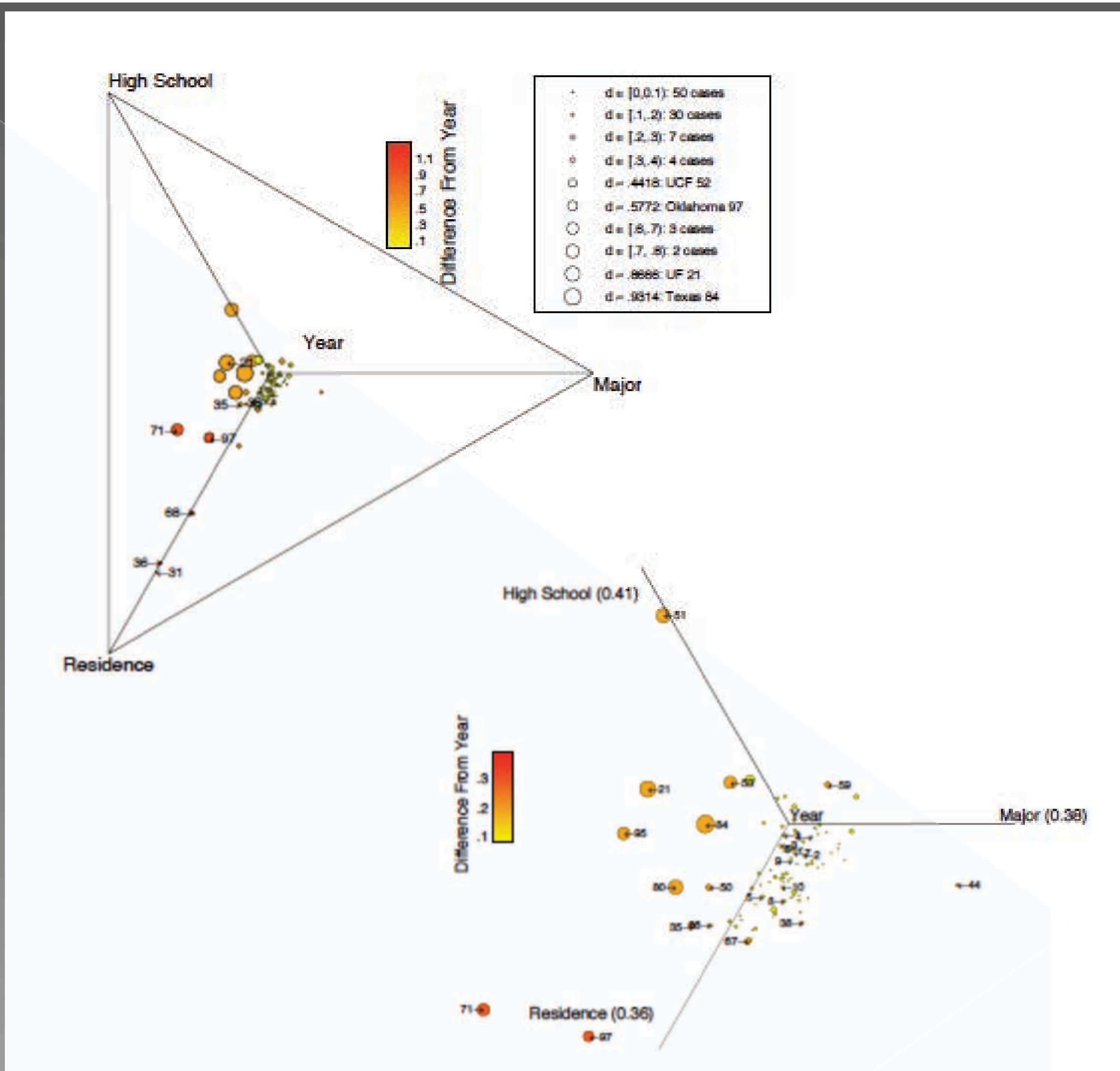
Z-Scores

		Caltech	Georgetown	Oklahoma	Princeton	UNC
Inclusion:	"Major"	3.962	5.885	3.799	15.03	8.044
	"Dorm/House"	200.8	148.8	71.00	58.26	113.0
	"Year"	6.727	1543	206.7	1058	778.2
	"High School"	-0.553	26.13	18.50	15.62	15.93
Pairwise:	"Major"	4.051	16.00	16.44	9.968	5.700
	"Dorm/House"	285.3	212.9	186.9	147.2	93.34
	"Year"	5.389	1837	286.1	1270	889.1
	"High School"	0.7695	4.247	22.54	2.888	37.22
Listwise:	"Major"	2.235	15.23	26.10	10.07	13.90
	"Dorm/House"	248.9	221.5	159.9	116.5	90.50
	"Year"	2.644	1913	251.2	997.3	475.7
	"High School"	0.3063	1.228	13.69	2.415	21.12

How do Universities Organize?

- Houses are important at Caltech (reality check for methodology)
- High school is more important at large state universities
- Class year is the most important factor at most universities and dorm is often a very strong secondary factor
- Major has varying importance at different universities


High School
Year
Major
Residence
Difference From Year
1.1
.9
.7
.5
.3
.1
d = [0,0.1): 63 cases
d = [.1,.2): 25 cases
d = [.2,.3): 2 cases
d = [.3,.4): 3 cases
d = .4459: FSU 53
d = [.5,.6): 3 cases
d = .6250: Texas 84
d = .7971: Auburn 71
d = .8283: Texas 80
High School (0.18)
Year
Major (0.19)
Residence (0.36)
Difference From Year
.3
.2
.1

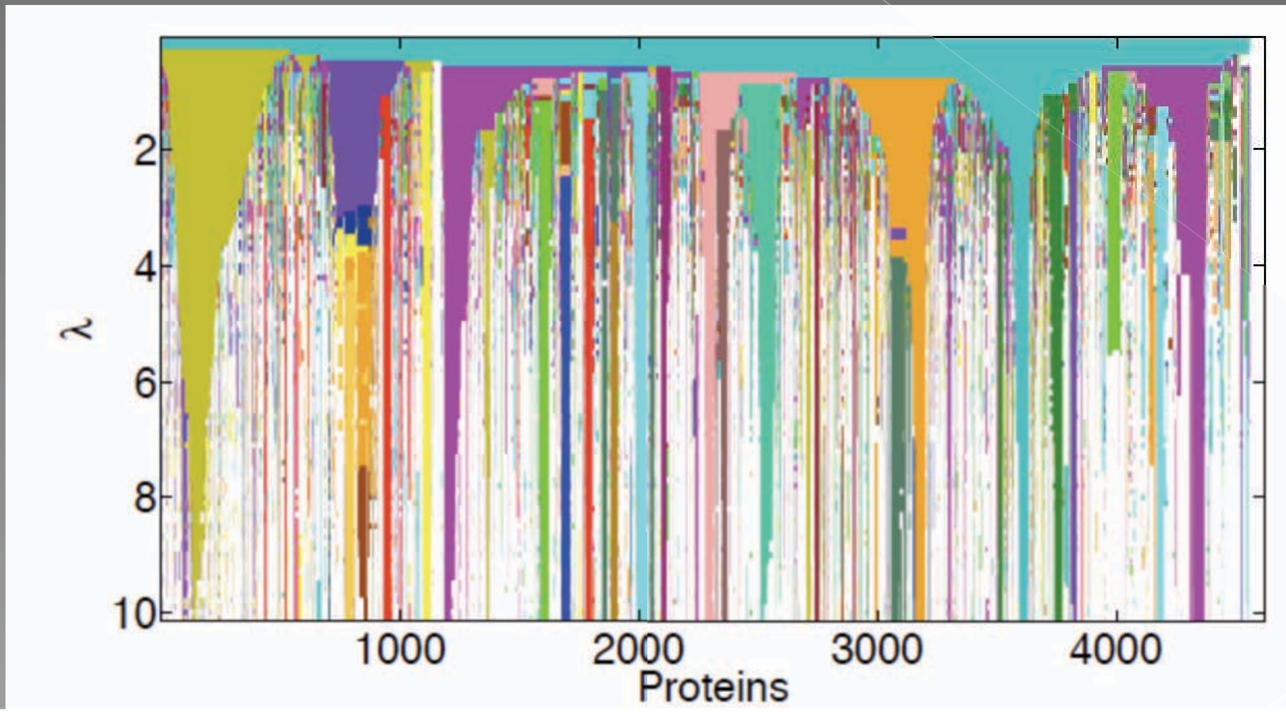

High School
Year
Major
Residence
Difference From Year
1.1
.9
.7
.5
.3
.1
d = [0,0.1): 50 cases
d = [.1,.2): 30 cases
d = [.2,.3): 7 cases
d = [.3,.4): 4 cases
d = .4418: UCF 52
d = .5772: Oklahoma 97
d = [.6,.7): 3 cases
d = [.7,.8): 2 cases
d = .8666: UF 21
d = .9314: Texas 84
High School (0.41)
Major (0.38)
Residence (0.36)
.3
.2
.1

Communities, Dynamics, and Function: A Whirlwind Tour

- Dynamics *on* networks
  - E.g., how does network structure affect dynamics, models of social influence, etc.
- Dynamics *of* networks
  - E.g., communities in evolving networks
    - Temporal dynamics
    - Dynamics with respect to parameters
- Developing some theory…
  - E.g., "multislice" networks, mesoscopic response functions, new methods to detect core-periphery structure, etc.

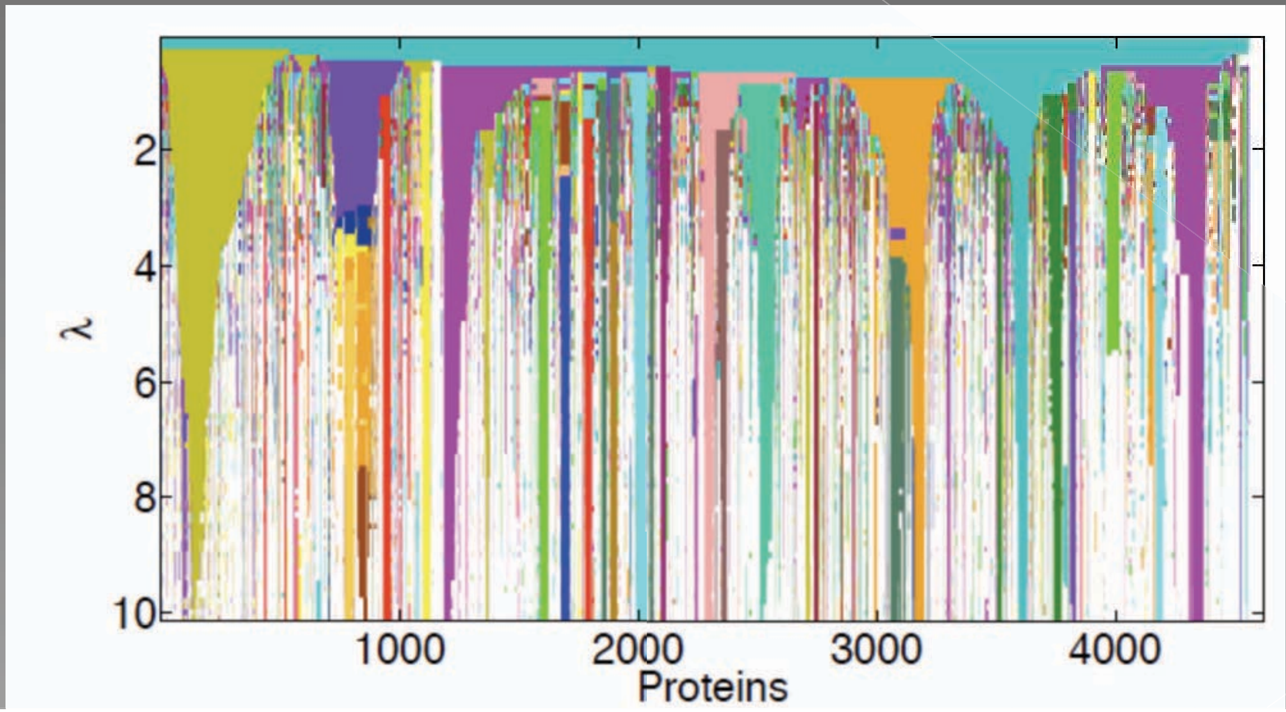
Parameter Dynamics of Communities

- A. C. F. Lewis, NSJ, MAP, & C. M. Deane, BMC Systems Biology **4**: 100 (2010)
- Protein-protein interaction networks
- Examine changes in communities with respect to resolution parameters
- Investigate biological properties of "persistent" communities
- Can network properties pick out functionally homogeneous communities?
  - Clustering coefficient does well (best among 49 tested properties), which is very nice given incompleteness of Gene Ontology (GO) annotations



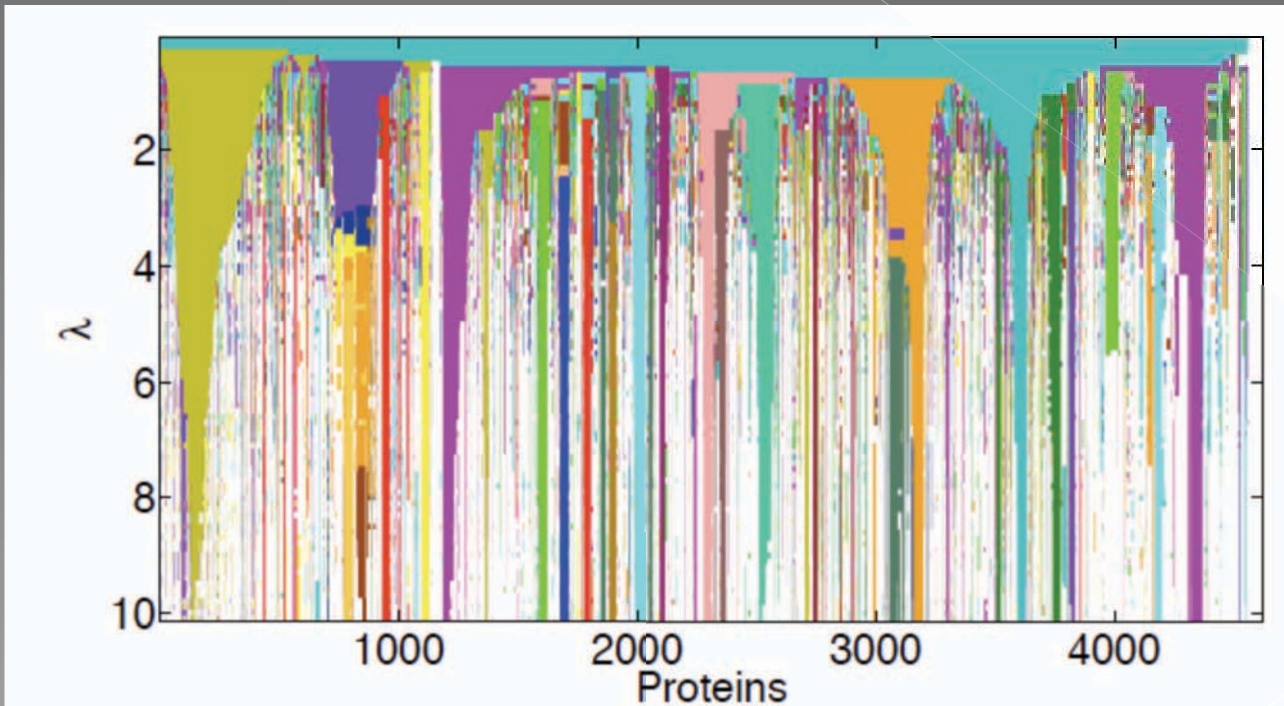
Parameter Dynamics of Communities

- A. C. F. Lewis, NSJ, MAP, & C. M. Deane, BMC Systems Biology **4**: 100 (2010)
- Protein-protein interaction networks
- Examine changes in communities with respect to resolution parameters
- Investigate biological properties of "persistent" communities
- Can network properties pick out functionally homogeneous communities?
  - Clustering coefficient does well (best among 49 tested properties), which is very nice given incompleteness of Gene Ontology (GO) annotations



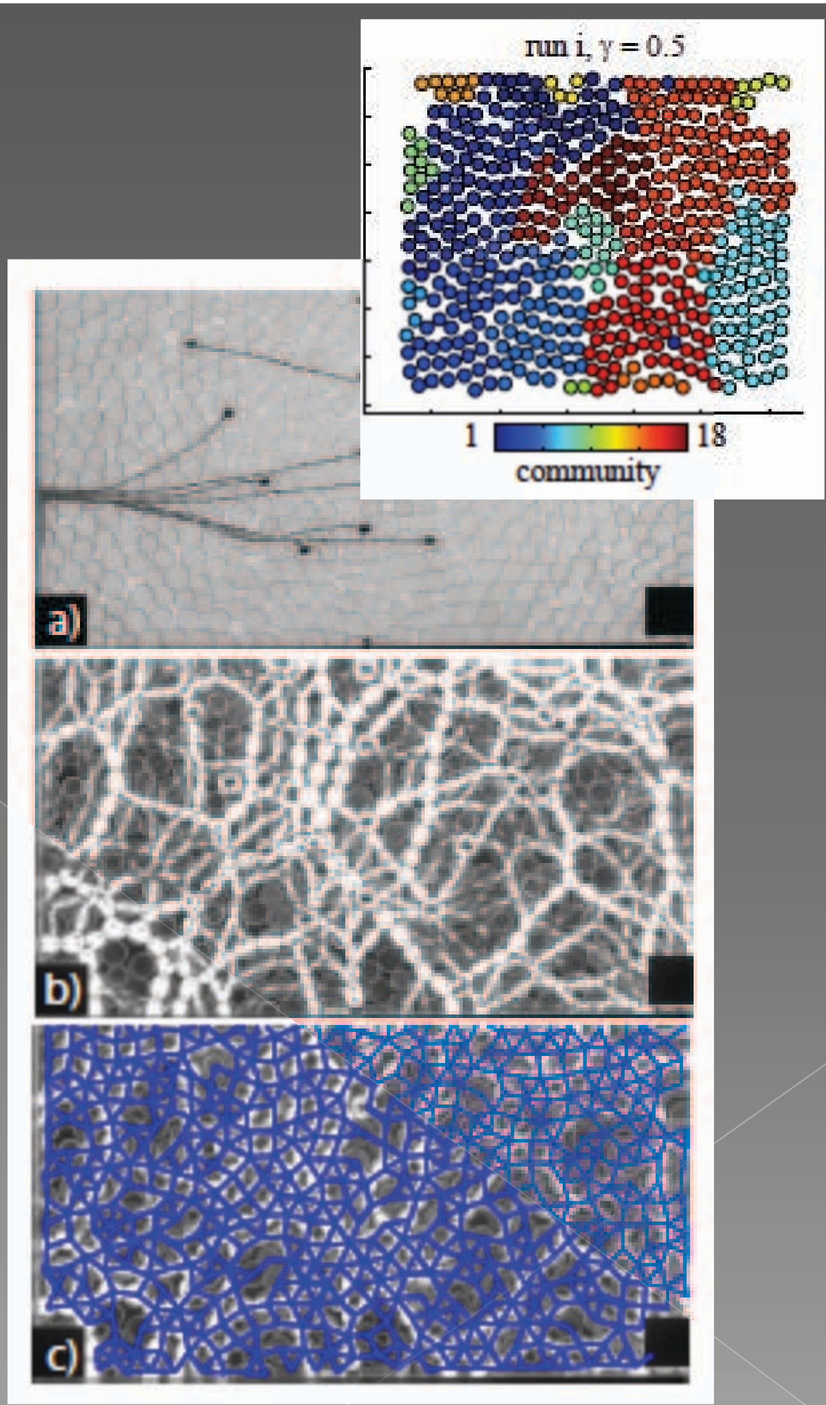
Parameter Dynamics of Communities

- A. C. F. Lewis, NSJ, MAP, & C. M. Deane, BMC Systems Biology **4**: 100 (2010)
- Protein-protein interaction networks
- Examine changes in communities with respect to resolution parameters
- Investigate biological properties of "persistent" communities
- Can network properties pick out functionally homogeneous communities?
  - Clustering coefficient does well (best among 49 tested properties), which is very nice given incompleteness of Gene Ontology (GO) annotations



Sound Propagation in Granular Force Networks

- D. S. Bassett, E. T. Owens, K. E. Daniels, & MAP, arXiv: 1110.1858
- 2D granular medium of photoelastic disks
- Two networks
  - Underlying topology (unweighted)
  - Forces (weighted)
- Meso-scale structures (communities) of both types of networks are crucial for characterizing sound propagation, illustrating that contact topology alone is insufficient



Core-Periphery Structure

- M. P. Rombach, MAP, J. H. Fowler, & PJM, arXiv:0212.2684 (2012)
- Core-periphery structure is a different type of mesoscopic structure from community structure.

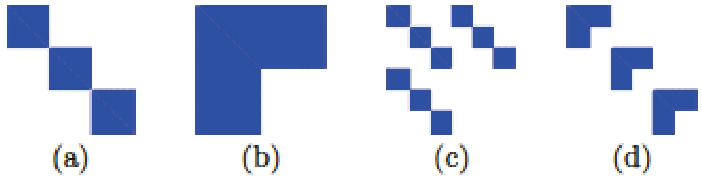
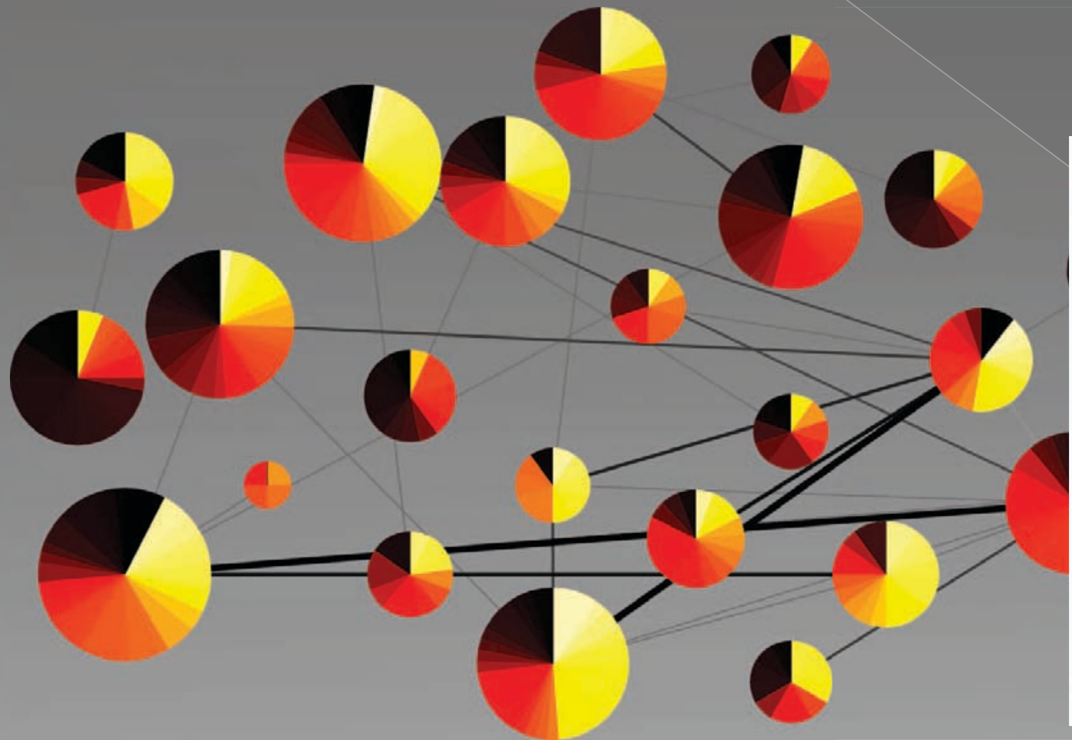


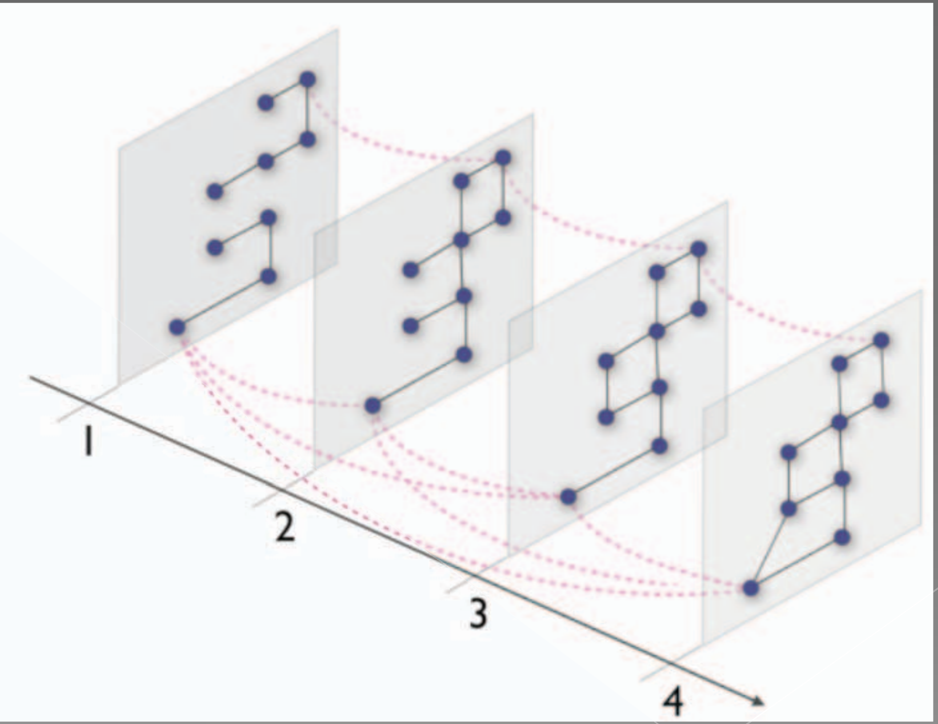
Fig. 2. Different network models. (a) community structure, (b) core-periphery structure, (c) global core-periphery / local community structure and (d) global community / local core-periphery structure.



Node	Core Score
King's Cross St. Pancras	1.0000
Baker Street	0.8339
Waterloo	0.8175
Willesden Junction	0.7973
Bank	0.7789
West Ham	0.7516
Green Park	0.7447
Oxford Circus	0.7200
Liverpool Street	0.7167
Paddington	0.6799

"Multislice" Community Detection

- PJM, T. Richardson, Kevin Macon, MAP, & JPO, Science **328**(5980): 876-878 (2010)
- Detect communities in networks in a general setting that incorporates time-dependence, parameter-dependence, and multiplexity
  - Normal connections in a single slice + connections between node j and itself in different slices
  - Slice = different resolution, different time, different type of link, etc.

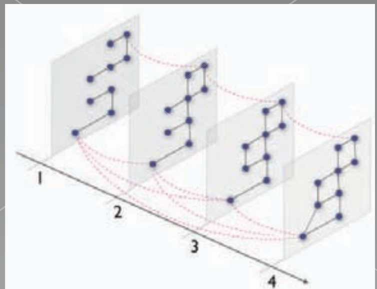


Random Walk on Multislice Network

- Undirected network slices: $A_{ijs} = A_{jis}$
- Undirected couplings: $C_{jrs} = C_{jsr}$
- Multislice strength: $\kappa_{js} = k_{js} + c_{js}$
- Density of random walkers in a node-slice:

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

- Steady-state probability distribution:
  - $p_{jr}^* = \kappa_{jr}/(2\mu)$
  - $2\mu = \Sigma_{jr}(\kappa_{jr})$



Obtain a Quality Function to Optimize

- Specify null model: Probability of sampling node-slice *is* conditional on whether the multislice structure allows one to step from node-slice *jr* to node-slice *is*:

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

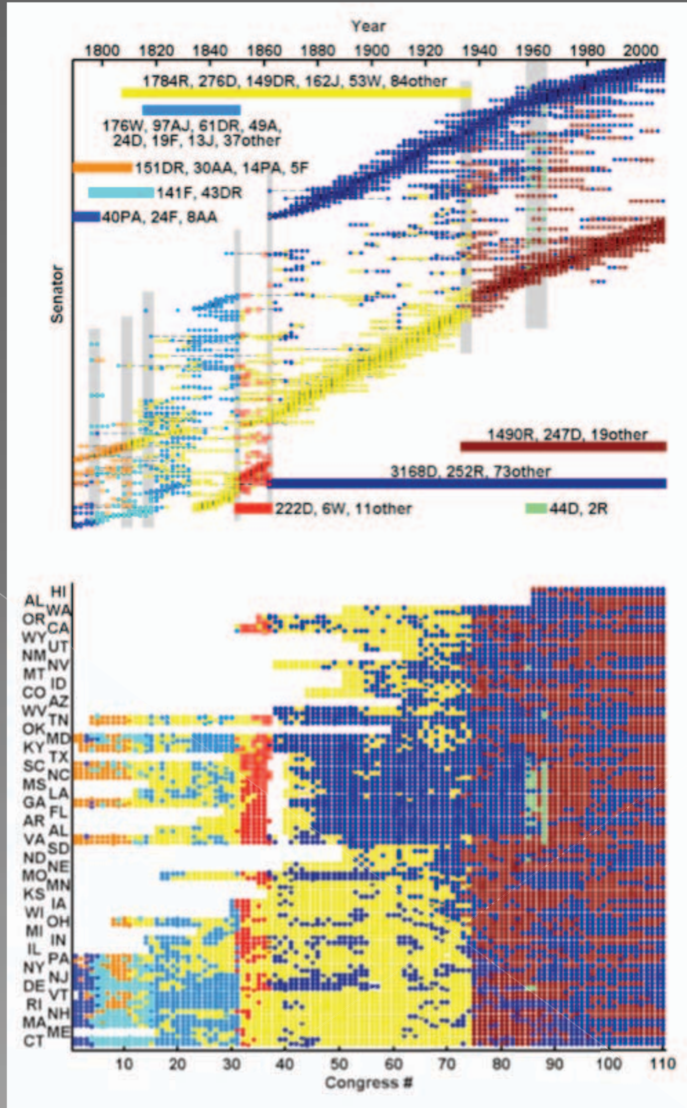
- Multislice modularity:

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

- Each slice has own resolution parameter γ_s

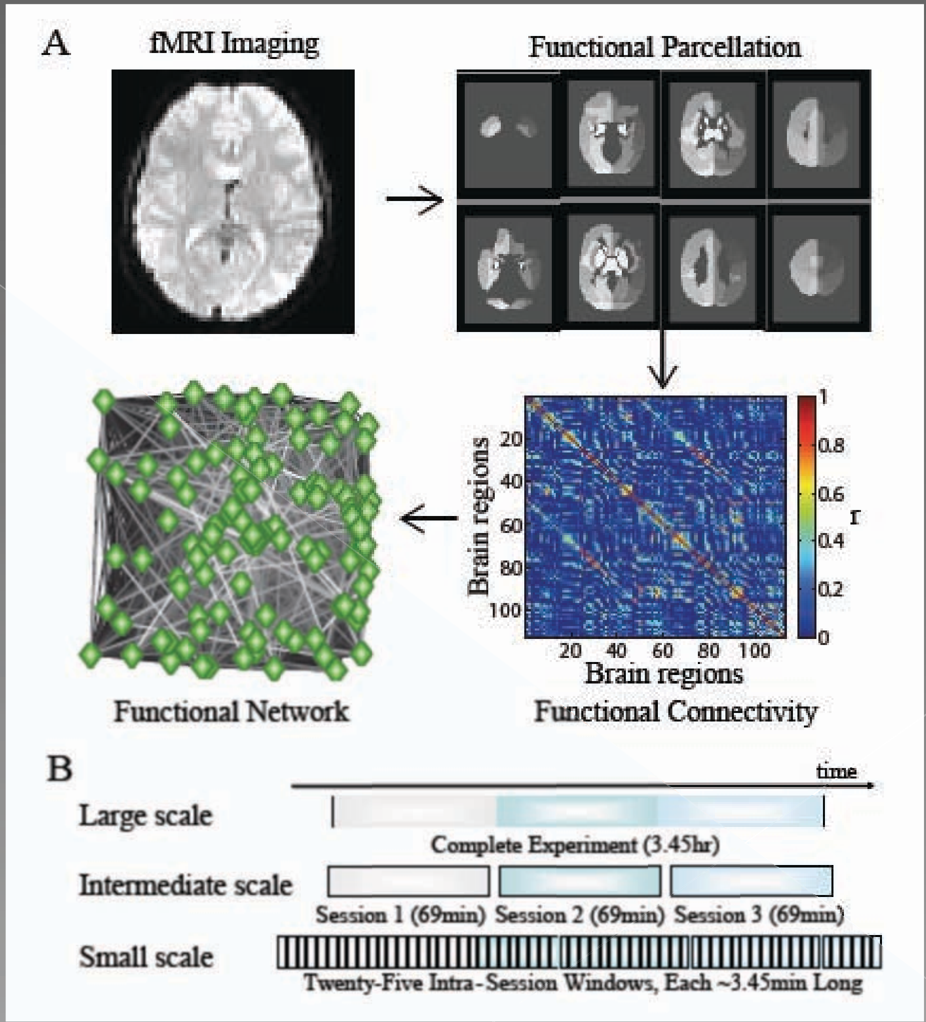
Example 1: U.S. Senate Voting

- Time-dependent network with over 200 years of roll call votes (1789-2008)
  - Weighted intra-slice edges based on voting similarity (computed separately for each slice)
  - Interslice edges for senators in consecutive 2-year congresses
- Colors = communities



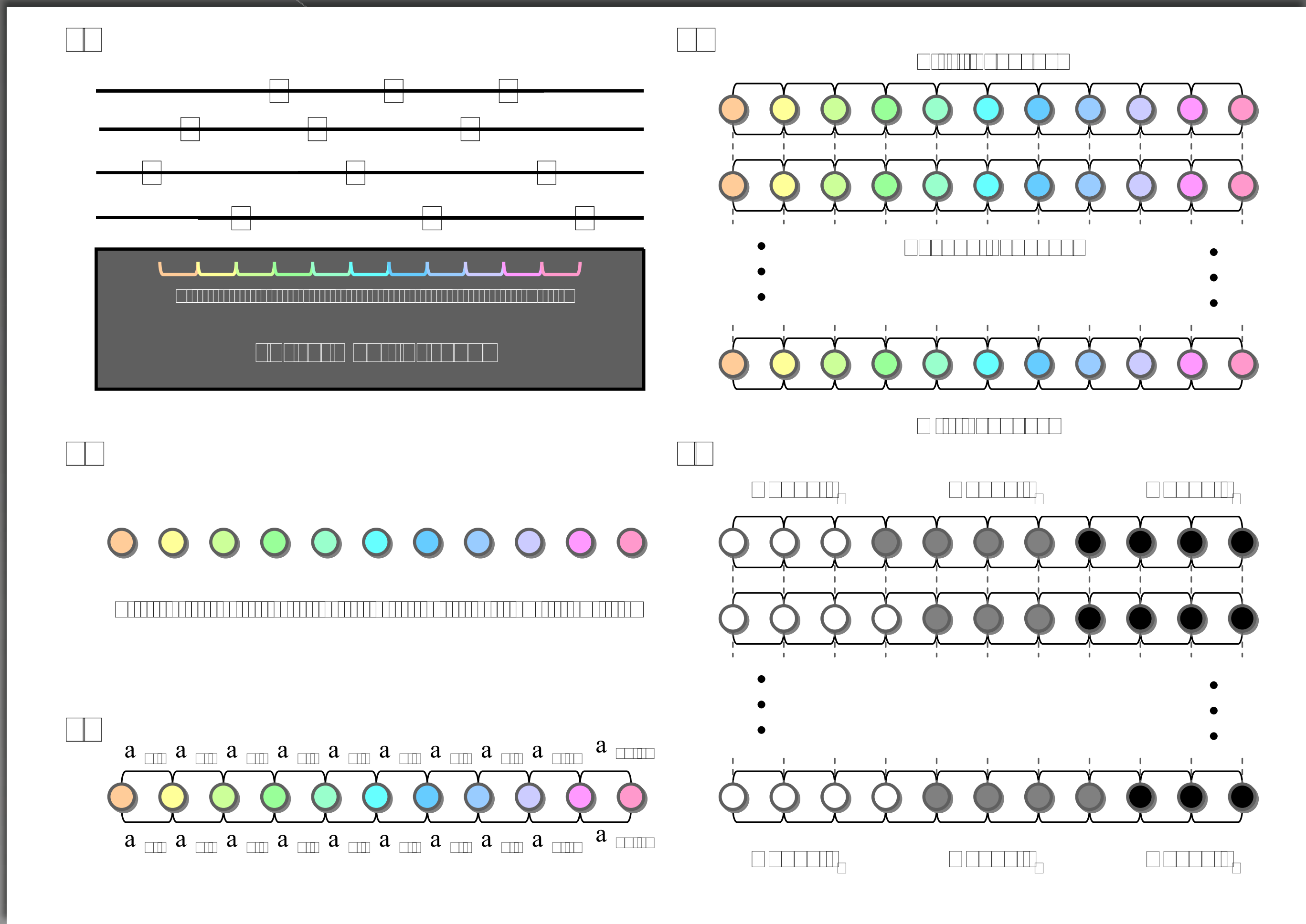
Example 2: Dynamic Reconfiguration of Human Brain Networks During Learning

- DSB, N. F. Wymbs, MAP, PJM, J. M. Carlson, & S. T. Grafton, PNAS **108**(18): 7641—7646 (2011)
- fMRI data: network from correlated time series
- Examine role of modularity in human learning by identifying dynamic changes in modular organization over multiple time scales
- Main result: Flexibility, as measured by allegiance of nodes to communities, in one session predicts amount of learning in future sessions



Example 3: Chunking

NFW, DSB, PJM, MAP, & STG, “Differential Recruitment of the Sensorimotor Putamen and Frontoparietal Cortex During Motor Chunking in Human”, to appear in Neuron (2012)



Conclusions

- Conclusions in limerick form*:

When detecting a network's communities,
try not to do it with impunity.
For it is not enough
to stop with that stuff.
Be sure to think about functionality.

- In other words…
- Most research on community structure:
  - Finds communities, possibly presents a new method, and stops.
- Another important consideration: Validating and/or studying the properties of communities once we have them
- * The audience at University of Limerick was far less amused by this than I thought they'd be.

Five Postdoc Positions to study "Multiplex Networks"

- My collaborators and I will soon be advertising five 3-year postdoc positions to study theory and applications of multiplex networks.
  - 1 postdoc at each of 5 universities
  - Investigators: Alex Arenas, Marc Barthelemy, James Gleeson, Yamir Moreno, Mason Porter