Lecture 7:
Generalized large-scale structure

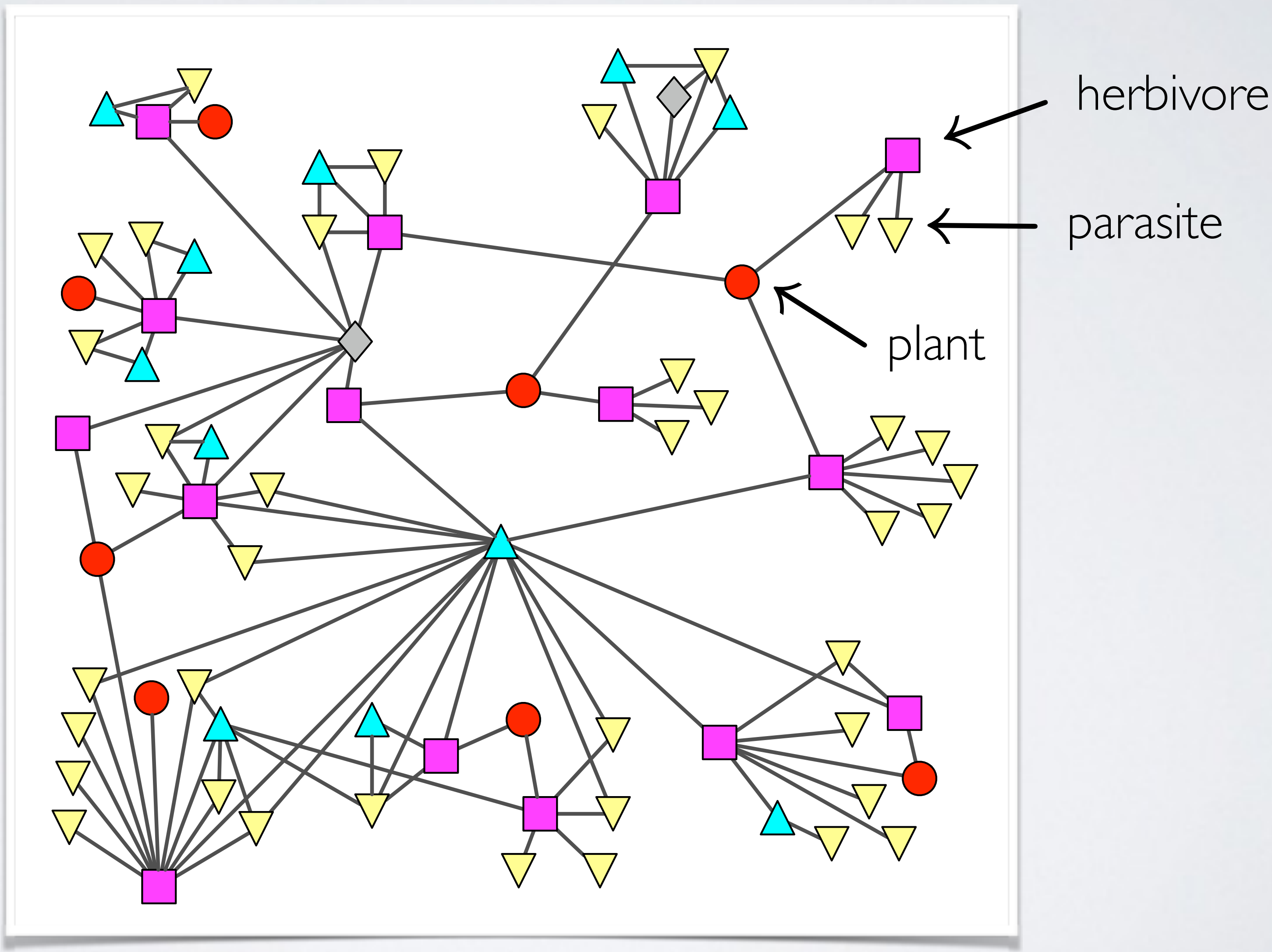
Aaron Clauset
@aaronclauset
Assistant Professor of Computer Science
University of Colorado Boulder
External Faculty, Santa Fe Institute

hierarchical communities

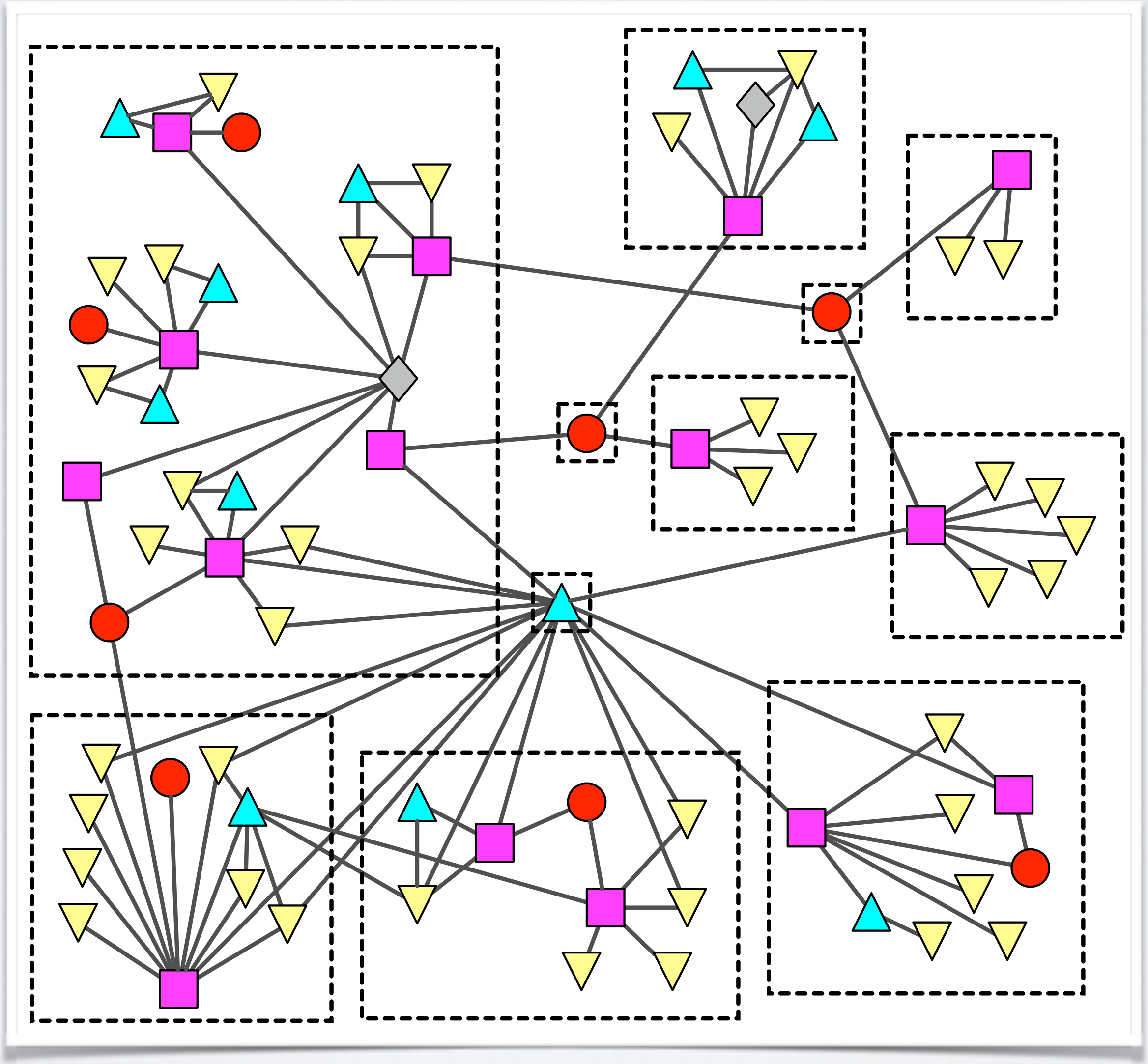
most communities are not random graphs

- groups within groups / groups of groups
- finding communities at one "level" of a hierarchy can obscure structure *above* or *below* that level

hierarchical communities

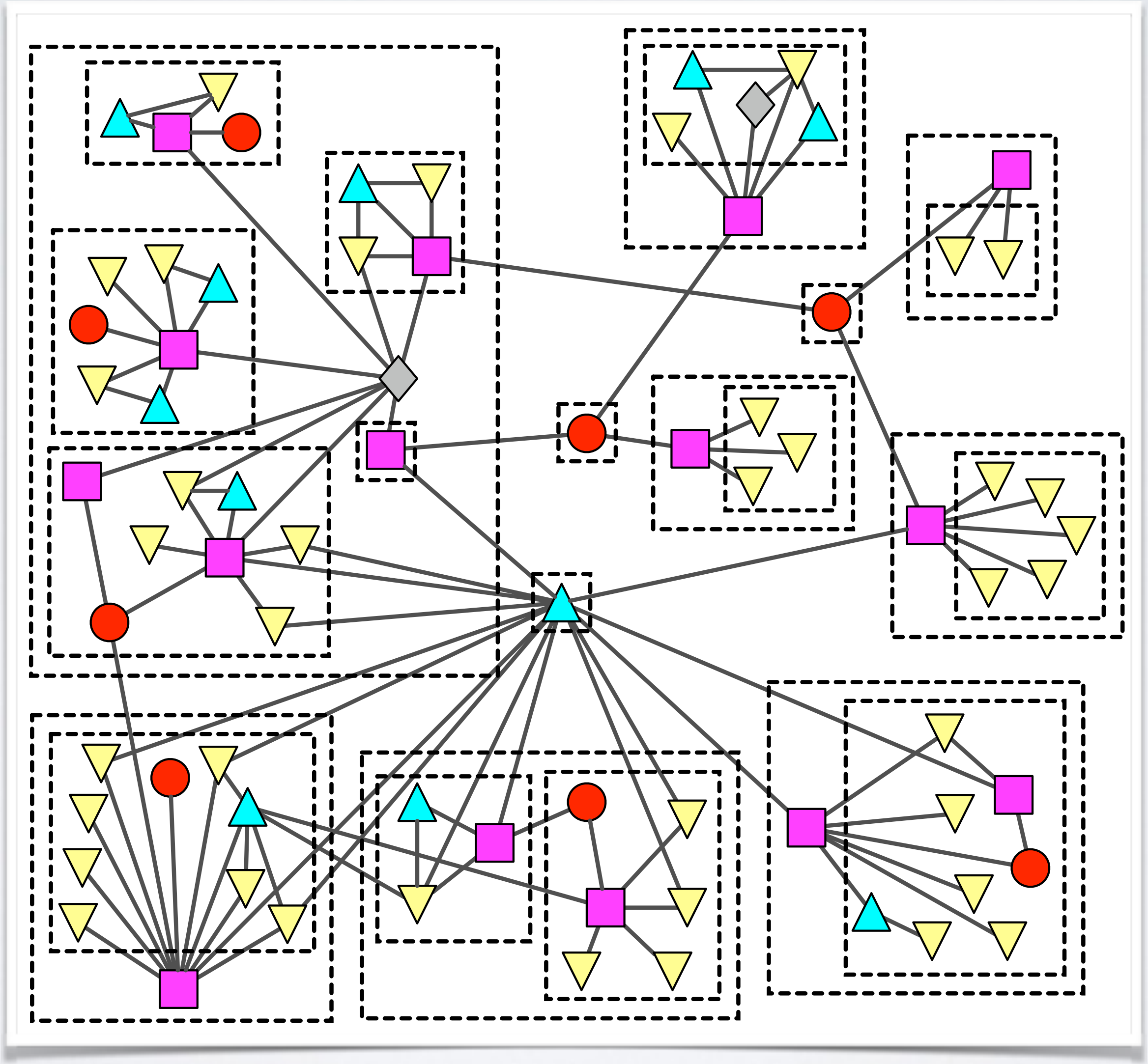


hierarchical communities



modules

hierarchical communities

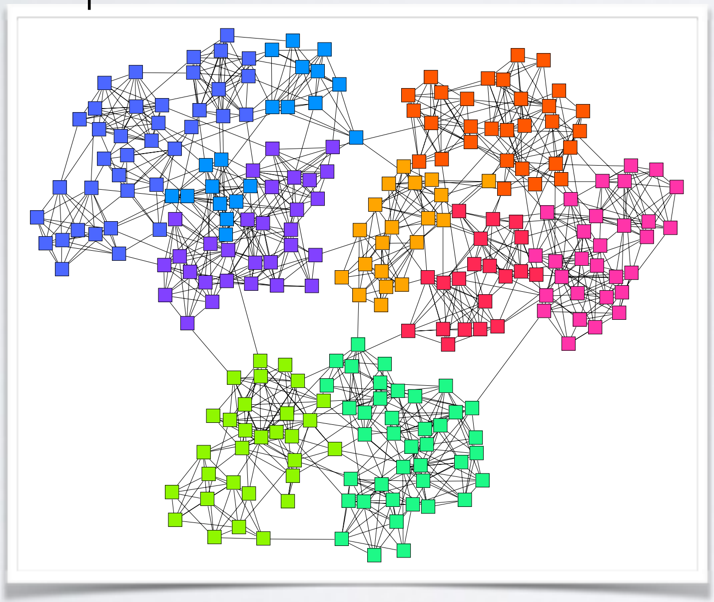


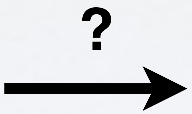
nested
modules

hierarchical communities

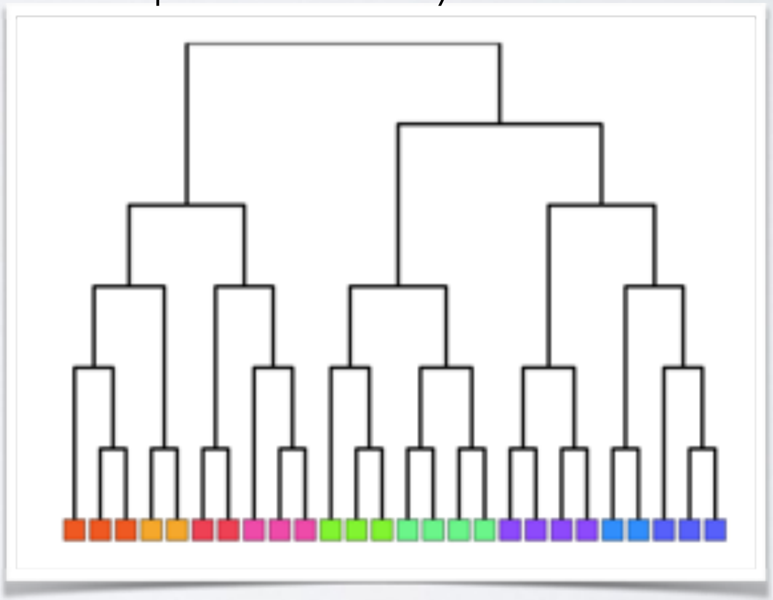
can we automatically extract such hierarchies?

step 1: network data





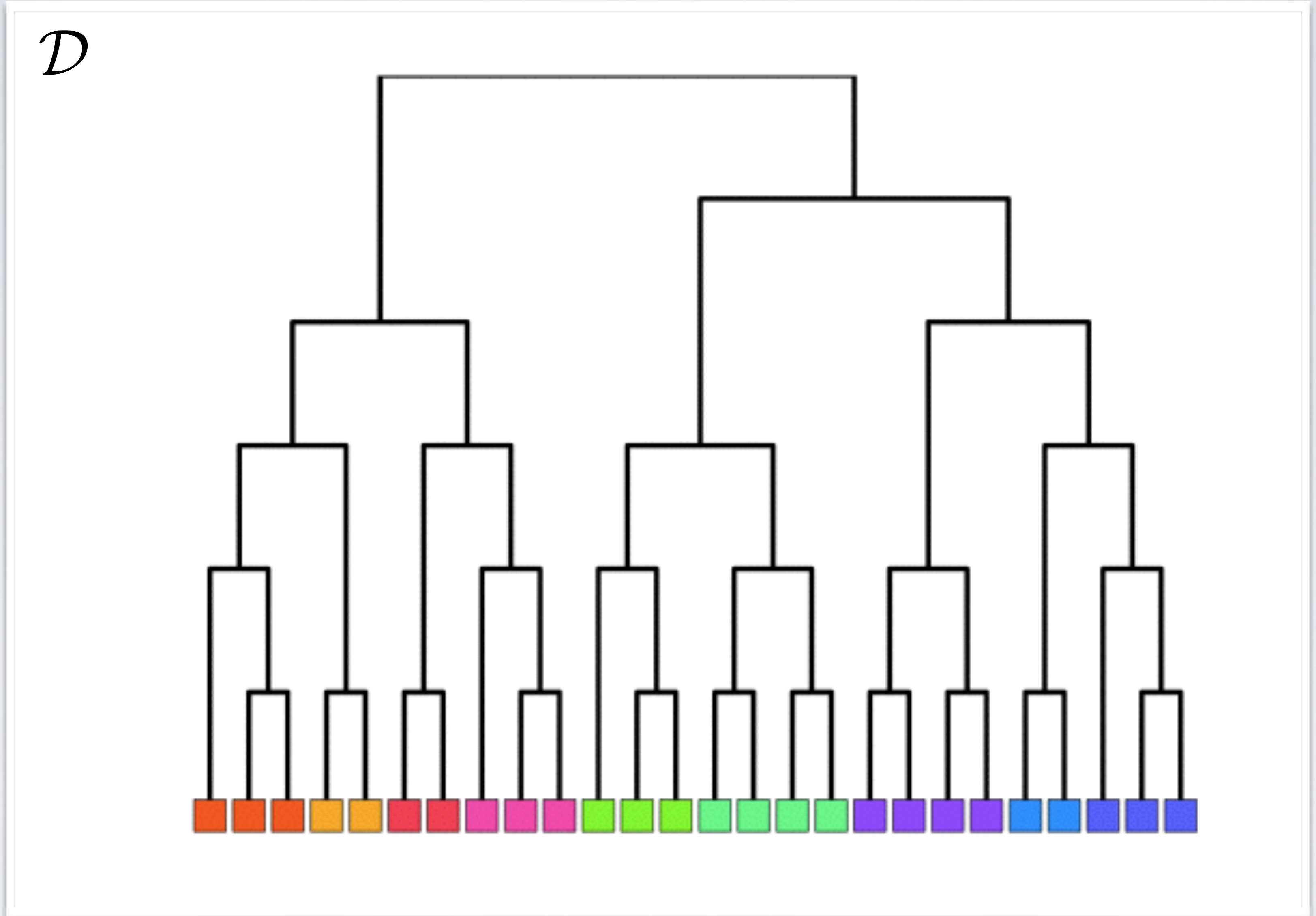
step 3: hierarchy



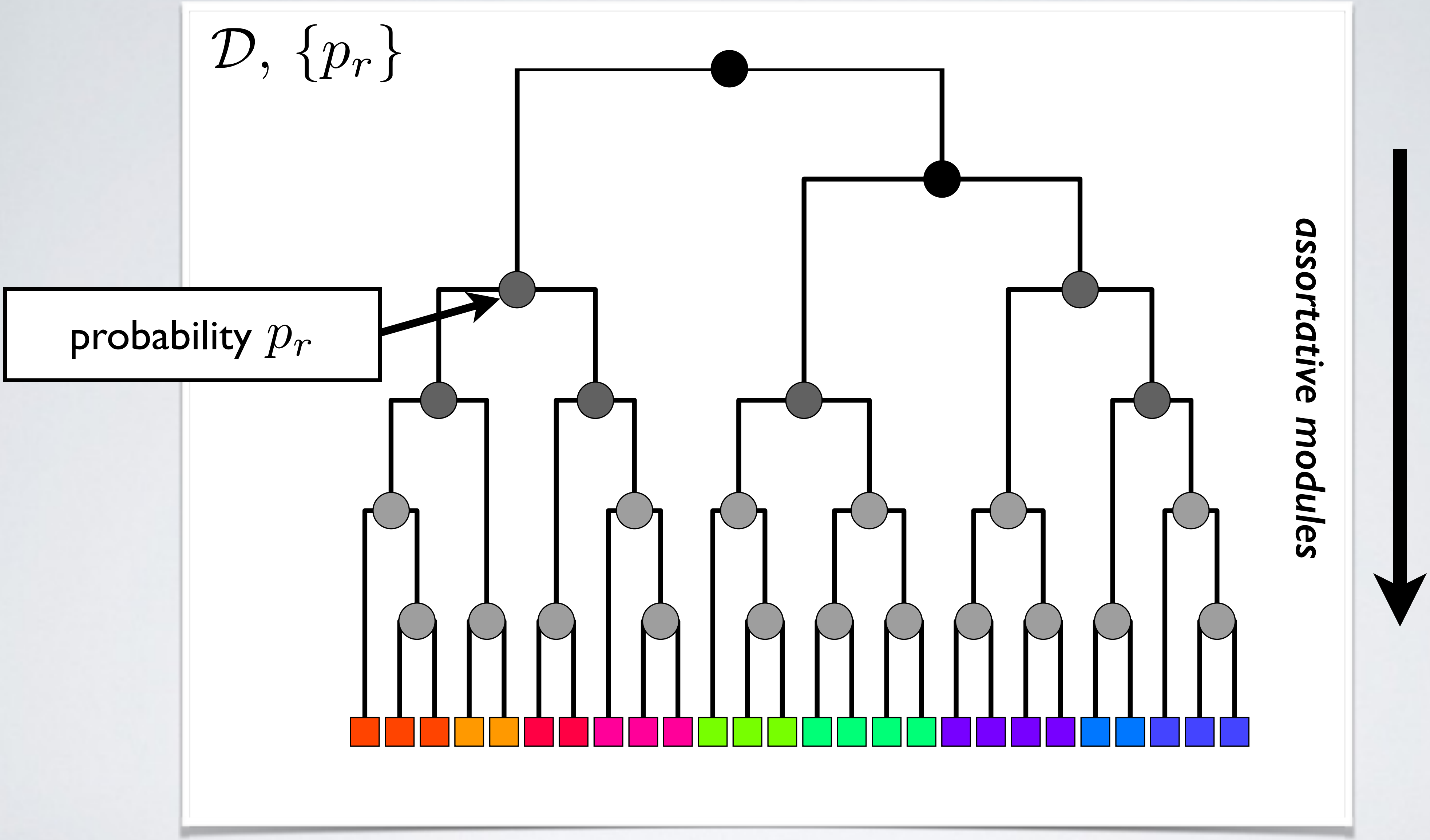
hierarchical communities

hierarchical random graph model

Clauset, Moore, Newman, *Nature* **453**, 98-101 (2008)
Clauset, Moore, Newman, *ICML* (2006)

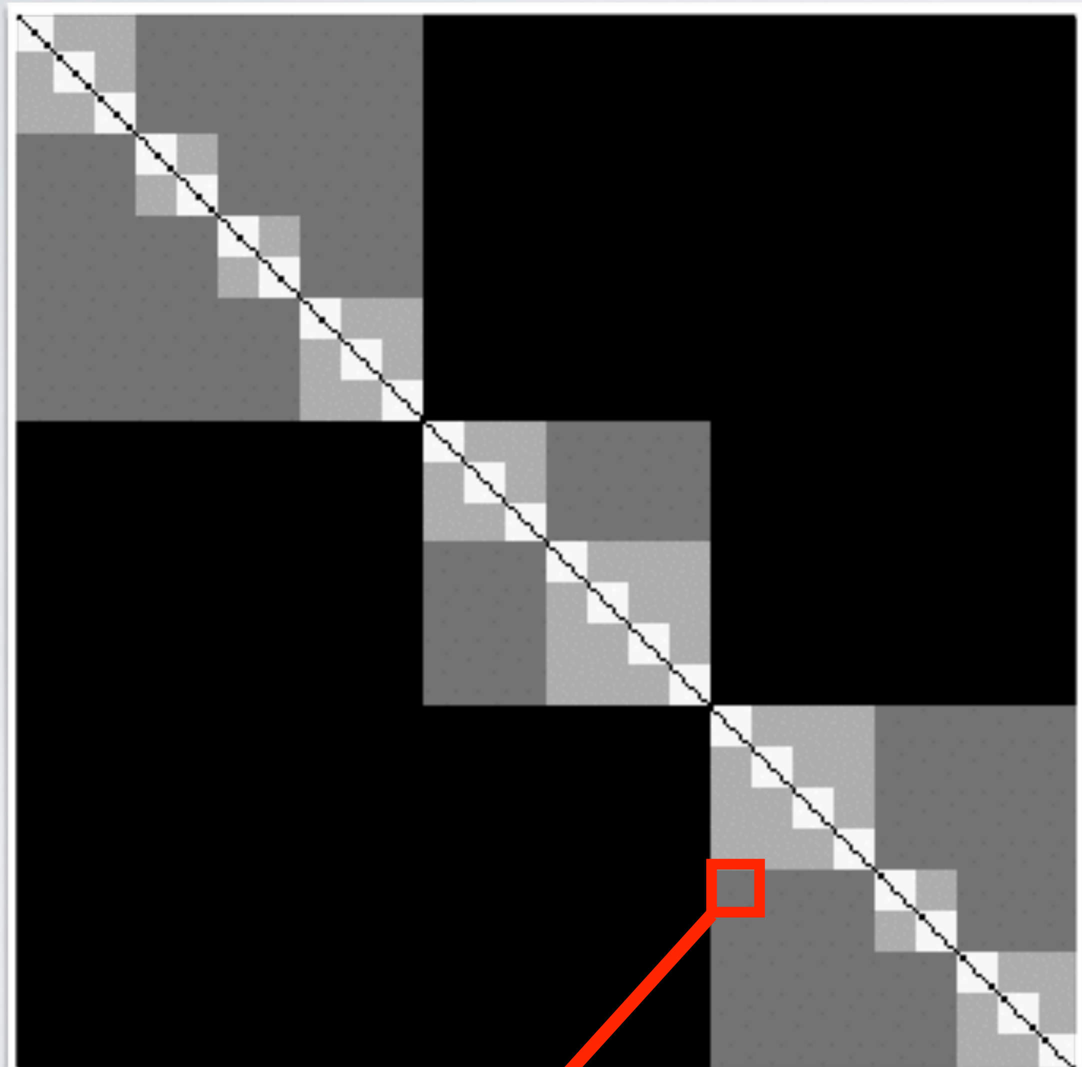


Clauset, Moore, Newman, *Nature* **453**, 98-101 (2008)
Clauset, Moore, Newman, *ICML* (2006)



Clauset, Moore, Newman, *Nature* **453**, 98-101 (2008)
Clauset, Moore, Newman, *ICML* (2006)

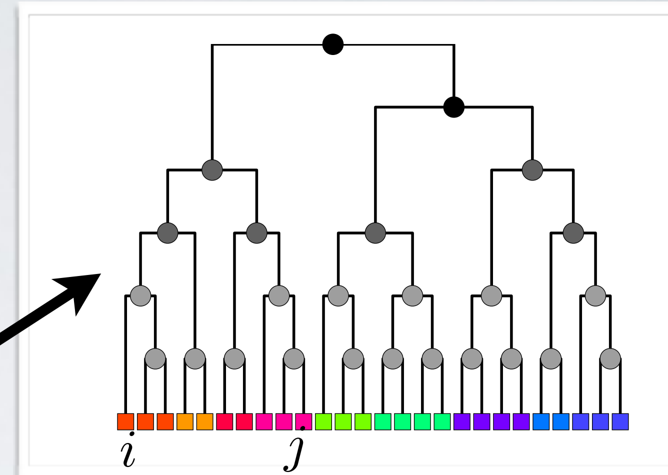
“inhomogeneous” random graph



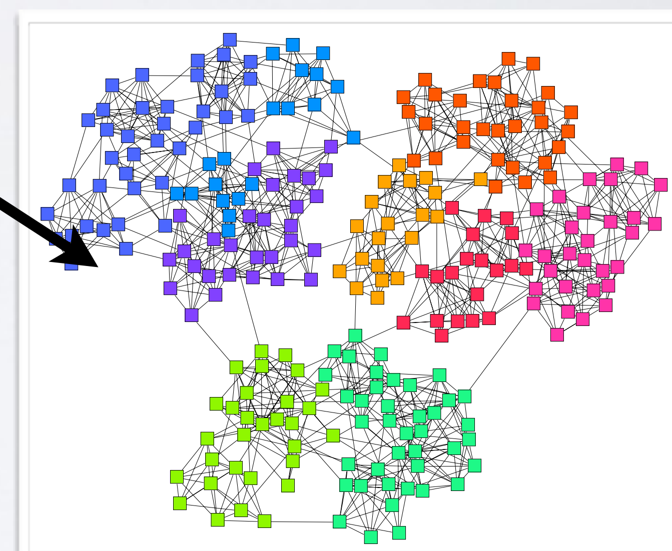
i

j

model



instance



$$\begin{aligned}\Pr(i, j \text{ connected}) &= p_r \\ &= p_{(\text{lowest common ancestor of } i,j)}\end{aligned}$$

hierarchical communities

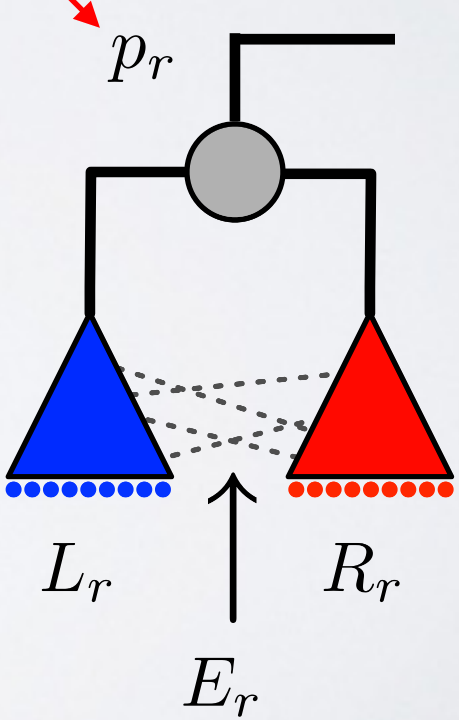
hierarchical random graph model

$$\Pr(A \mid \mathcal{D}, \{p_r\}) = \prod_r \underbrace{p_r^{E_r} (1 - p_r)^{L_r R_r - E_r}}$$

L_r = number nodes in left subtree

R_r = number nodes in right subtree

E_r = number edges with r as lowest common ancestor



Clauset, Moore, Newman, *Nature* **453**, 98-101 (2008)
Clauset, Moore, Newman, *ICML* (2006)

hierarchical communities

hierarchical communities

generalizing from a single example

- given graph A, estimate model parameters $\mathcal{D}, \{p_r\}$
- sample new graphs from posterior distribution $\Pr(G \mid \mathcal{D}, \{p_r\})$

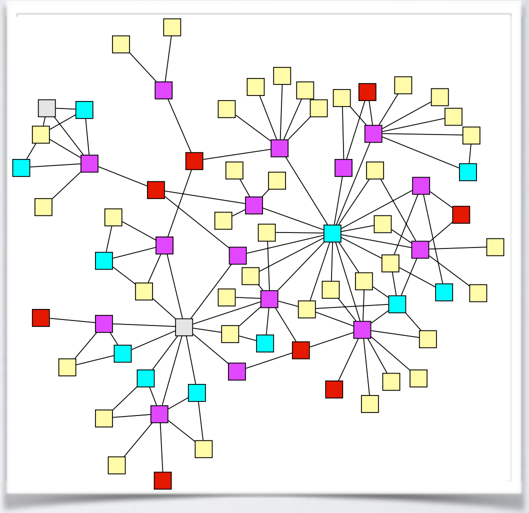
checking the models

compare resampled graphs with original data

check

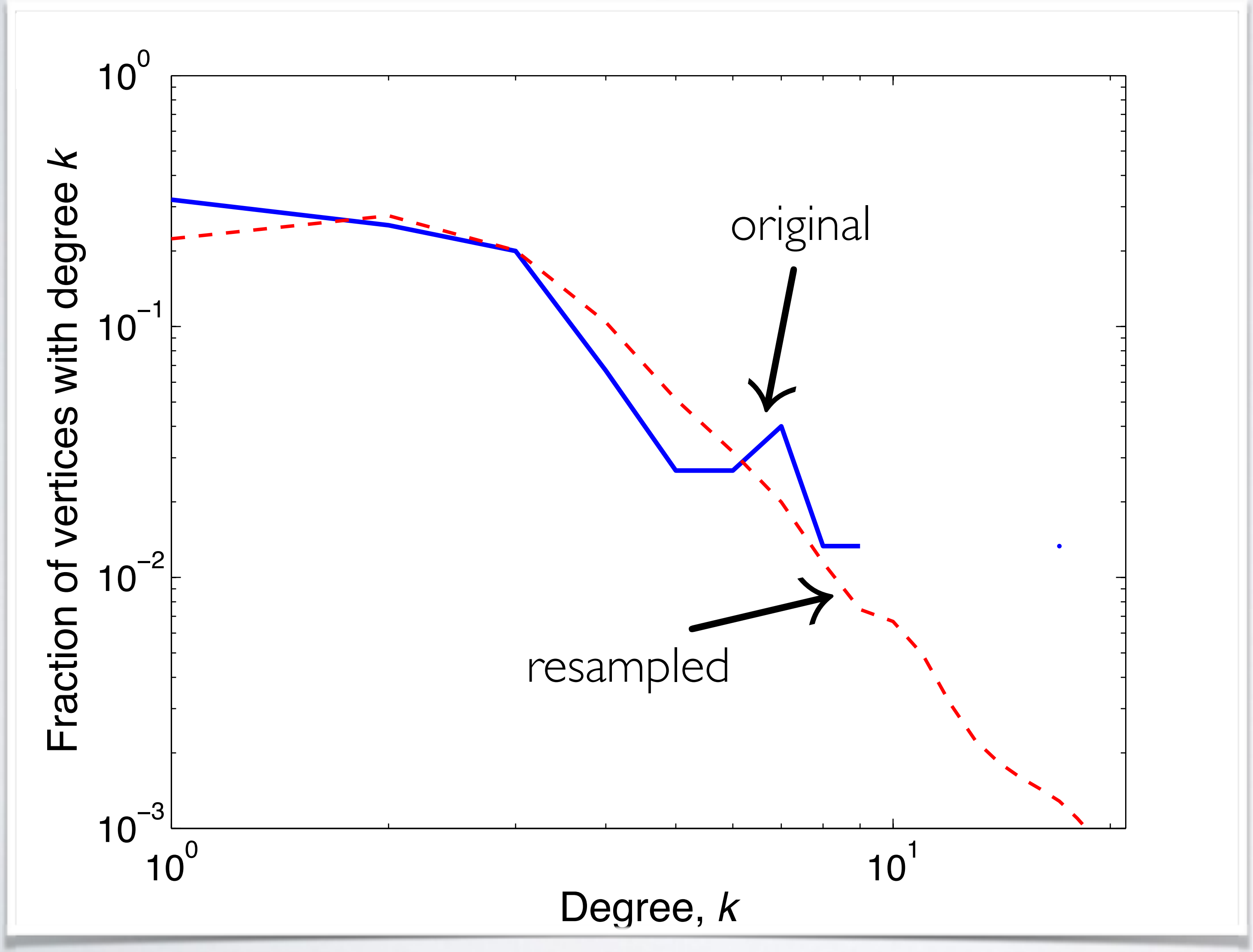
1. degree distribution
2. clustering coefficient
3. geodesic path lengths

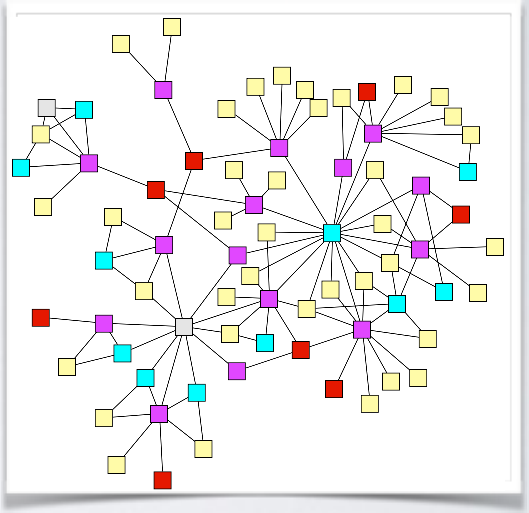
hierarchical communities



hierarchical communities

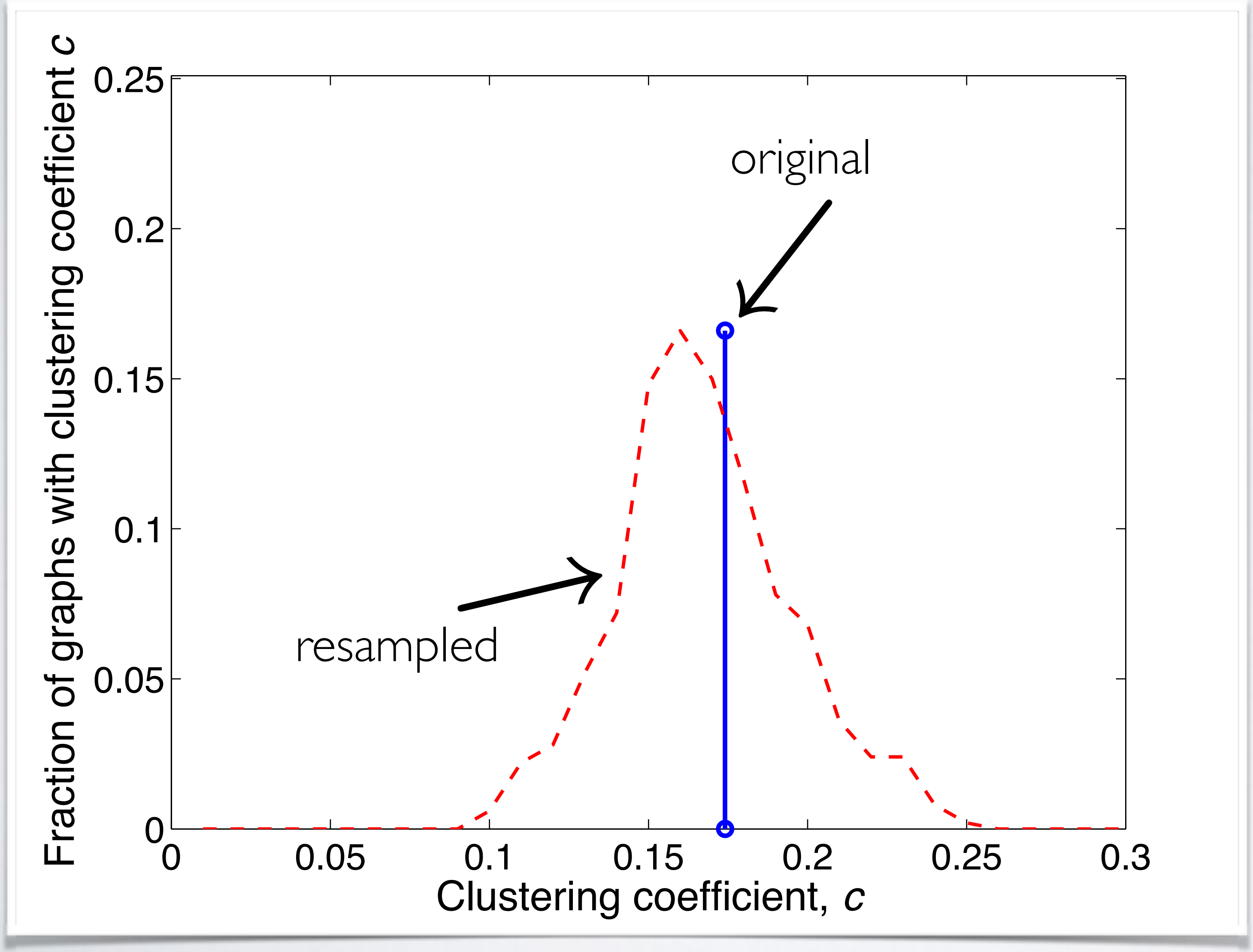
degree distribution

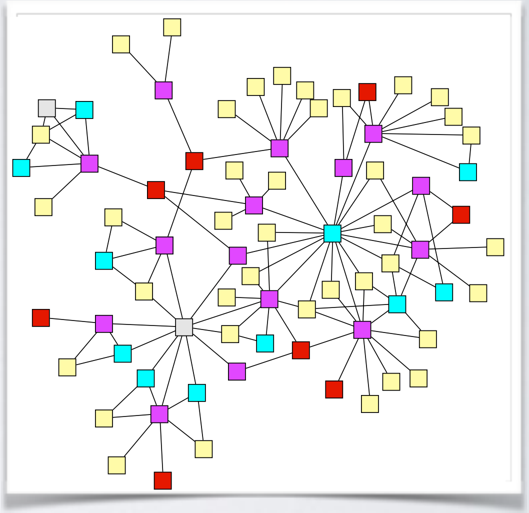




hierarchical communities

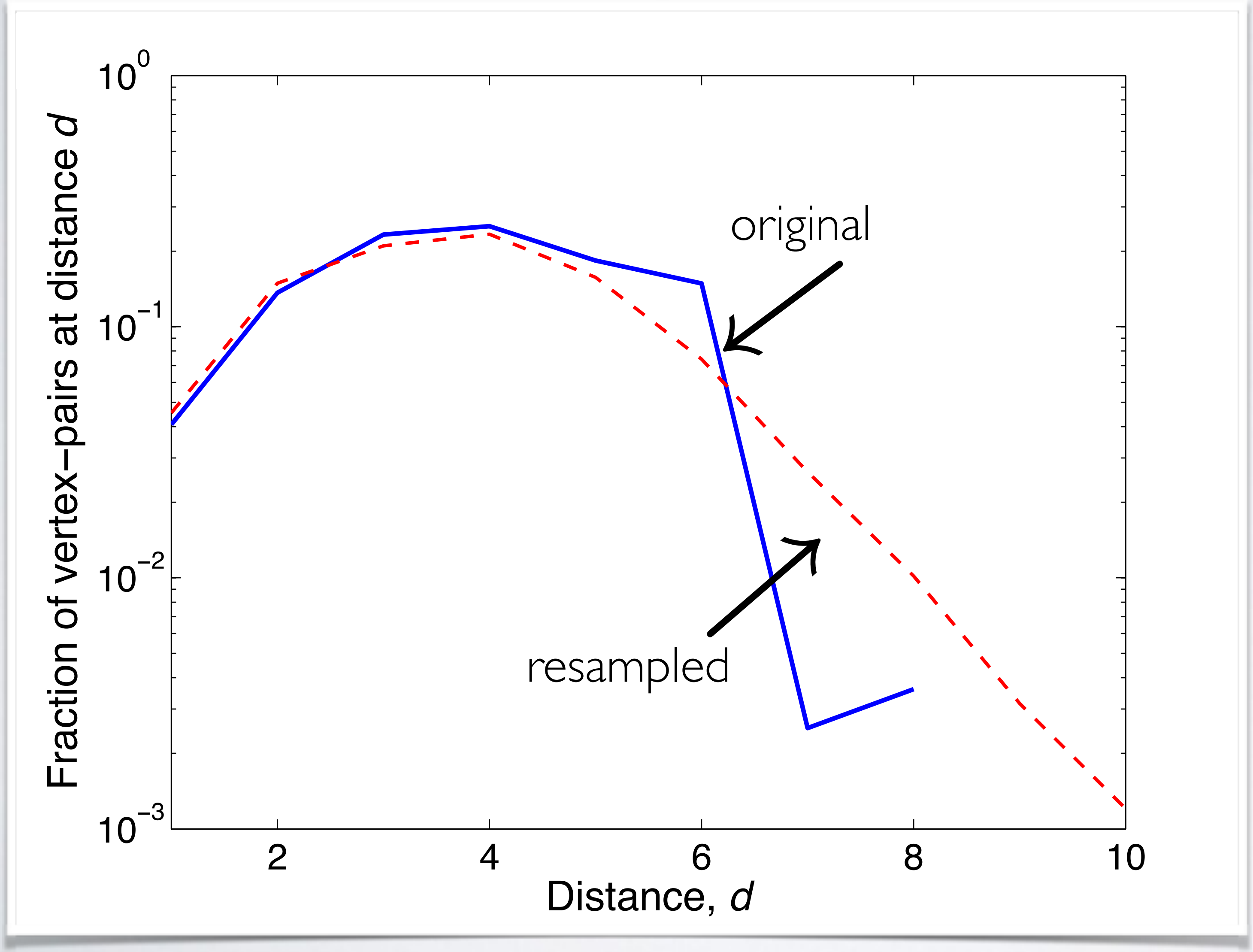
density of triangles

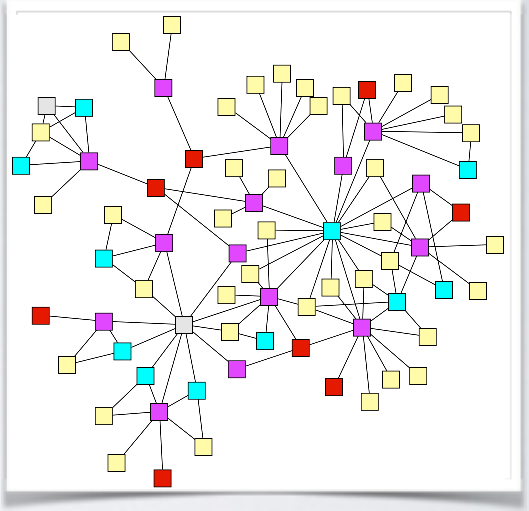




hierarchical communities

geodesic distances

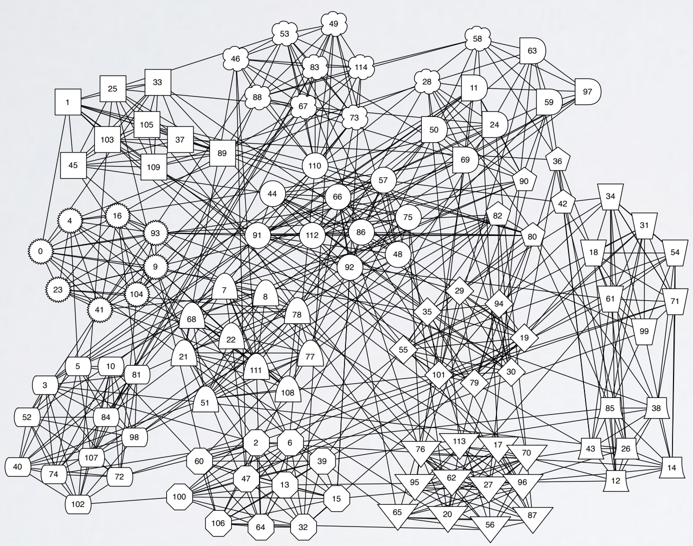




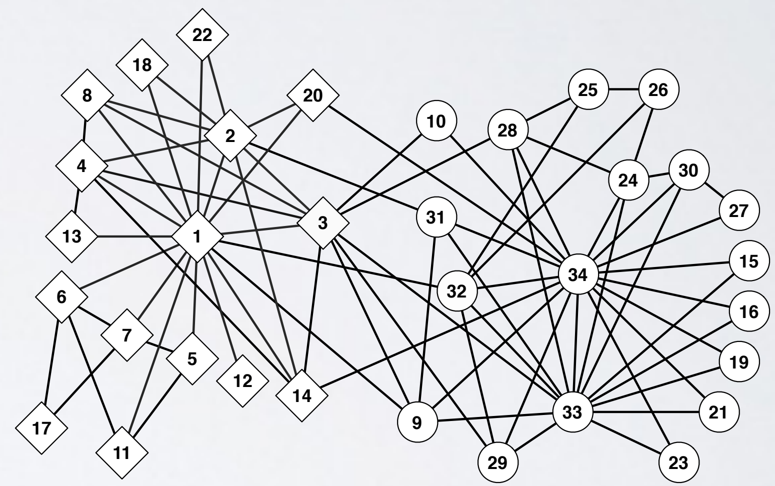
hierarchical communities

inspecting the dendrograms

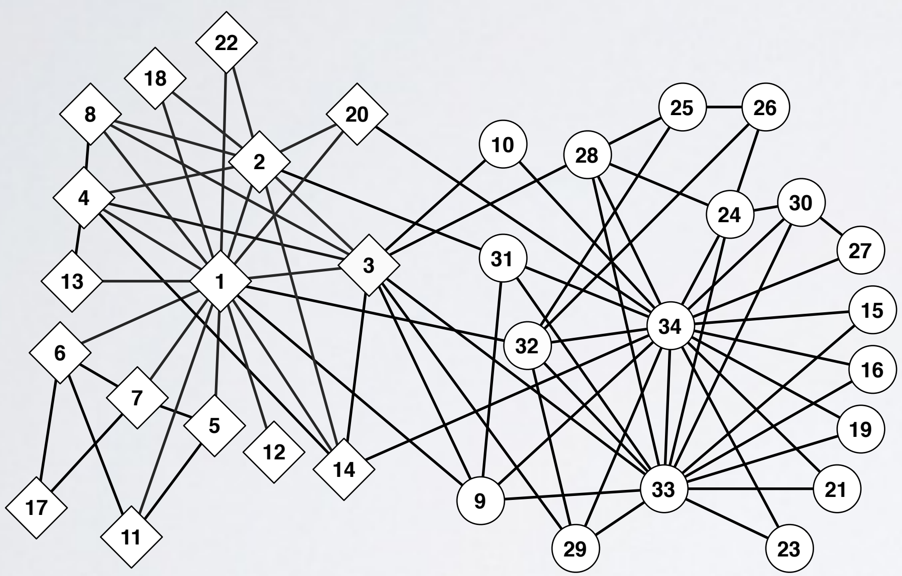
NCAA Schedule 2000

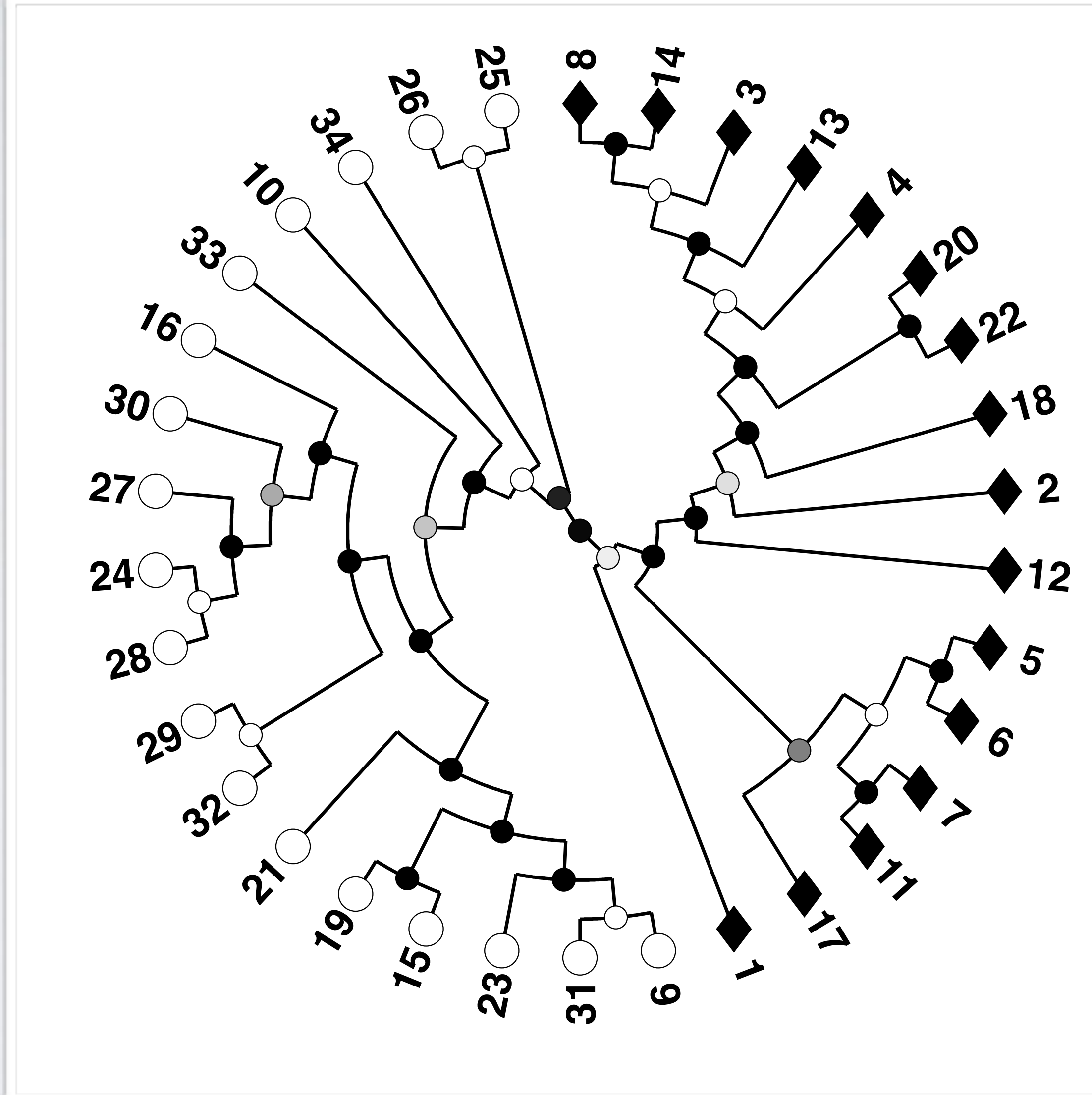


Zachary's Karate Club



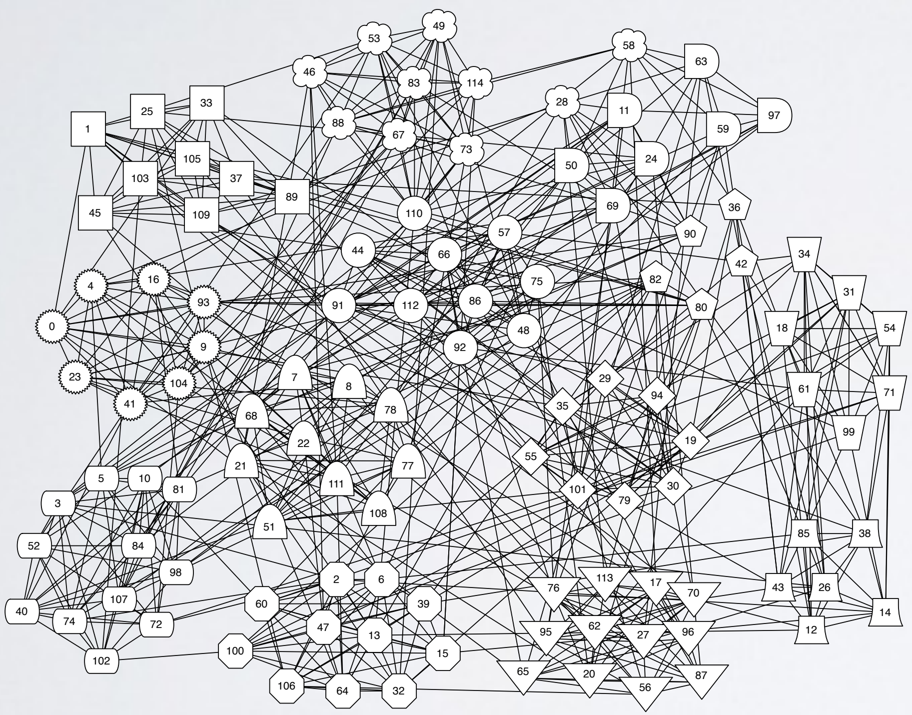
hierarchical communities

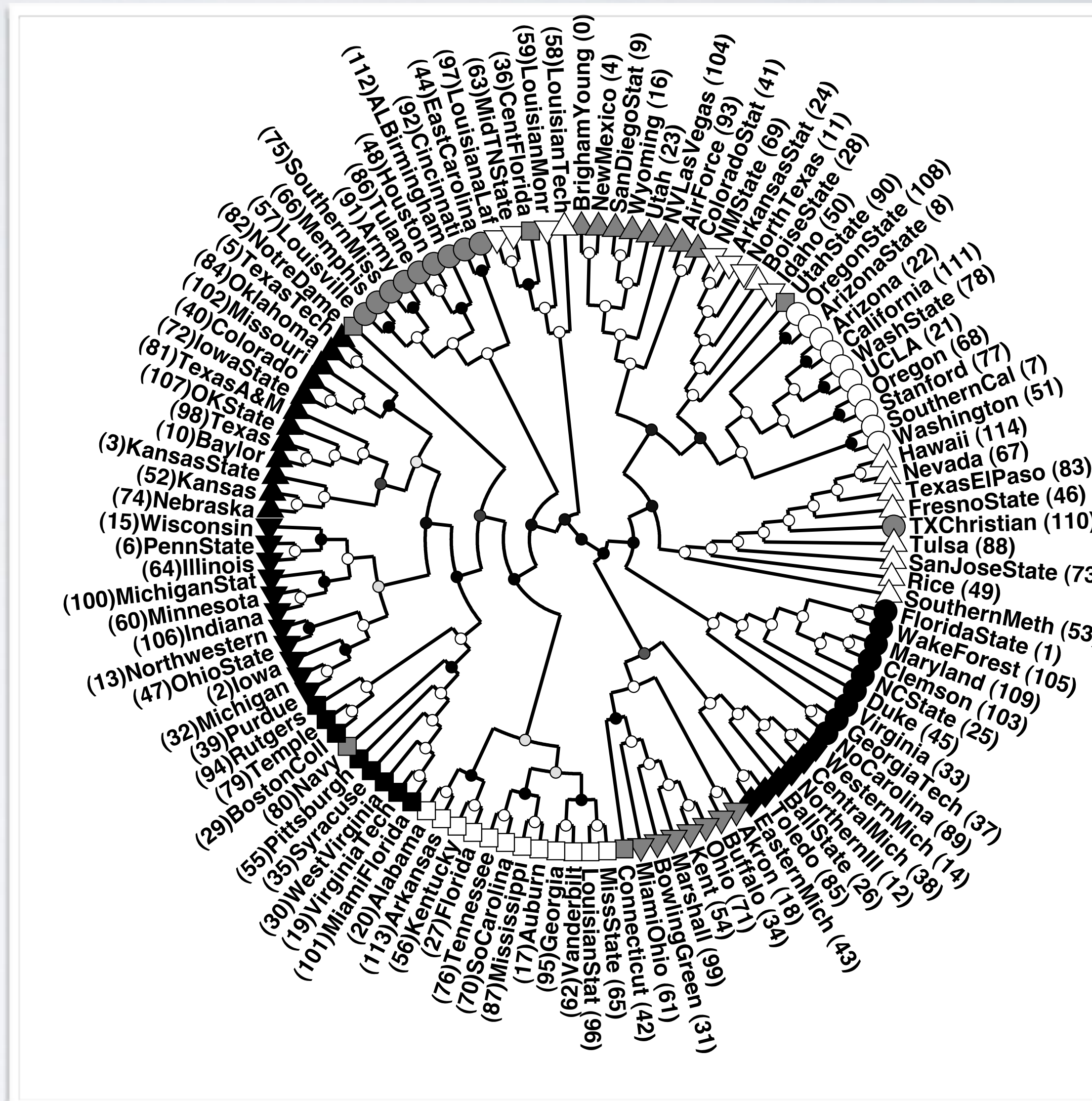




MAP

hierarchical communities





MAP

hierarchical communities

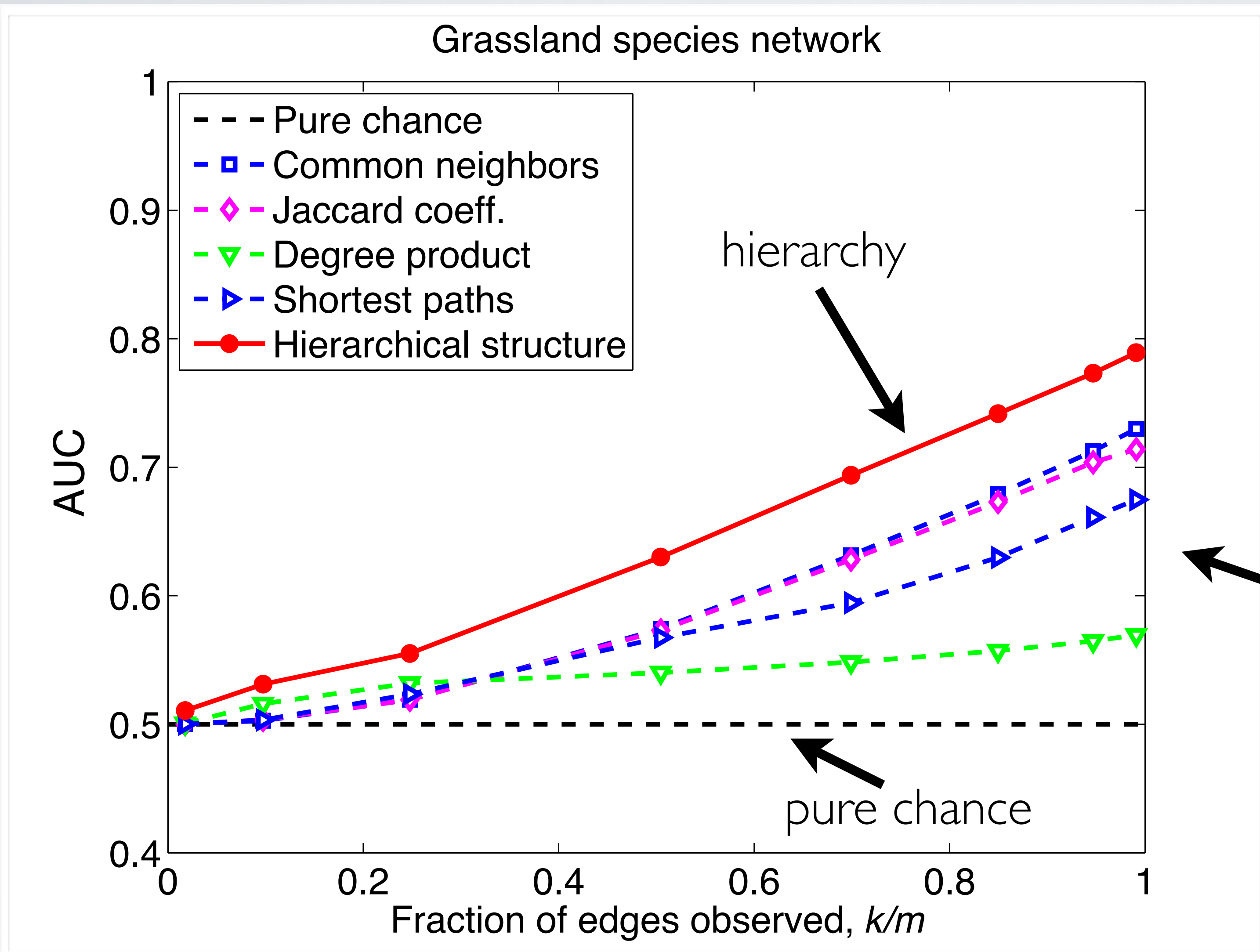
link prediction in networks

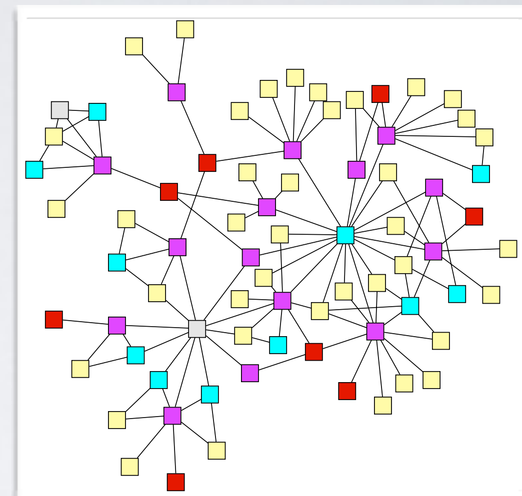
- many networks are sampled
- social nets, foodwebs, protein interactions, etc.
- generative models provide estimate of $\Pr(A_{ij} \mid \theta)$
  for either $A_{ij} = 0$ (missing links) or $A_{ij} = 1$ (spurious links)
- like cross-validation: hold out some adjacencies, $\{A_{ij}\}$
  measure accuracy of algorithm on these

now many approaches to link prediction:

- Liben-Nowell & Kleinberg (2003)
- Goldberg & Roth (2003)
- Szilágyi et al. (2005)
- Guimera & Sales-Pardo (2009)
- and many others

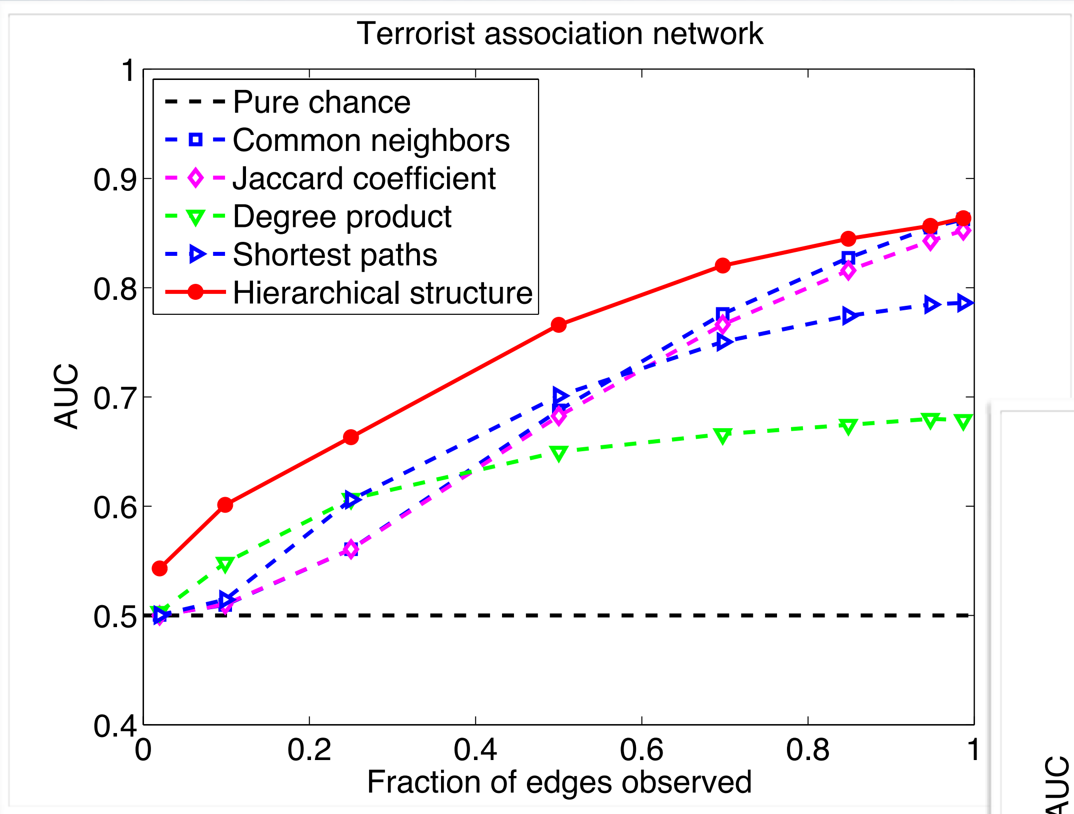
hierarchical communities

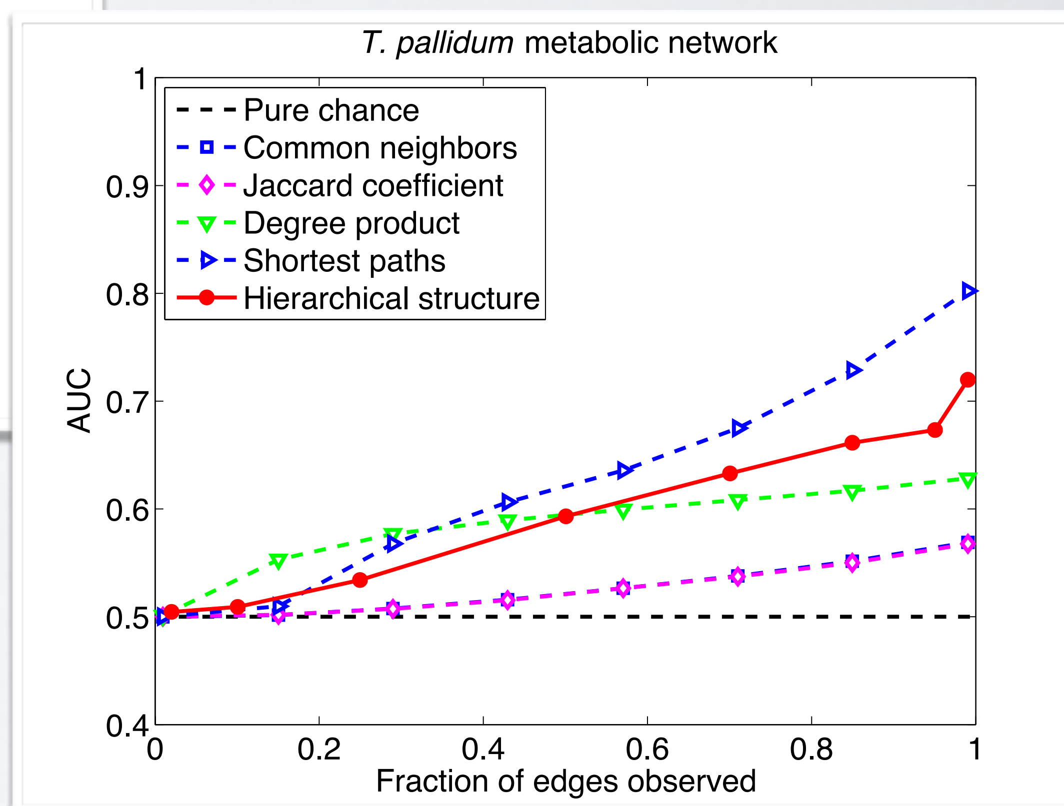




simple predictors

hierarchical communities





No Free Lunch theorem

NFL: *averaged over all possible inputs, every [optimization] algorithm performs equally poorly*

Peel et al. (2016) recently proved a No Free Lunch theorem for community detection

- this implies a spectrum of *specialized* vs. *general* algorithms
- **general algorithms** are very flexible (like the SBM) and can learn a wide variety of structural patterns, but are "weak" at doing so
- **specialized algorithms** are less flexible and can make more assumptions, e.g., look only for assortative groups, but are very "strong" when applied to inputs that match their assumptions
- the link prediction results show evidence of this: the hierarchical model does pretty well on all three problems, but is not always the best predictor

Peel et al., Preprint, arxiv:1608.05878 (2016)

hierarchical communities

other approaches

hierarchical communities

other approaches

PHYSICAL REVIEW X **4,** 011047 (2014)

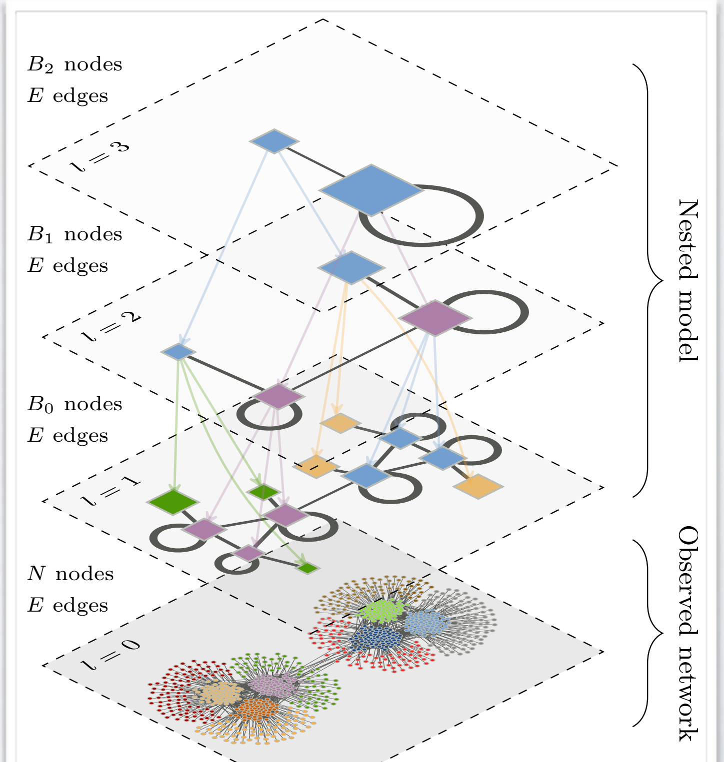
Hierarchical Block Structures and High-Resolution Model Selection in Large Networks

Tiago P. Peixoto*

Institut für Theoretische Physik, Universität Bremen, Hochschulring 18, D-28359 Bremen, Germany

edge counts e_{rs} among blocks are another network

fit another SBM to these, repeat

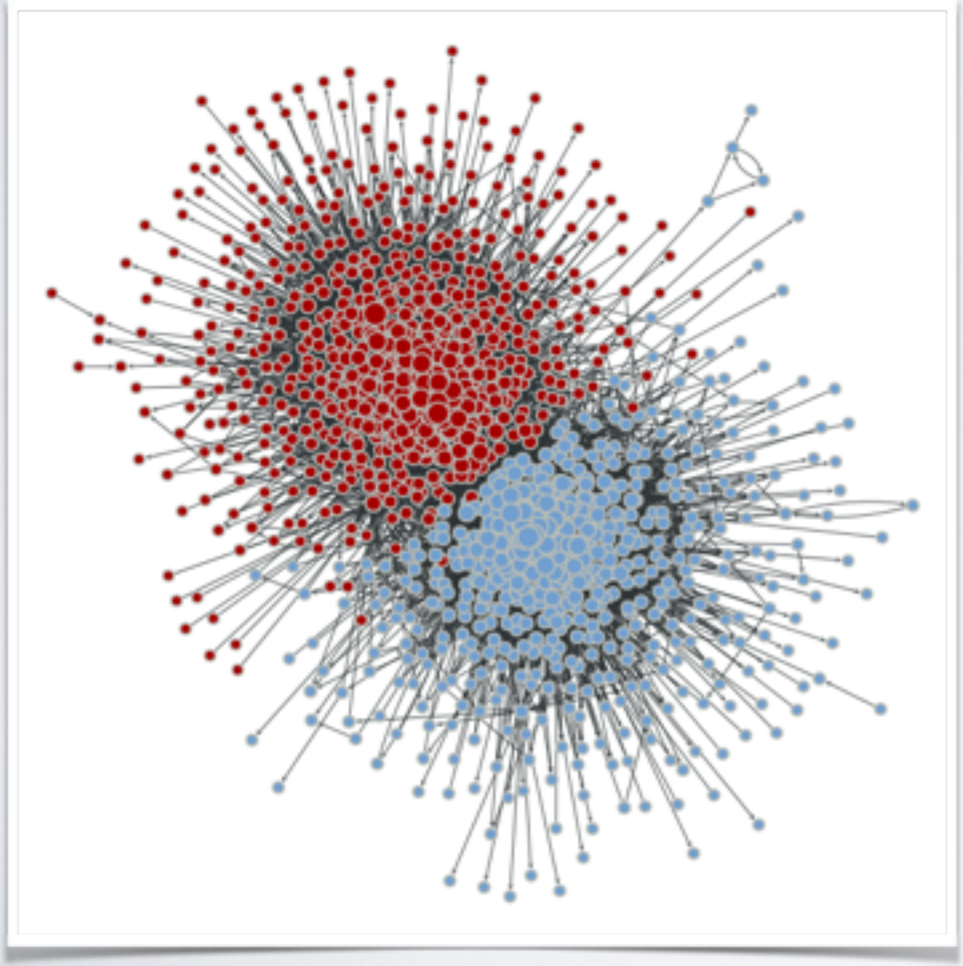


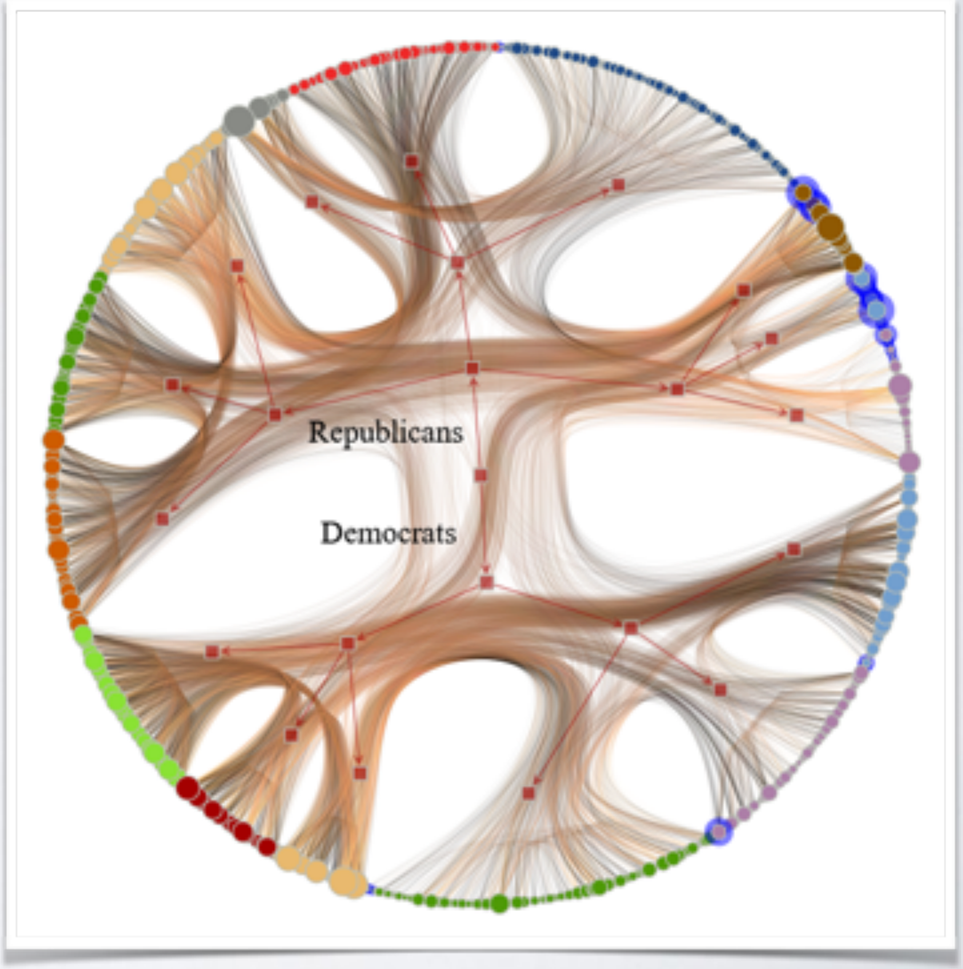
Peixoto, *Phys. Rev. X* **4**, *011047* (2014)

hierarchical communities

other approaches (hierarchical SBM)

political blogs (2004) network





Peixoto, *Phys. Rev. X* **4**, *011047* (2014)

limits of statistical inference

limits of statistical inference

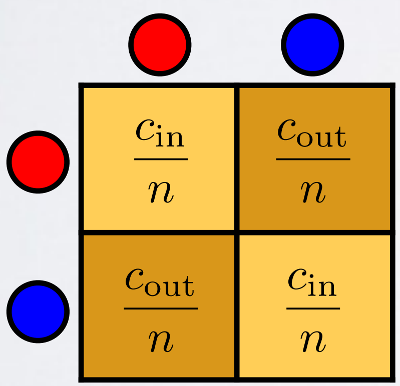
community structure in networks

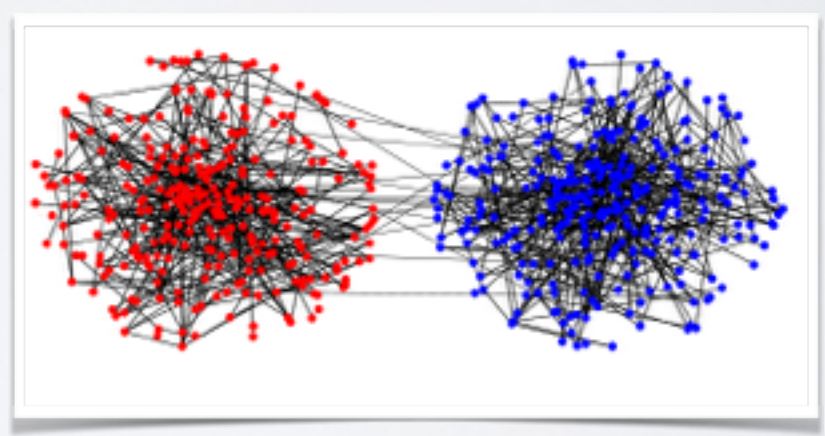
- dozens of algorithms for finding it
- generative models among the most powerful
- *how methods fail is as important as how they succeed*
- even if communities exist in a network, they may not be detectable

limits of statistical inference

planted partition problem

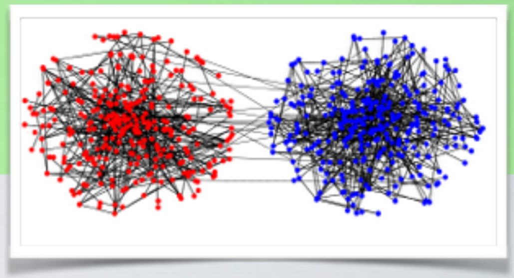
- synthetic data with known communities
- 2 groups, equal sized
- mean degree c
- parameterized strength of communities $\epsilon = c_{\text{out}}/c_{\text{in}}$





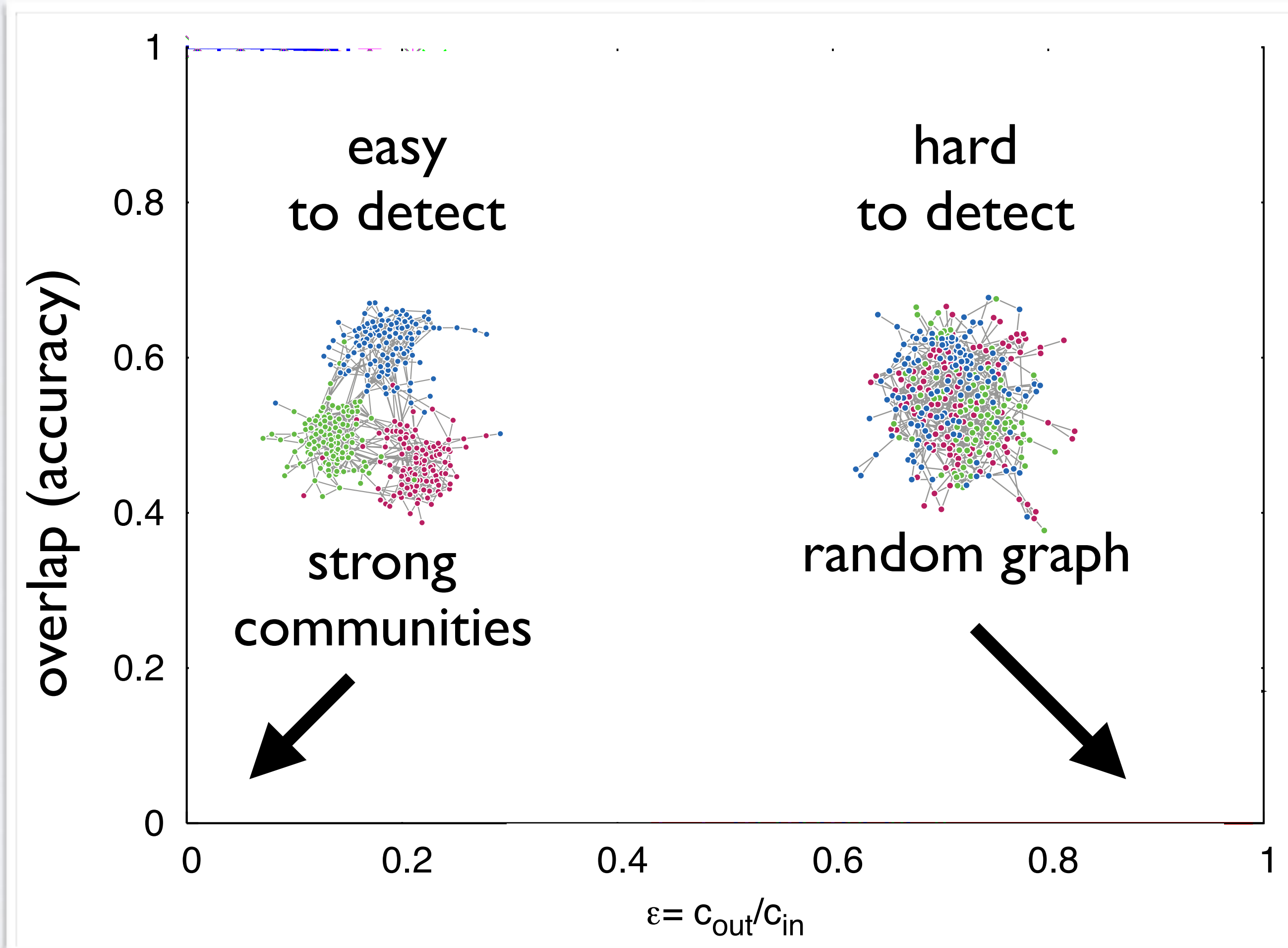
Decelle, Krzakala, Moore, & Zdeborová, *Phys. Rev. Lett.* **107**, 065701 (2011)

limits of statistical inference



planted partition problem

- synthetic data with known communities
- 2 groups, equal sized
- mean degree c



Decelle, Krzakala, Moore, & Zdeborová, *Phys. Rev. Lett.* **107**, 065701 (2011)

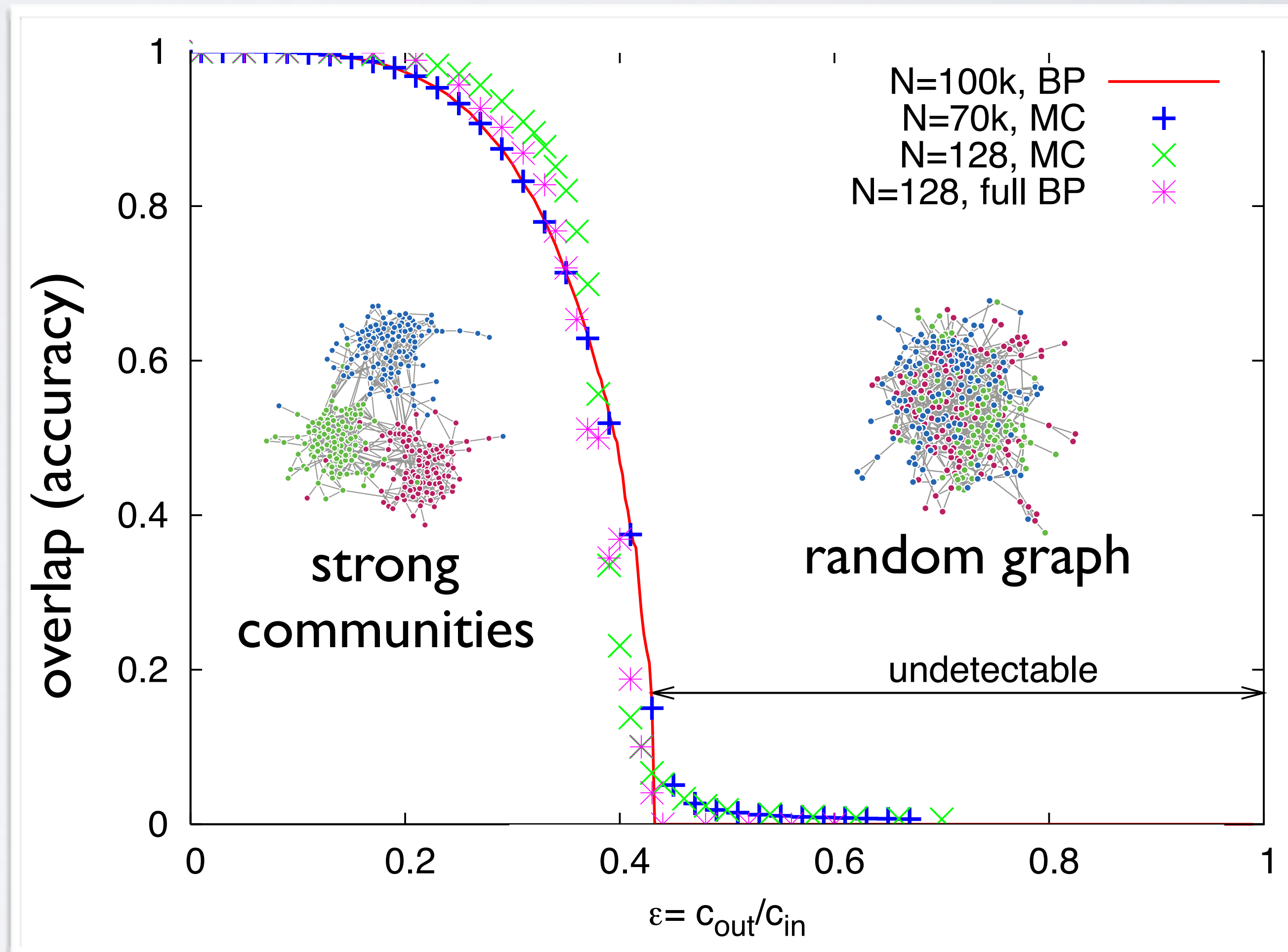
limits of statistical inference

planted partition problem

- synthetic data with known communities
- 2 groups, equal sized
- mean degree c

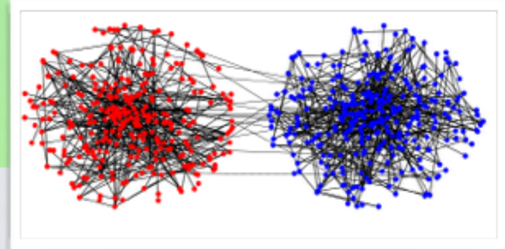
- 2nd order phase transition in detectability
- overlap goes to 0 for

$$\epsilon \geq \frac{c - \sqrt{c}}{c + \sqrt{c}(k-1)}$$



Decelle, Krzakala, Moore, & Zdeborová, *Phys. Rev. Lett.* **107**, 065701 (2011)

limits of statistical inference



planted partition problem

- for 2 groups, phase transition is information theoretic

  no algorithm can exist that detects these communities (better than chance)

- when communities are strong, most algorithms succeed
- when networks & communities are very sparse = trouble

- recently generalized to dynamic networks (Ghasemian et al. 2015)
- hierarchical block models (Peixoto 2014) and node metadata (Newman & Clauset 2016) both improve detectability

Decelle, Krzakala, Moore, & Zdeborová, *Phys. Rev. Lett.* **107**, 065701 (2011)
Ghasemian et al., arxiv:1506.0679 (2015)
Peixoto, *Phys. Rev. X* **4**, *011047* (2014)
Newman & Clauset, *Nature Communications,* to appear (2016)

fin