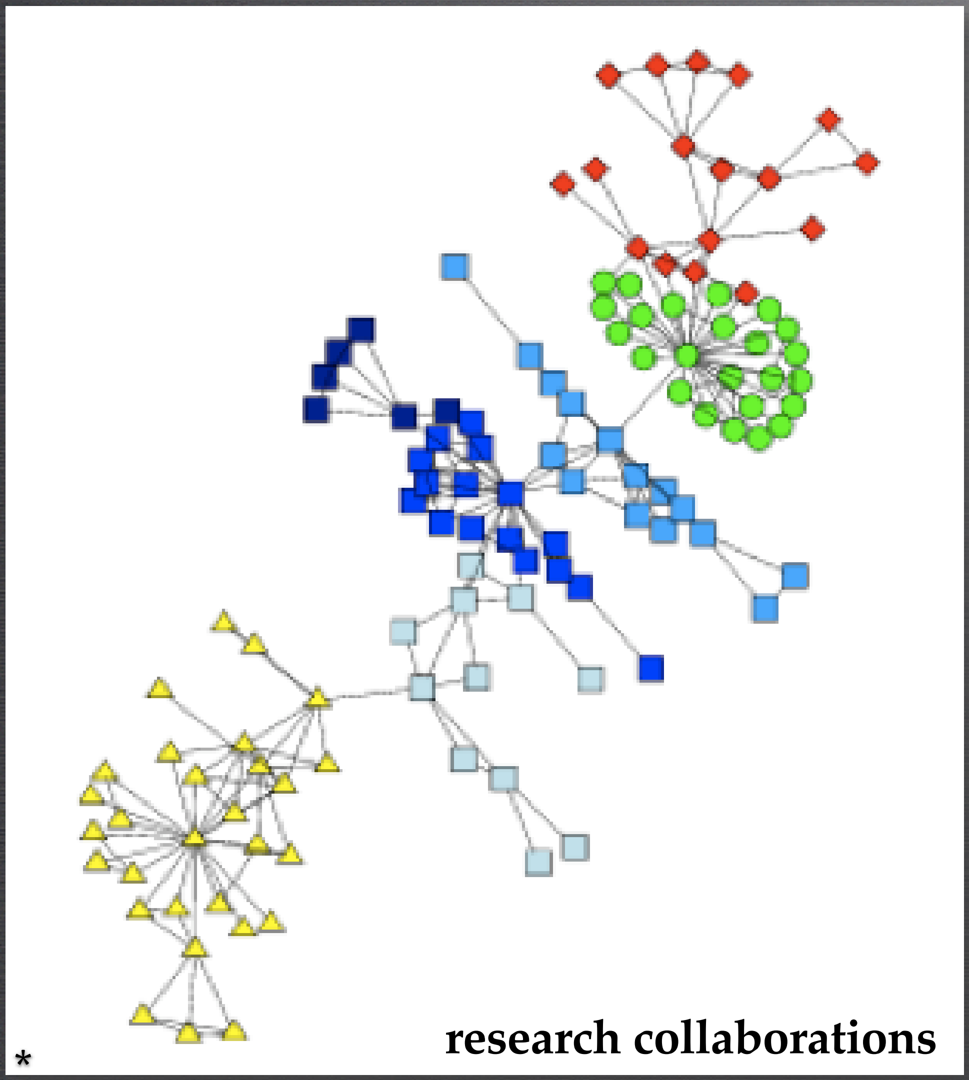
The Hierarchical Structure of Networks

Aaron Clauset
Santa Fe Institute

4 August 2008
SFI / CAIDA Workshop
Networks and Navigation

First, Some Pictures

social groups or communities



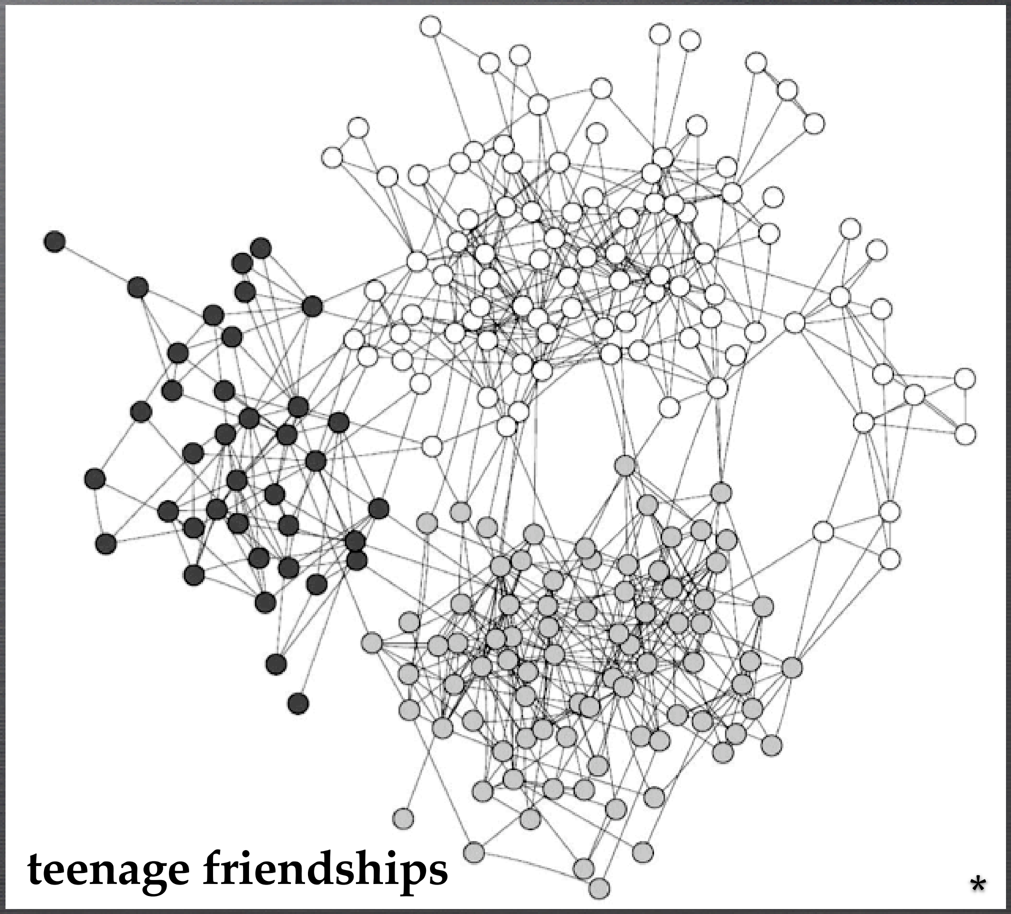
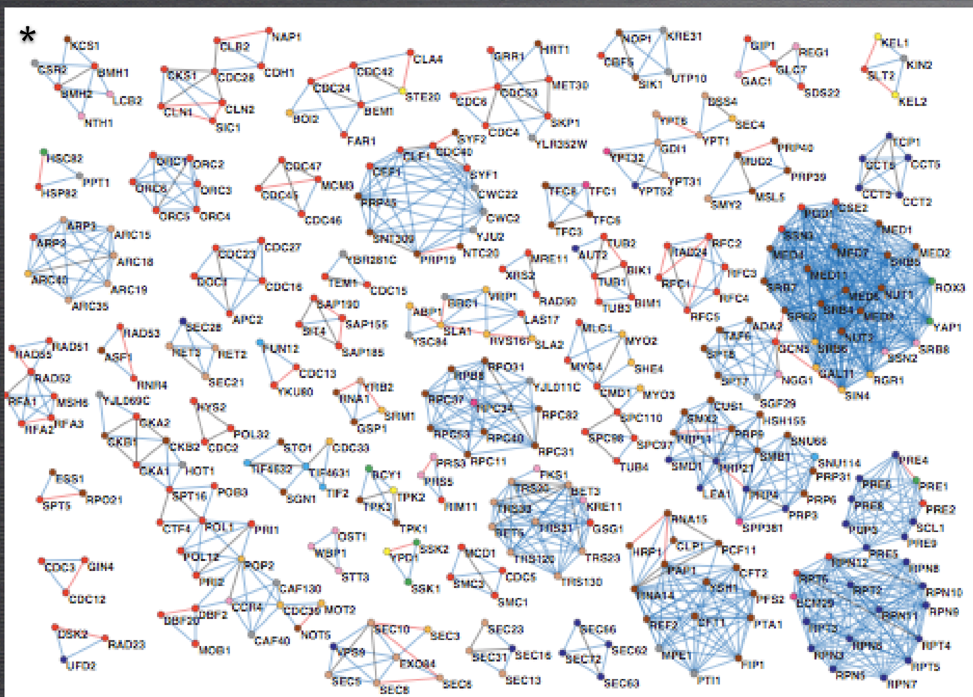


*image stolen from elsewhere

functional(?) clusters, hierarchies



proteins

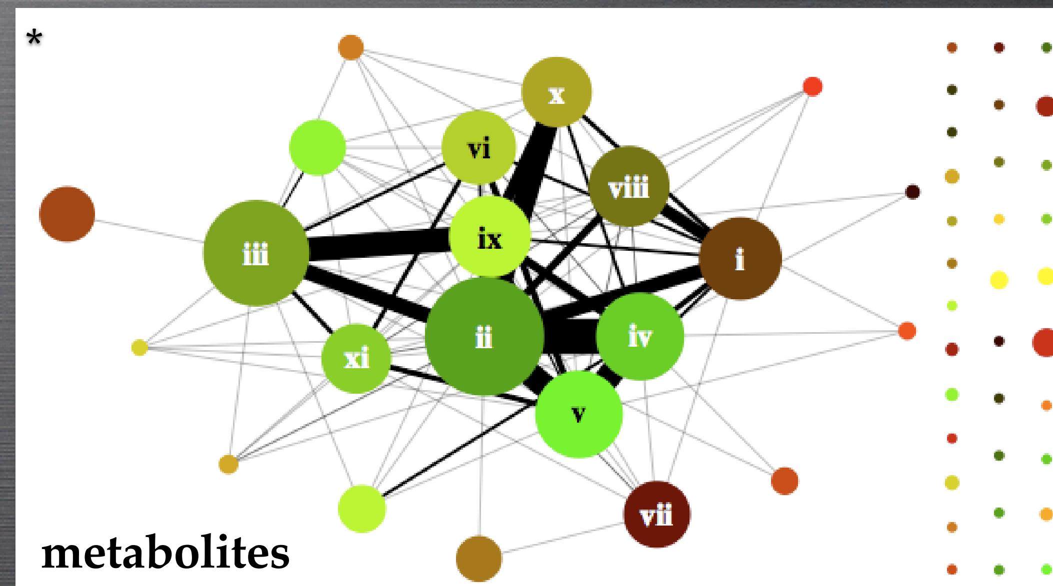
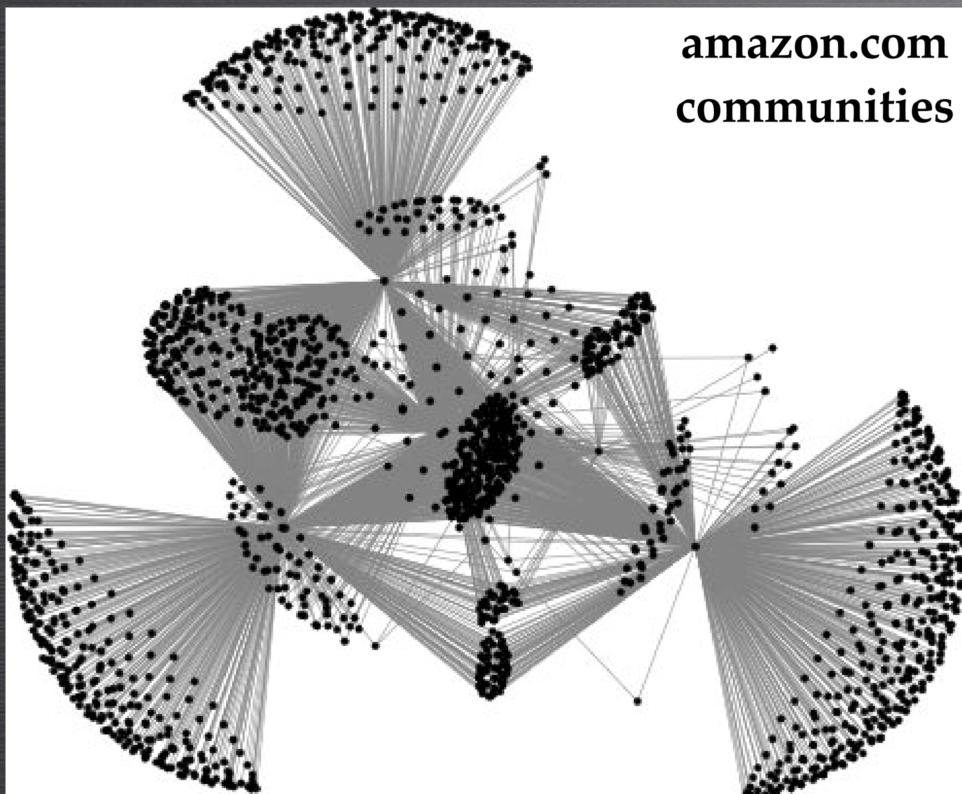


*image stolen from elsewhere

co-purchasing (topical?) groups



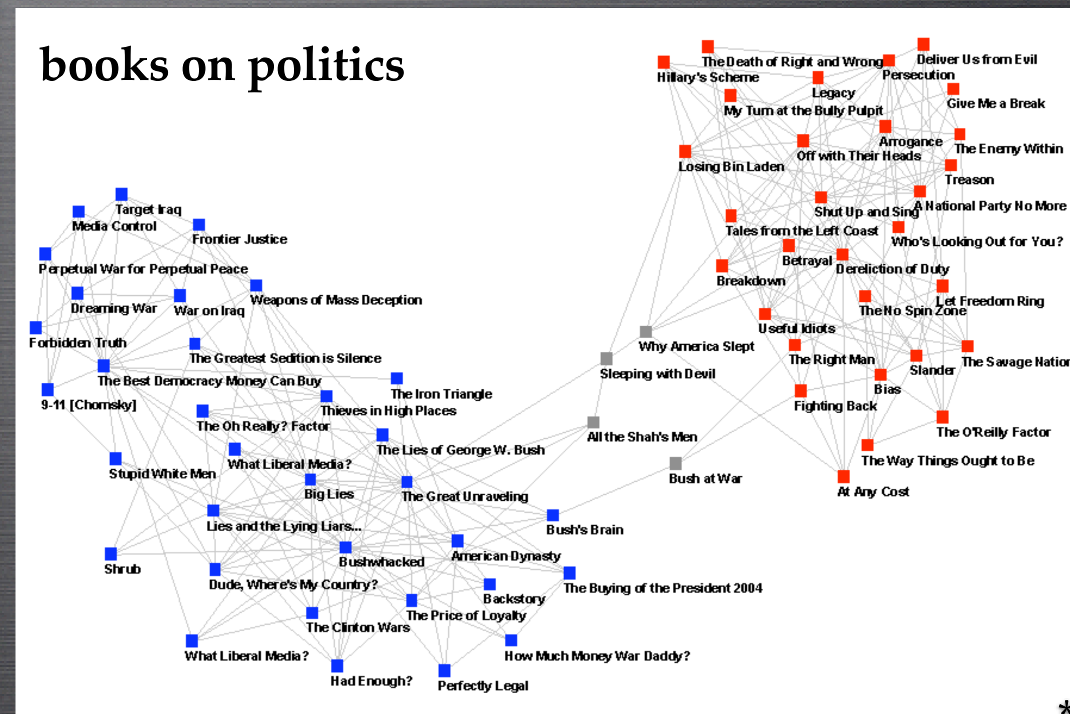


*image stolen from elsewhere

A Question

How can we extract

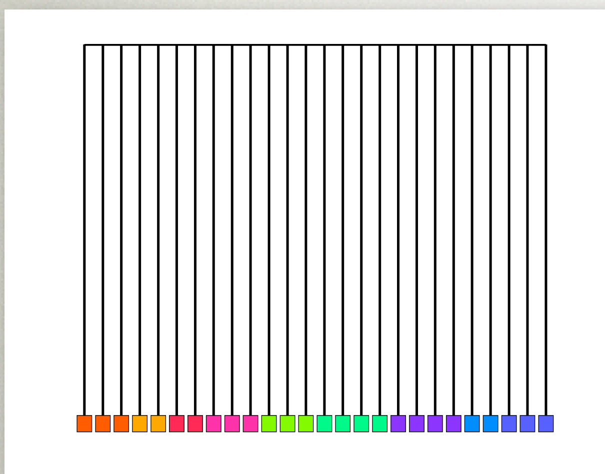
- structural patterns
- at many scales
- in a rigorous fashion

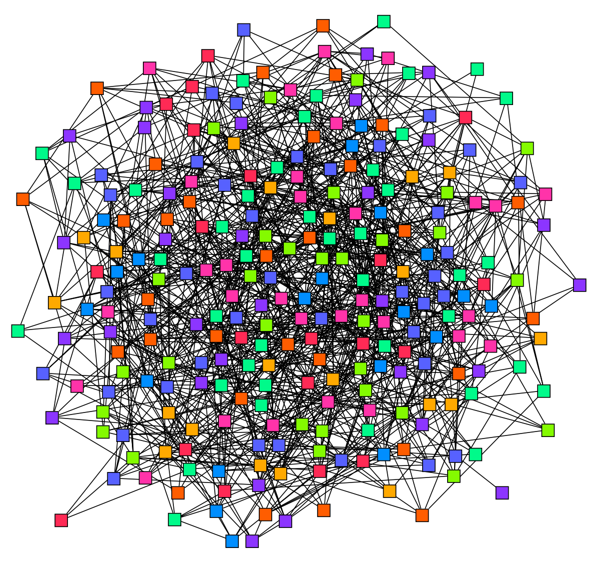
from complex networks?

What is Structure?

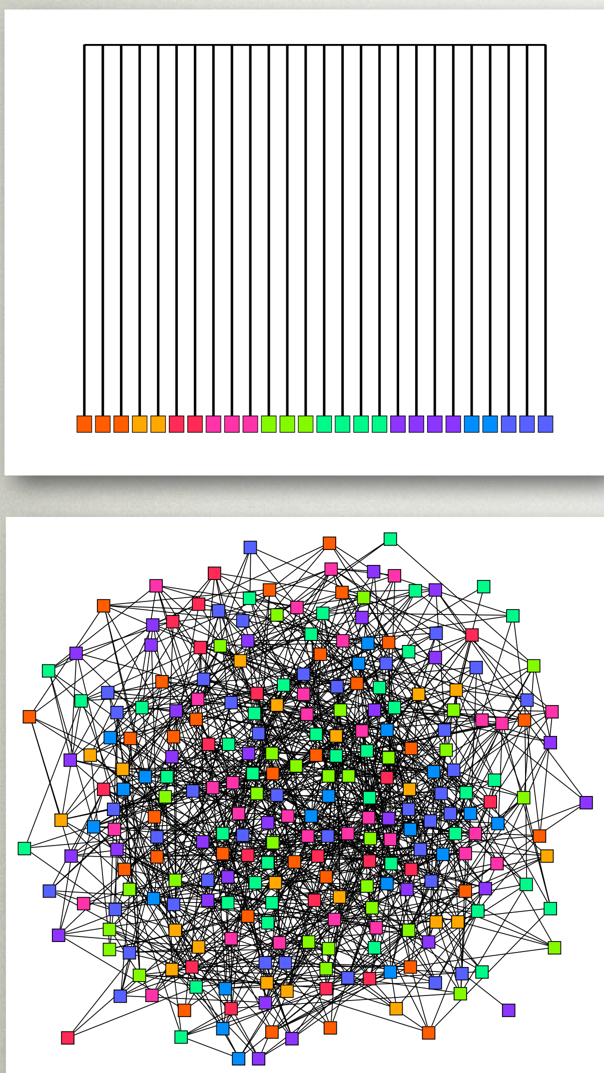
some stylized ideas

no structure

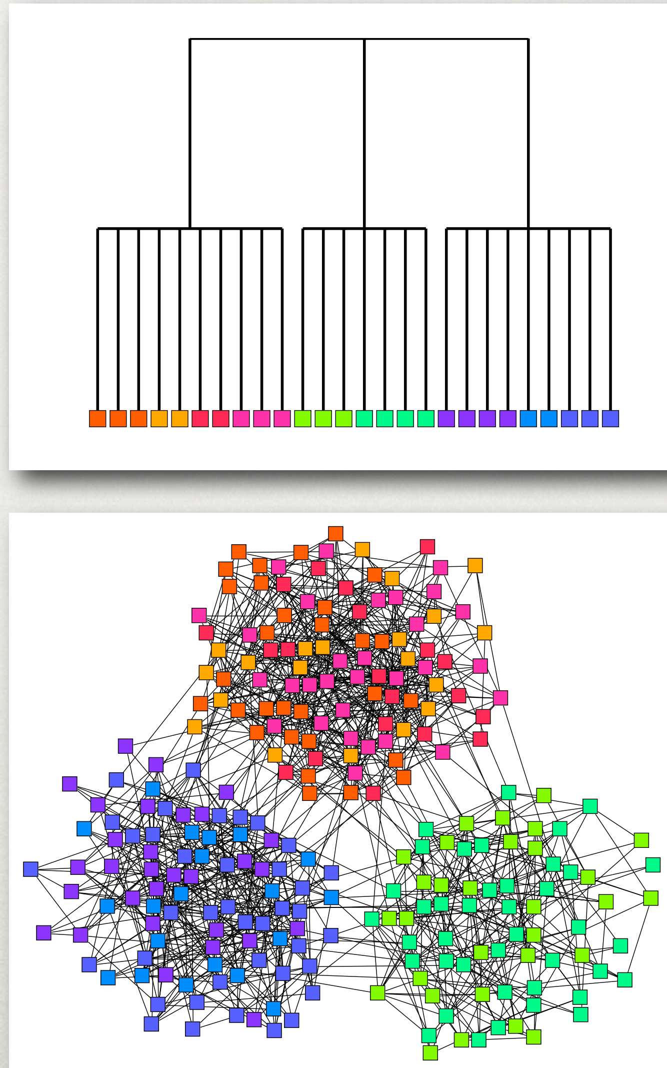




no structure

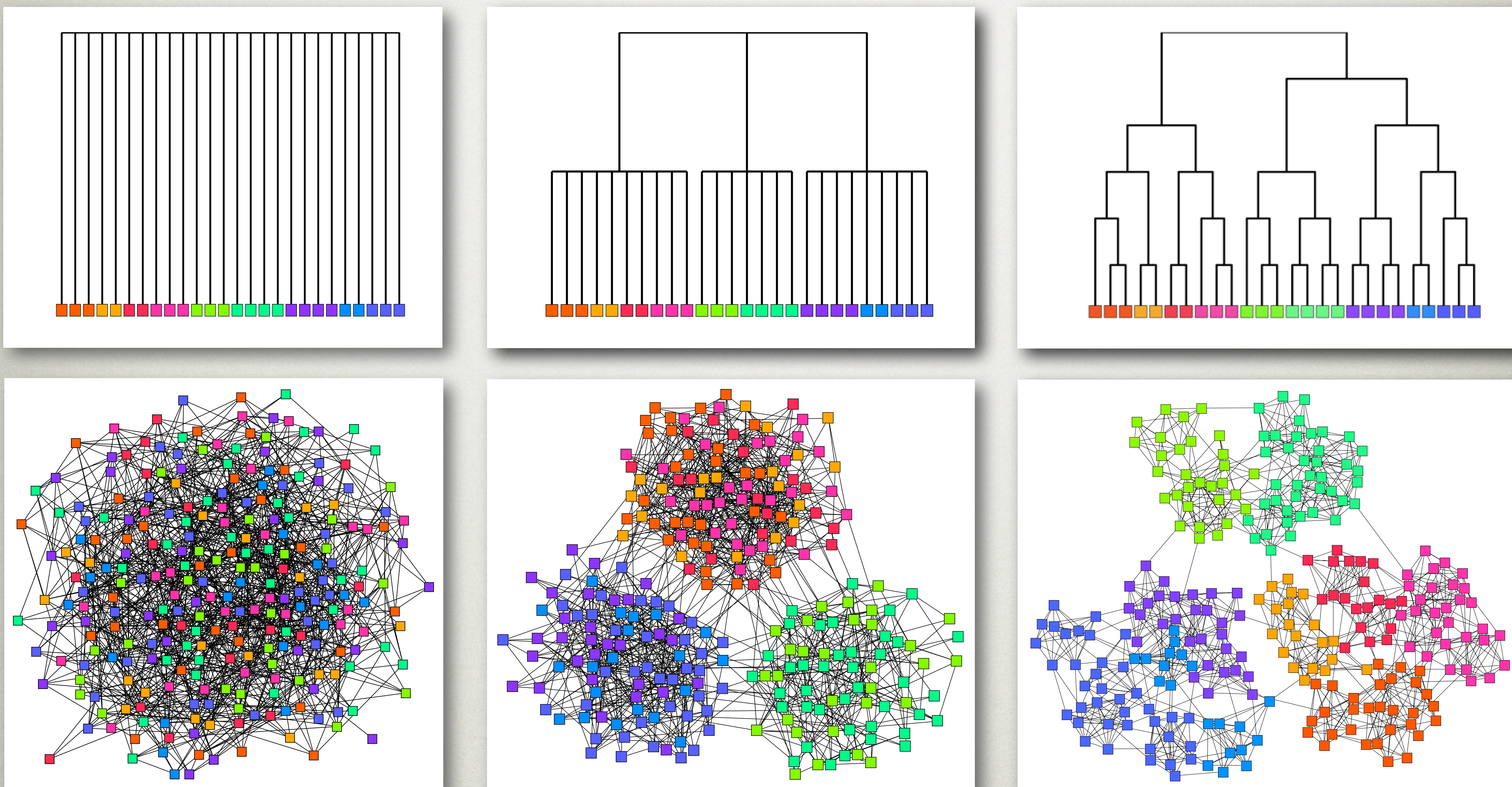


modular structure



one scale

no structure **modular structure** **hierarchical structure**



one scale **multi-scale**

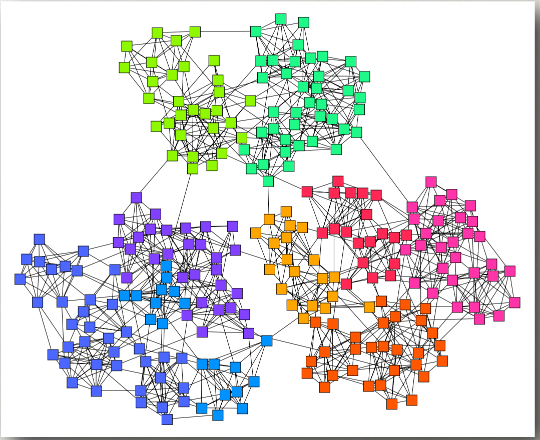
A Question

How can we extract

- **hierarchical structure**
- in a rigorous fashion

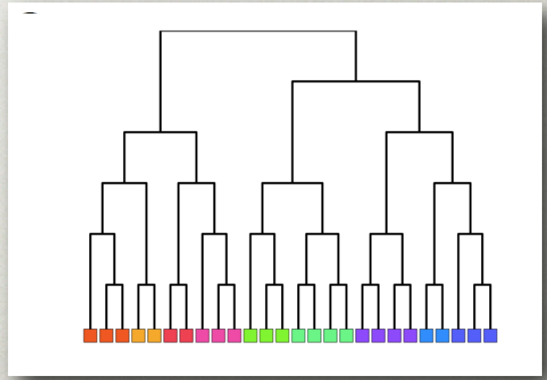
from complex networks?

network data



?

hierarchy

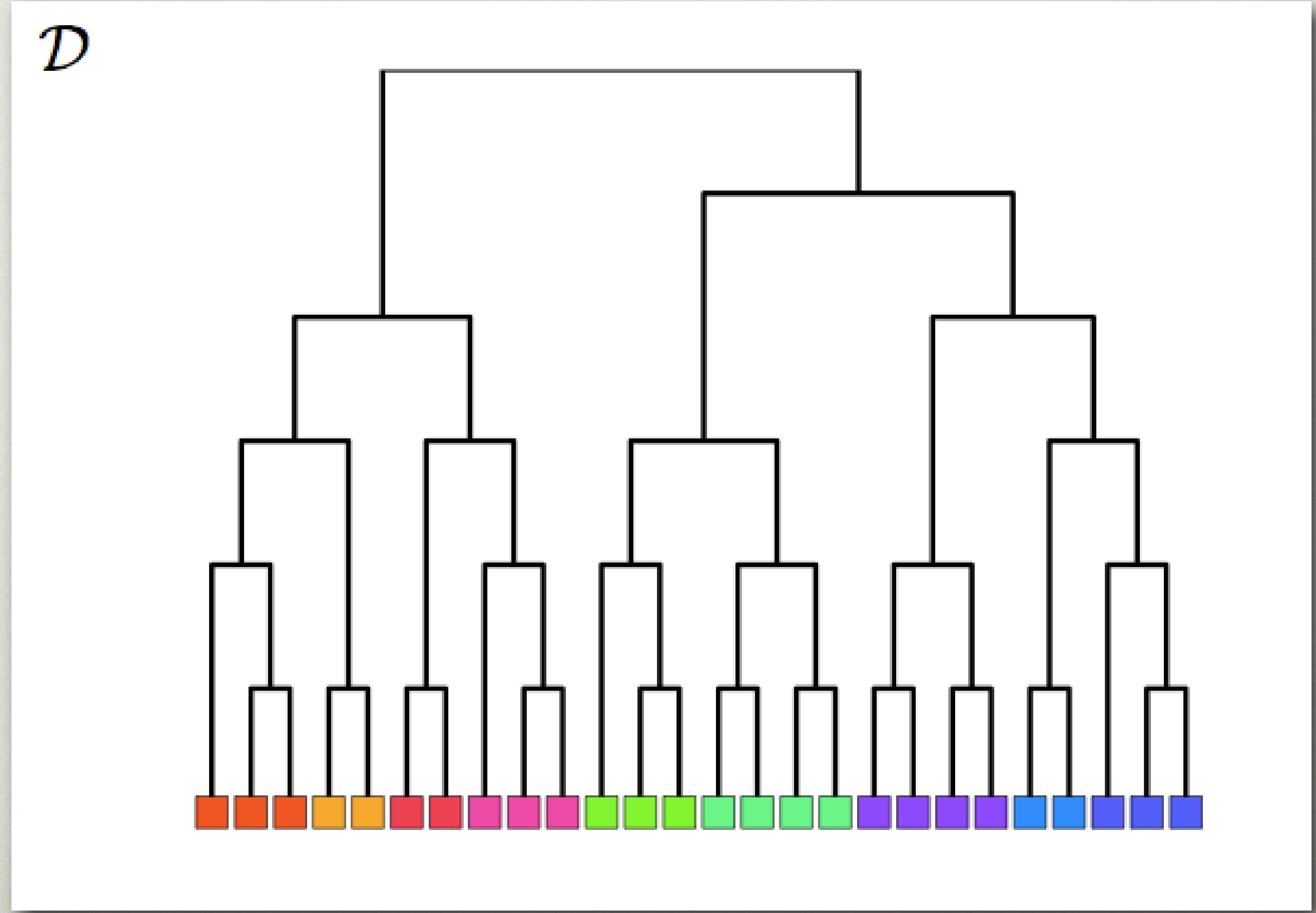


One Approach

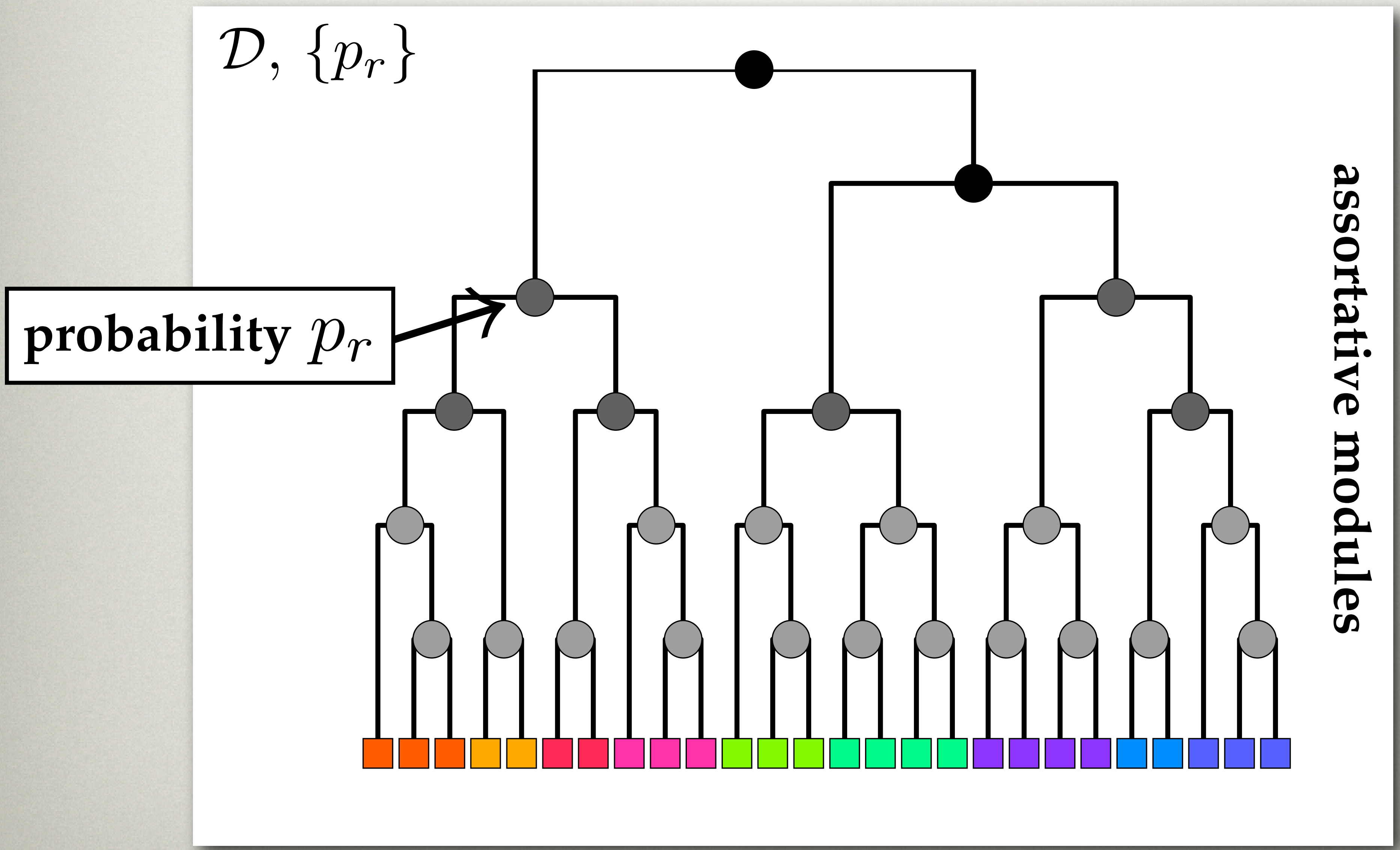
Model-based inference

1. describe how to generate hierarchies (a model)
2. “fit” model to empirical data
3. test “fitted” model
4. extract predictions + insight

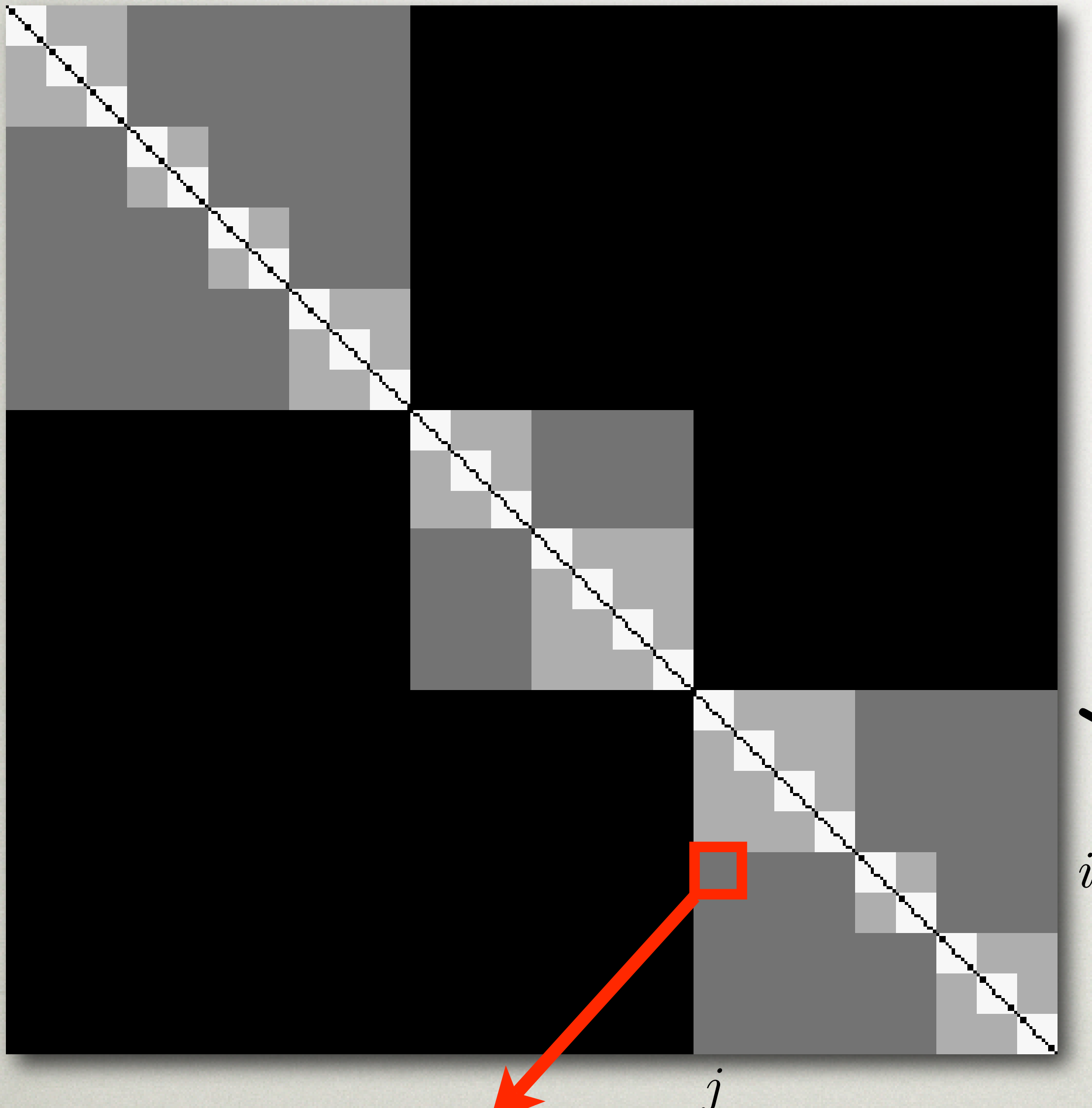
A Model of Hierarchy



A Model of Hierarchy

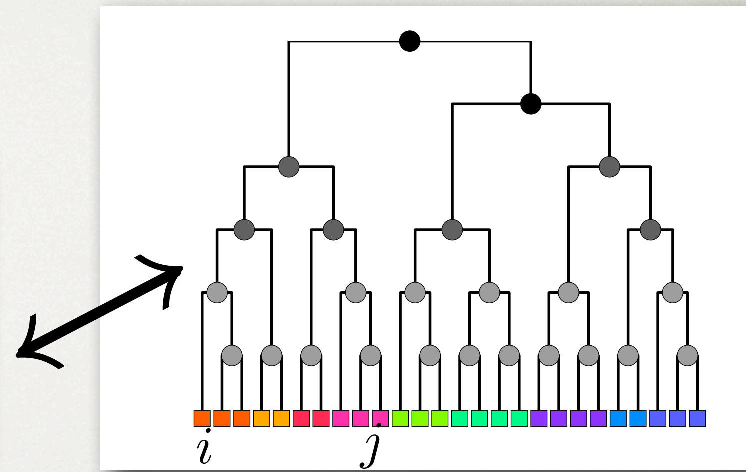


“inhomogeneous” random graph

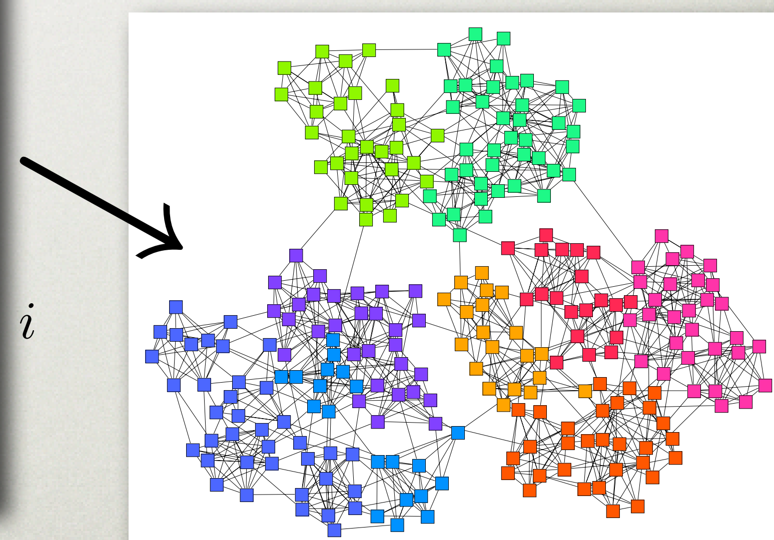


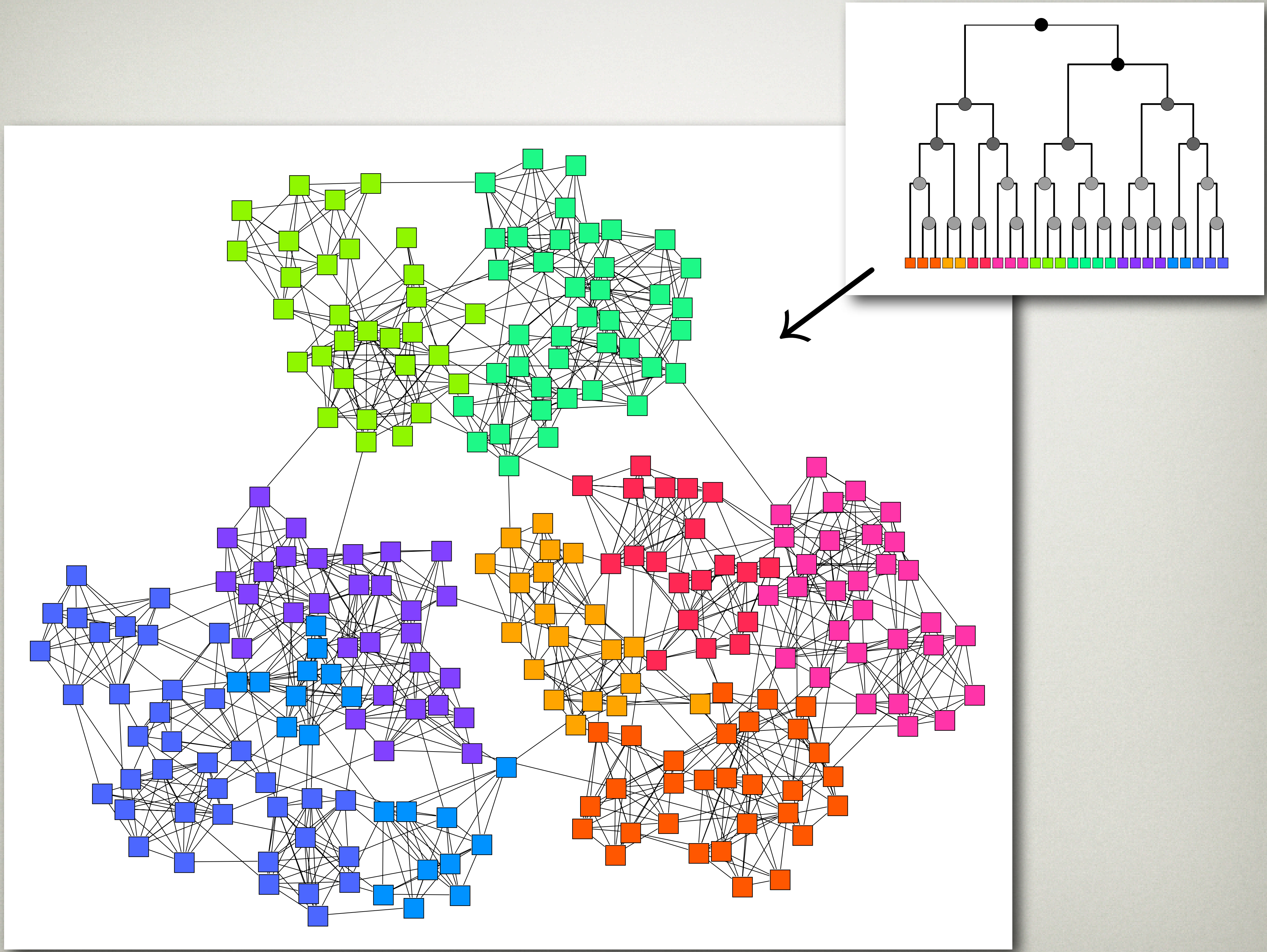
$$\Pr(i,j \text{ connected}) = p_r = p_{(\text{lowest common ancestor of } i,j)}$$

model



instance





Model Features

- explicit model = explicit assumptions
- very flexible (many parameters)
- captures structure at all scales
- arbitrary mixtures of assortativity, disassortativity
- learnable directly from data

Learning From Data

- We use a **Bayesian approach:**
- likelihood function $\mathcal{L} = \Pr(\text{ data } | \text{ model })$

  $\mathcal{L}$ scores **quality** of model

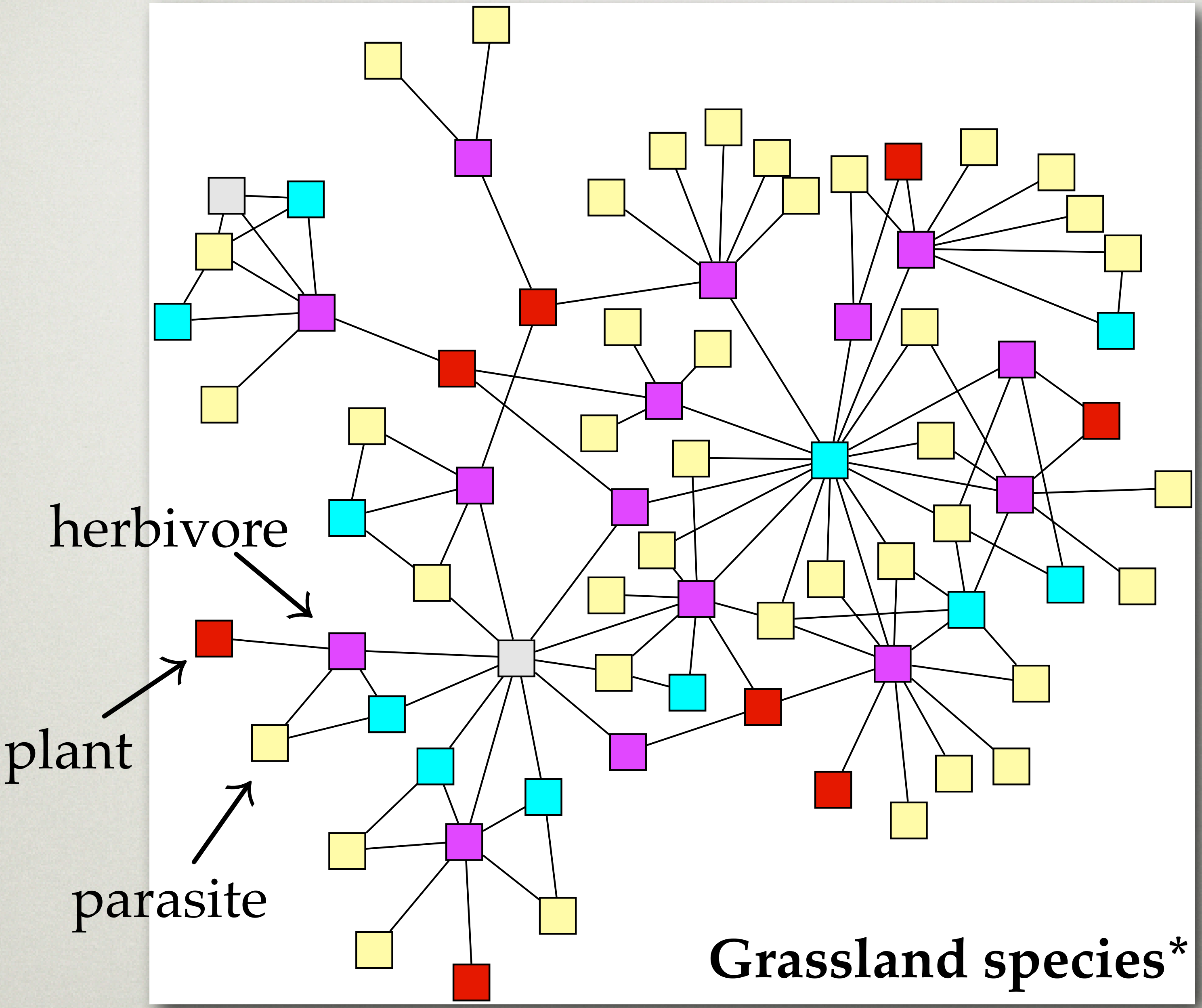
- sample **high quality** models via MCMC
- technical details in arXiv : *physics/0610051* and *Nature* **453**, p98 (2008)

From Graph to Ensemble

From Graph to Ensemble

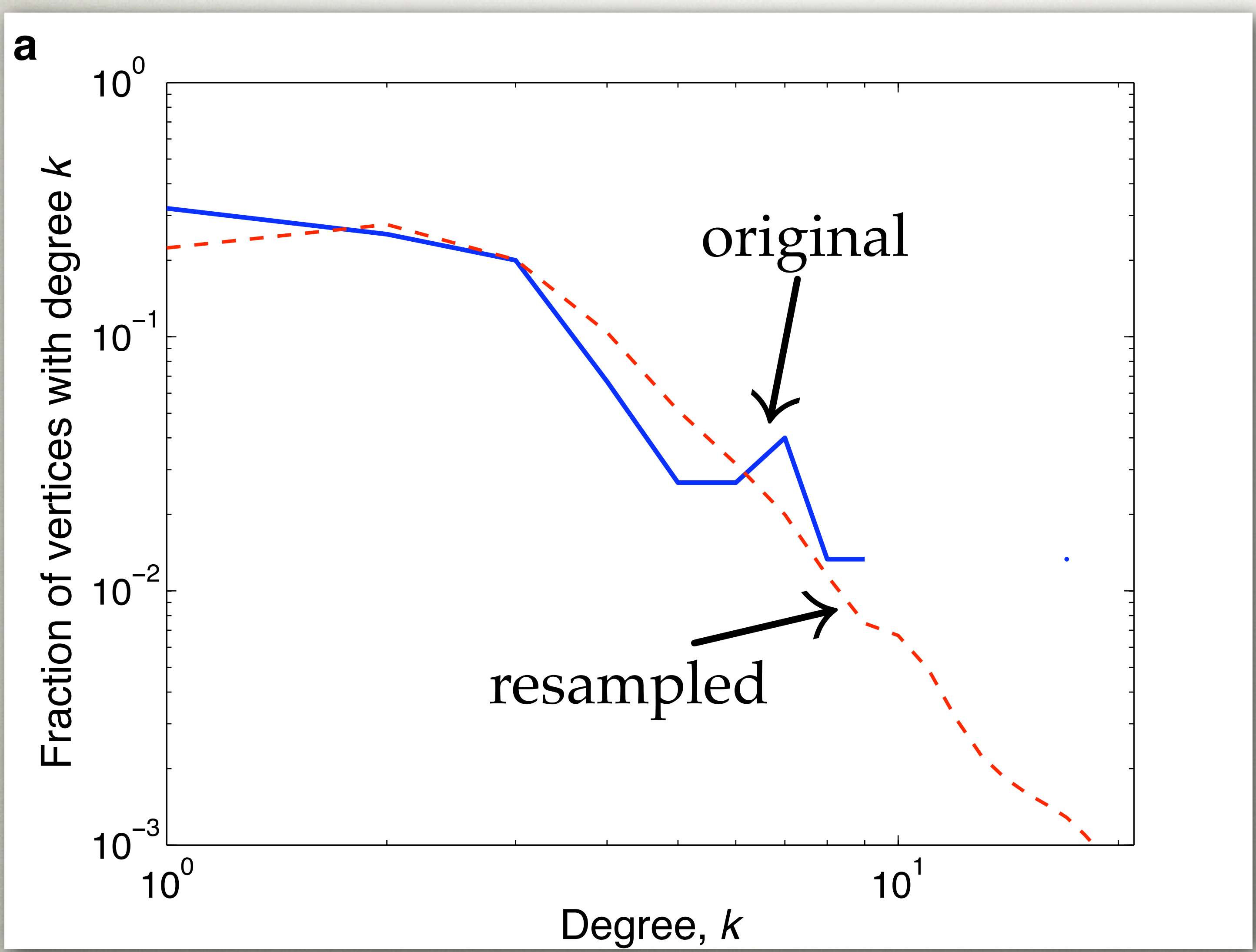
- Given graph G
- run MCMC to equilibrium
- then, for each sampled $\mathcal{D}$, draw a **resampled** graph G' from ensemble

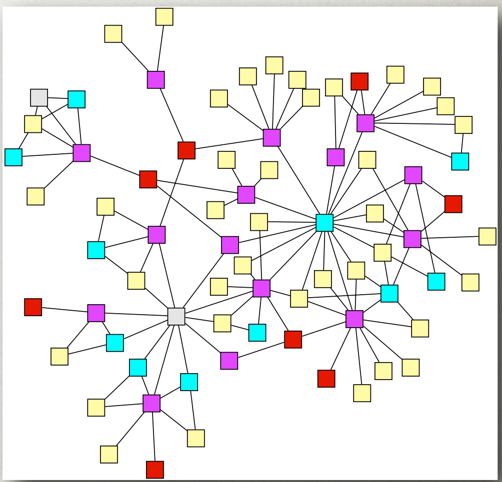
A test: do resampled graphs look like original?



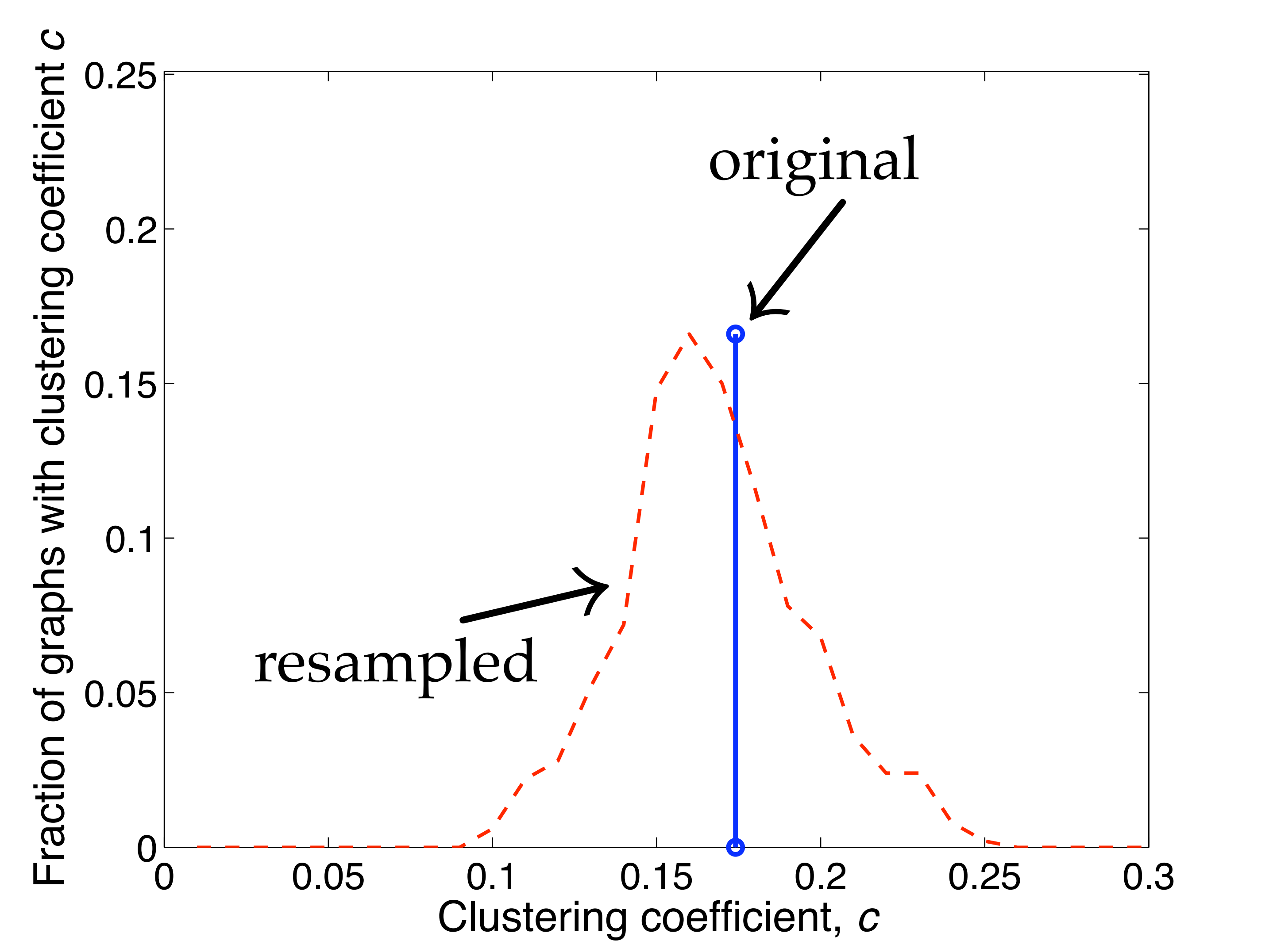
*thank you: Jennifer Dunne

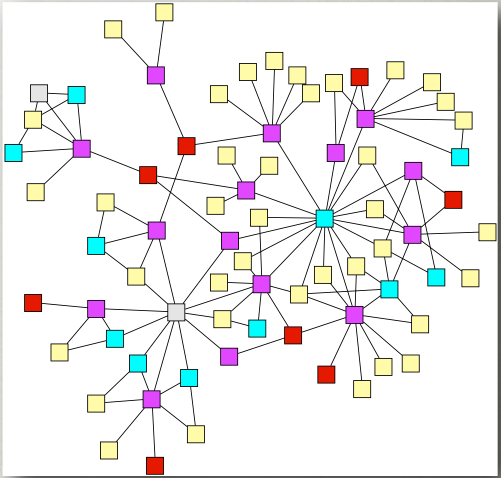
Degree Distribution



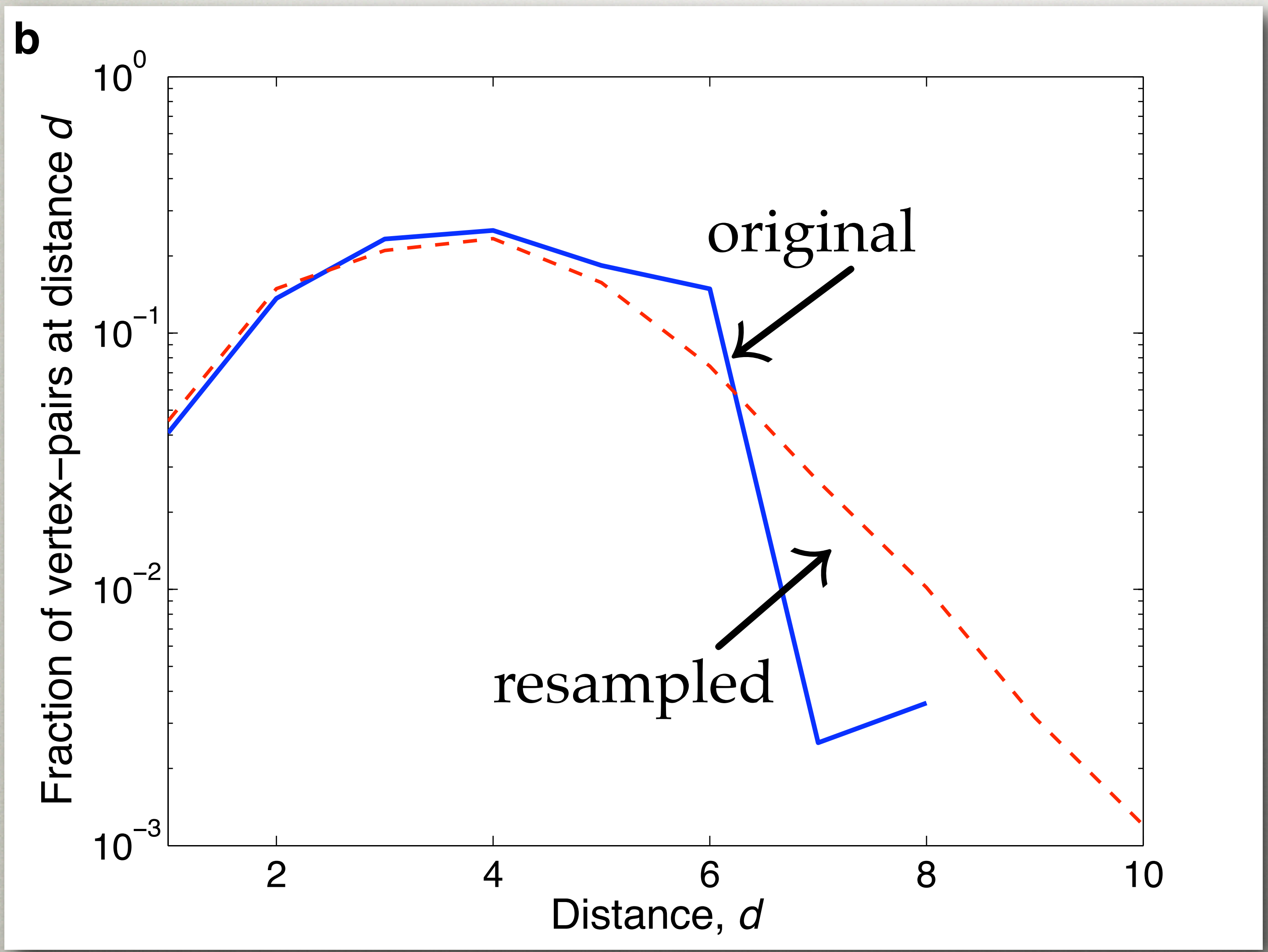


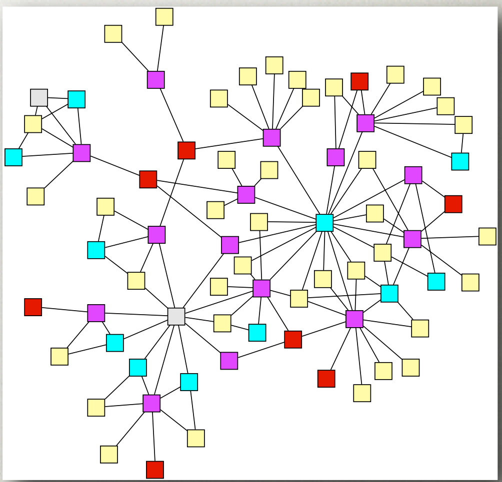
Clustering Coefficient





Distance Distribution





Missing Links

A test: can model predict missing links?

Predicting is Hard

- remove k edges from G
- how easy to guess a missing link?

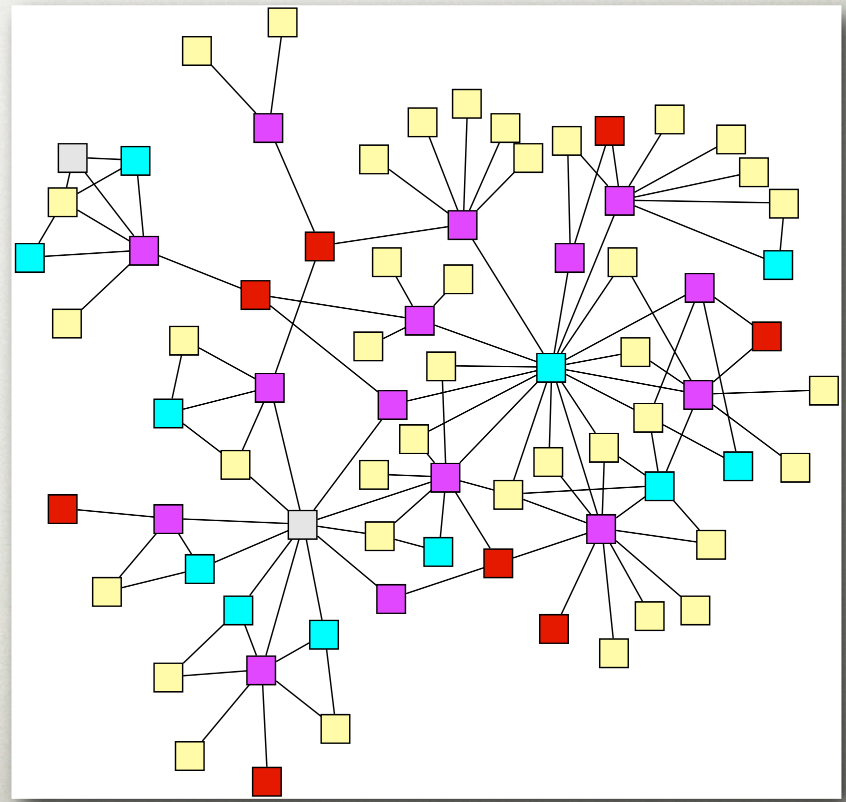
$$p_{\text{guess}} \approx \frac{k}{n^2 - m + k}$$

$$= O(n^{-2})$$

$$n = 75$$

$$m = 113$$

$$p_{\text{guess}} = k/(2662 + k)$$



Predicting Missing Links

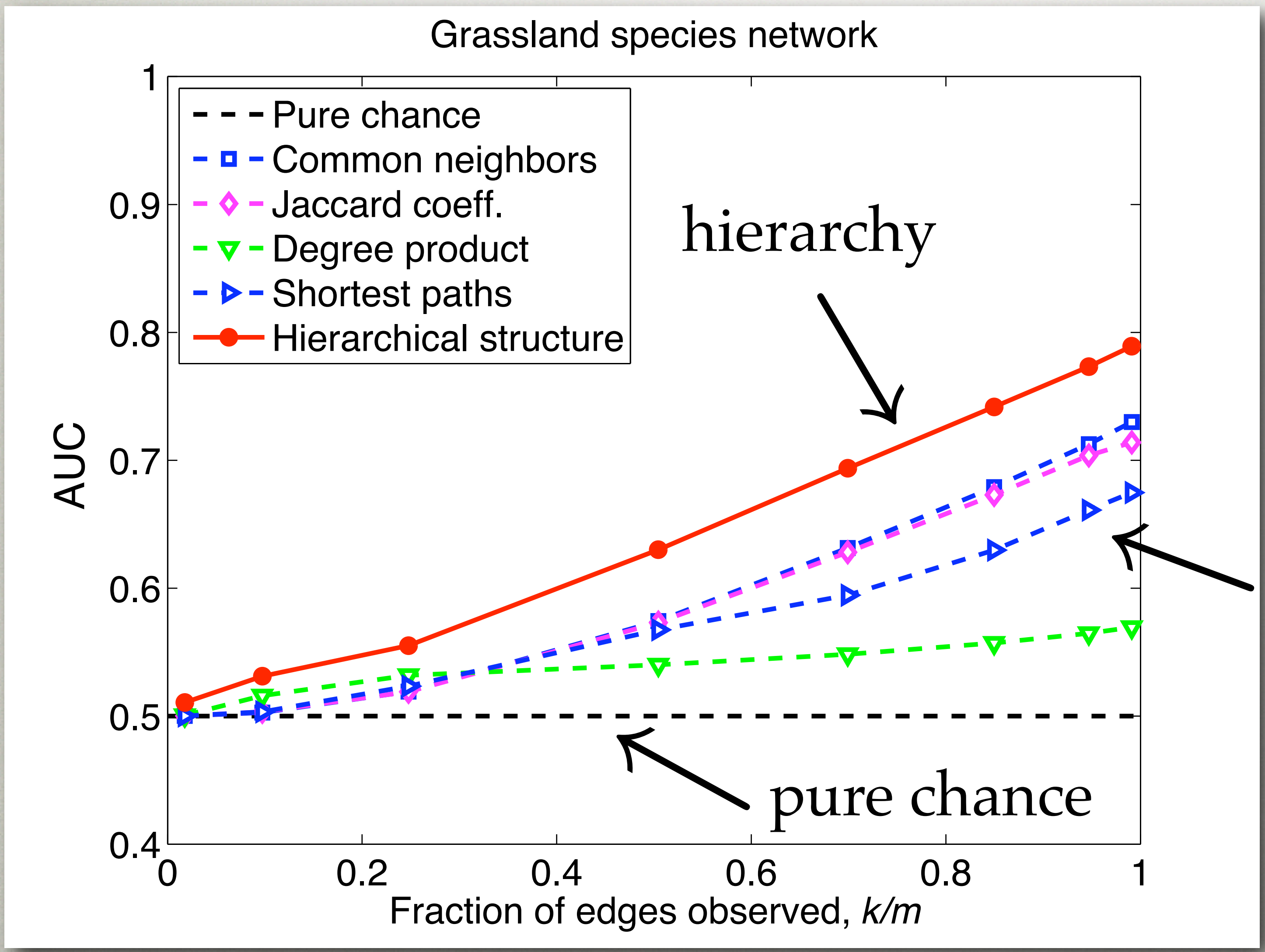
- Given incomplete graph G
- run MCMC to equilibrium
- then, over sampled $\mathcal{D}$, compute average $\langle p_r \rangle$ for links $(i, j) \notin G$
- predict links with high $\langle p_r \rangle$ values are missing

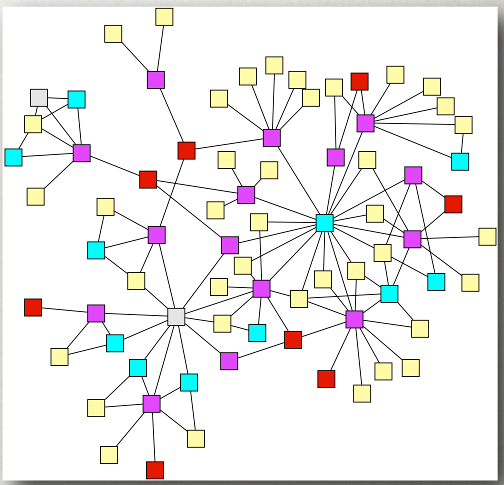
Test idea via leave-*k*-out cross-validation

perfect accuracy: AUC = 1

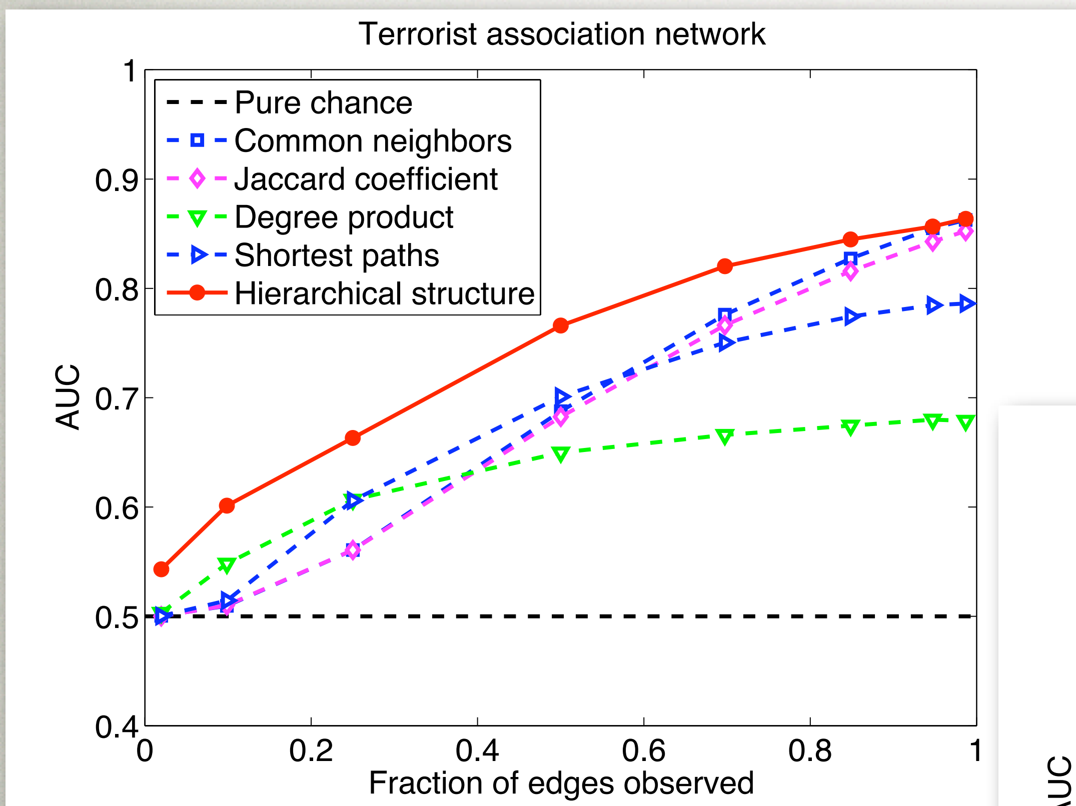
no better than chance: AUC = 1/2

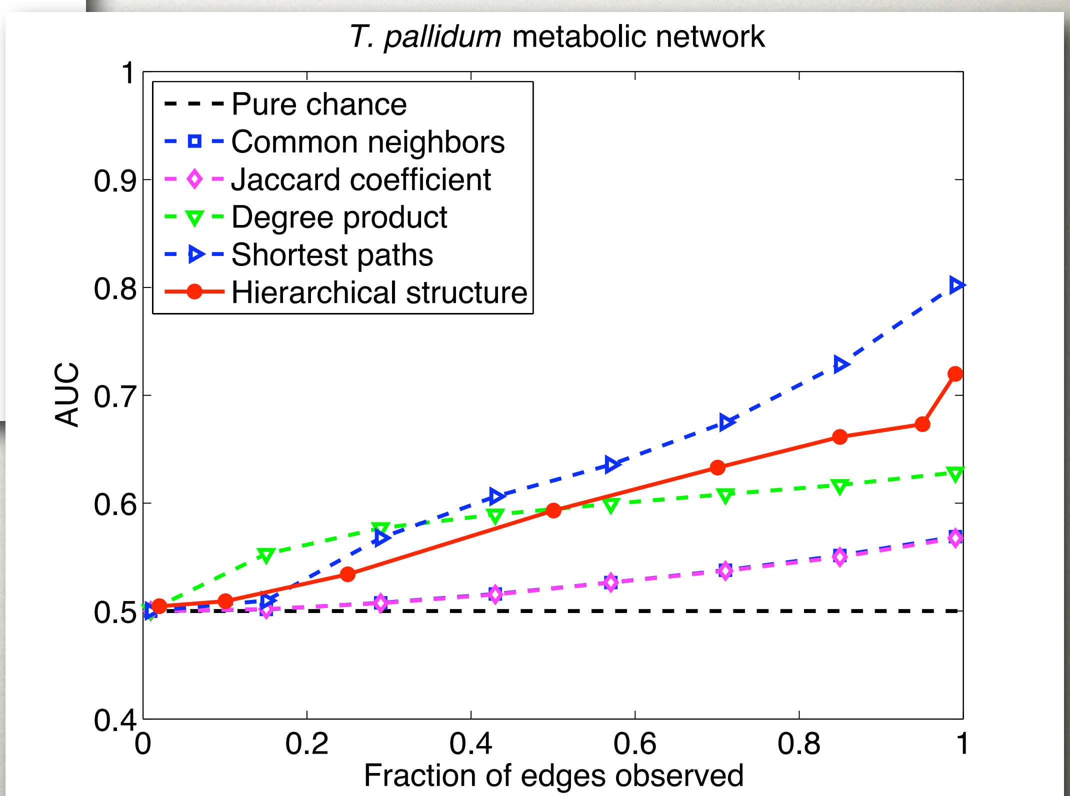
Missing Structure





Other Networks





Summary

- Many real networks are hierarchically modular
- Hierarchies can
  - model multi-scale structure
  - generalize a single network
  - predict missing links
- Model-based inference is very powerful

Acknowledgments:

C. Moore, M.E.J. Newman, C.H. Wiggins, and C.R. Shalizi

Fin

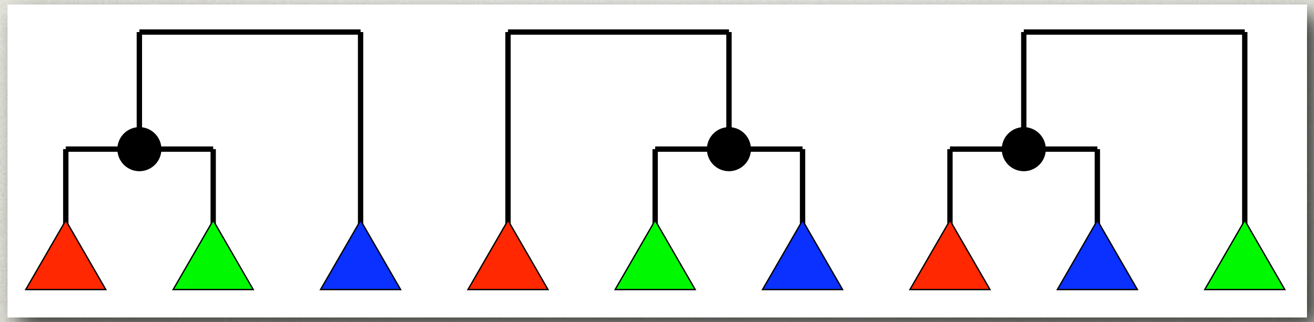
Markov chain Monte Carlo (MCMC)

Given $\mathcal{D}$, choose random internal node

Choose random reconfiguration of subtrees [ergodicity]

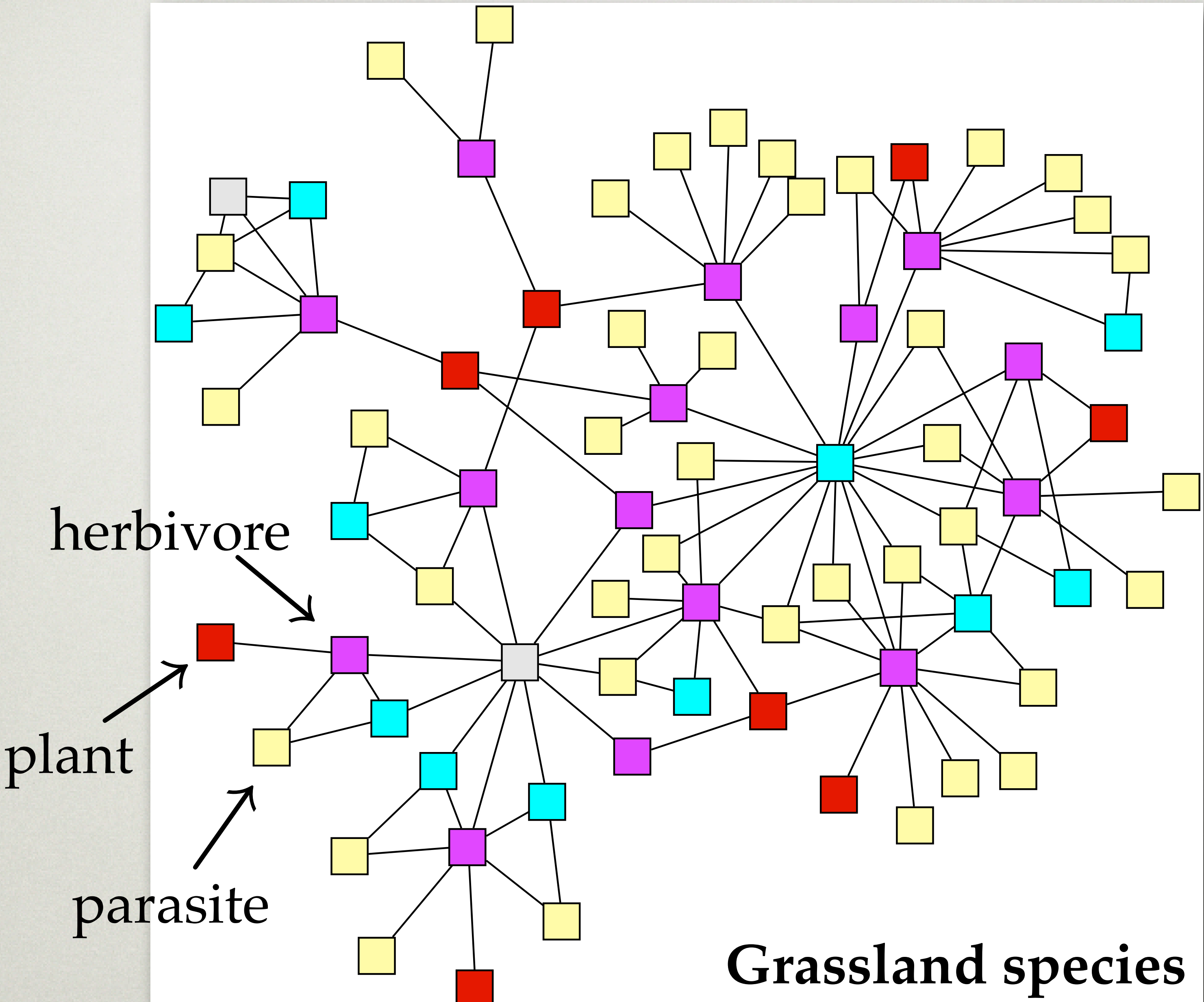
Recompute probabilities $\{p_r\}$ and likelihood $\mathcal{L}$

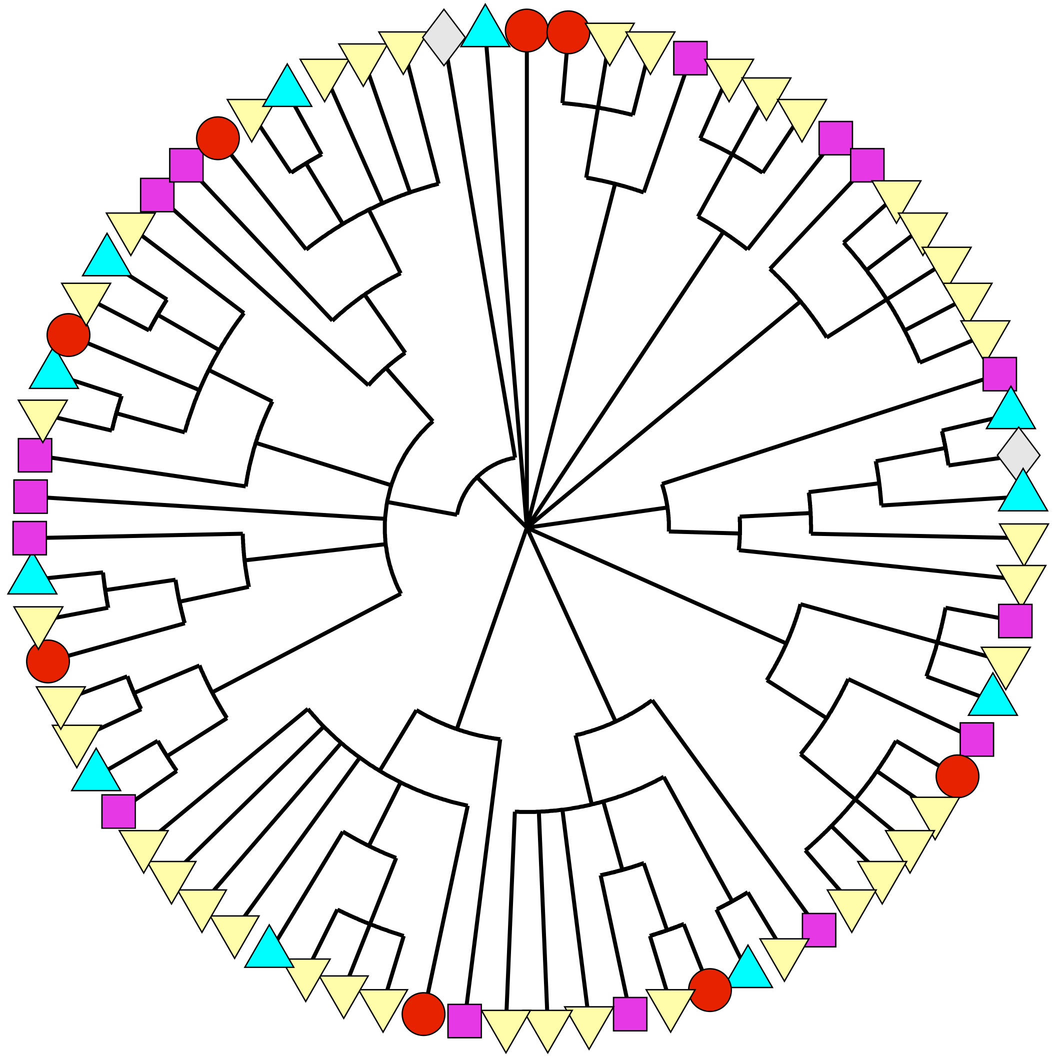
Sampling states according to their likelihood [detailed balance]



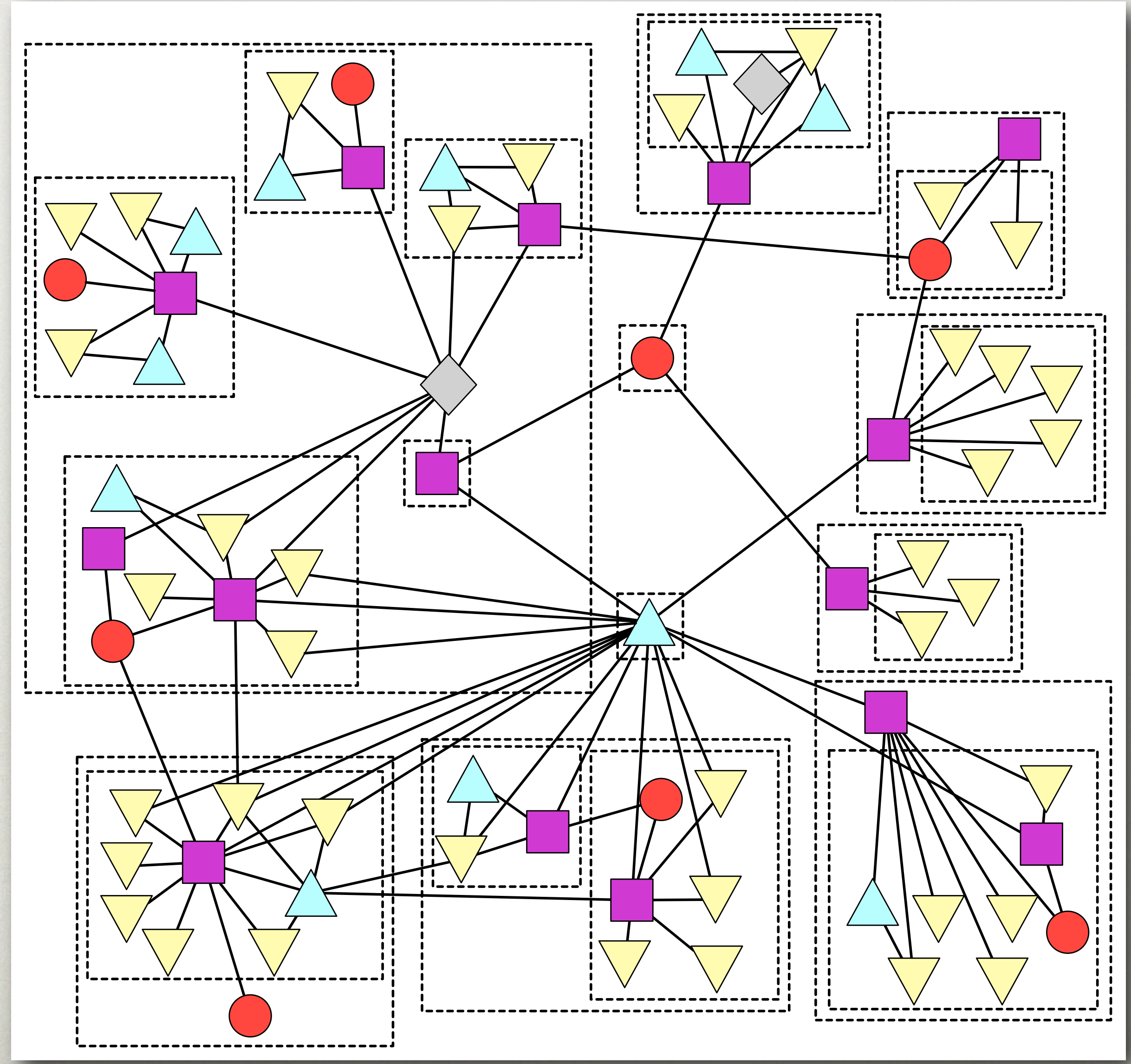
three subtree configurations

(up to relabeling)

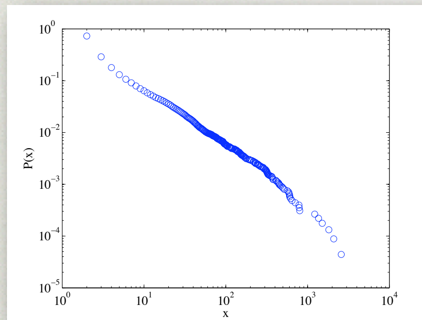

herbivore
plant
parasite
Grassland species



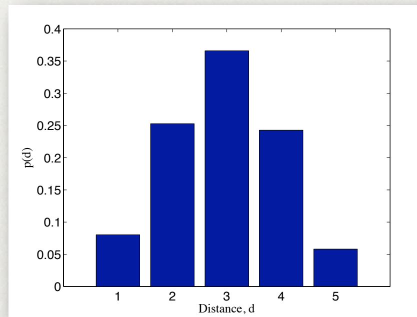
Graph Resampling



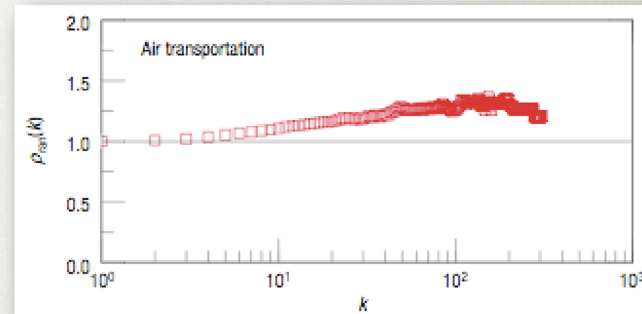
1. Summary Statistics



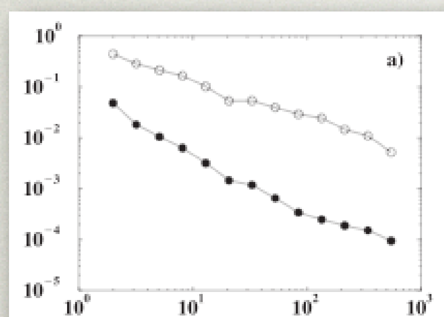
degree distribution



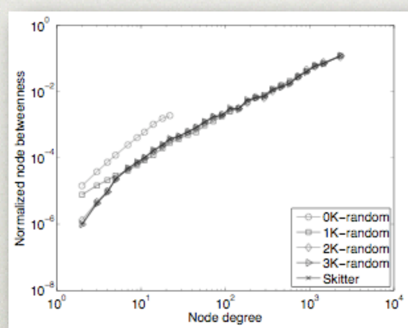
distance distribution



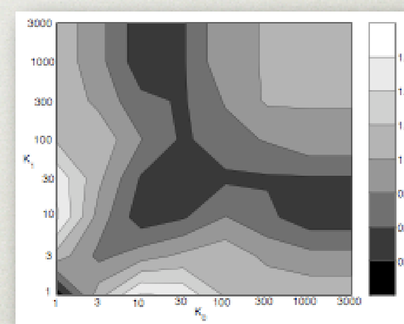
rich-club distribution



short-loop distribution



betweenness function



degree-degree correlations

... etc.

1. Summary Statistics

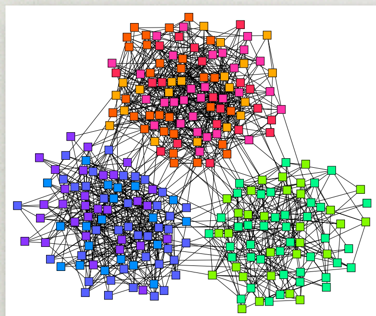
The good

- good for exploratory analysis
- often quick calculations

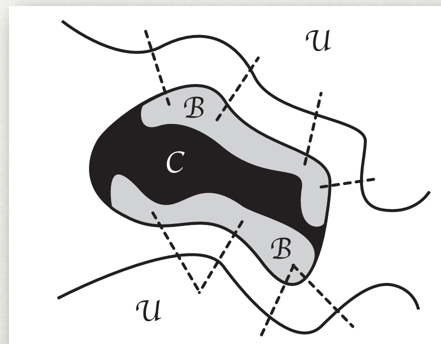
The bad

- throw away important information
- can make different networks **appear** similar
- what are **right** statistics to measure?
- different statistics often highly correlated
- indirect measures of large-scale structure, function

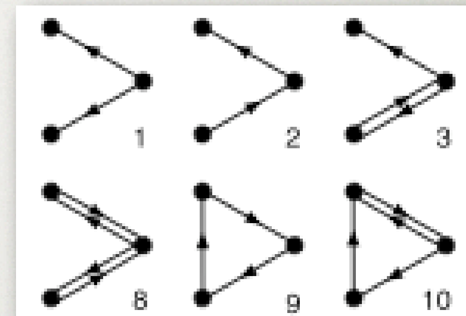
2. Algorithmic Analysis



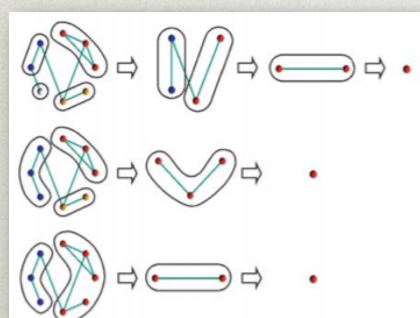
global modularity Q



local modularity R



network motifs



box covering

clique covering

... etc.

2. Algorithmic Analysis

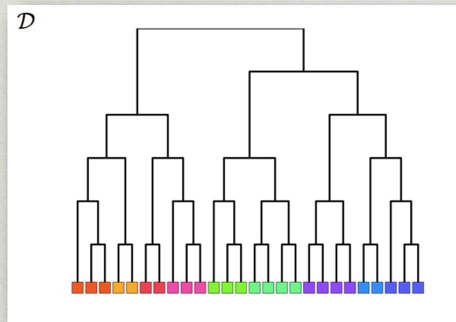
The good

- good for exploratory analysis
- illustrate large-scale structure, heterogeneity

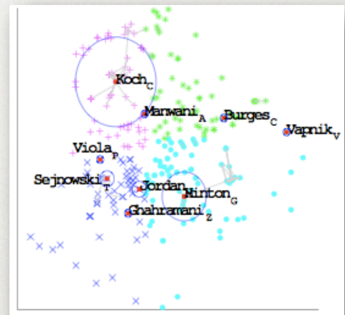
The bad

- often (NP-)hard optimizations
- can be sensitive to noise, uncertainty
- *ad hoc* or heuristic measures of structure, function
  - algorithm = theory
  - implied physics often unclear

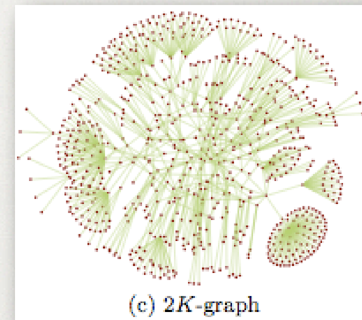
3. Statistical Inference



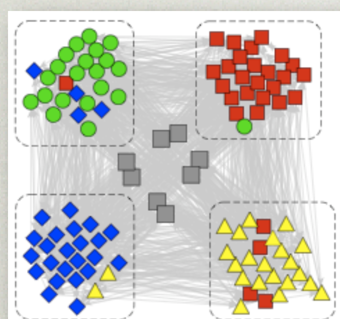
hierarchical random graphs



latent space models



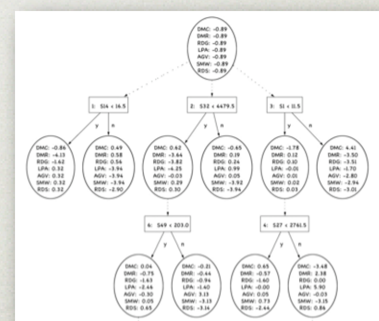
correlation reconstruction



community mixtures

$$I(X;Y) = H(X) - H(X|Y)$$

information bottlenecks



network classification

3. Statistical Inference

The good

- model-based measures of structure
- concrete, testable predictions
- better robustness to noise, uncertainty
- well-grounded in computer science, statistics

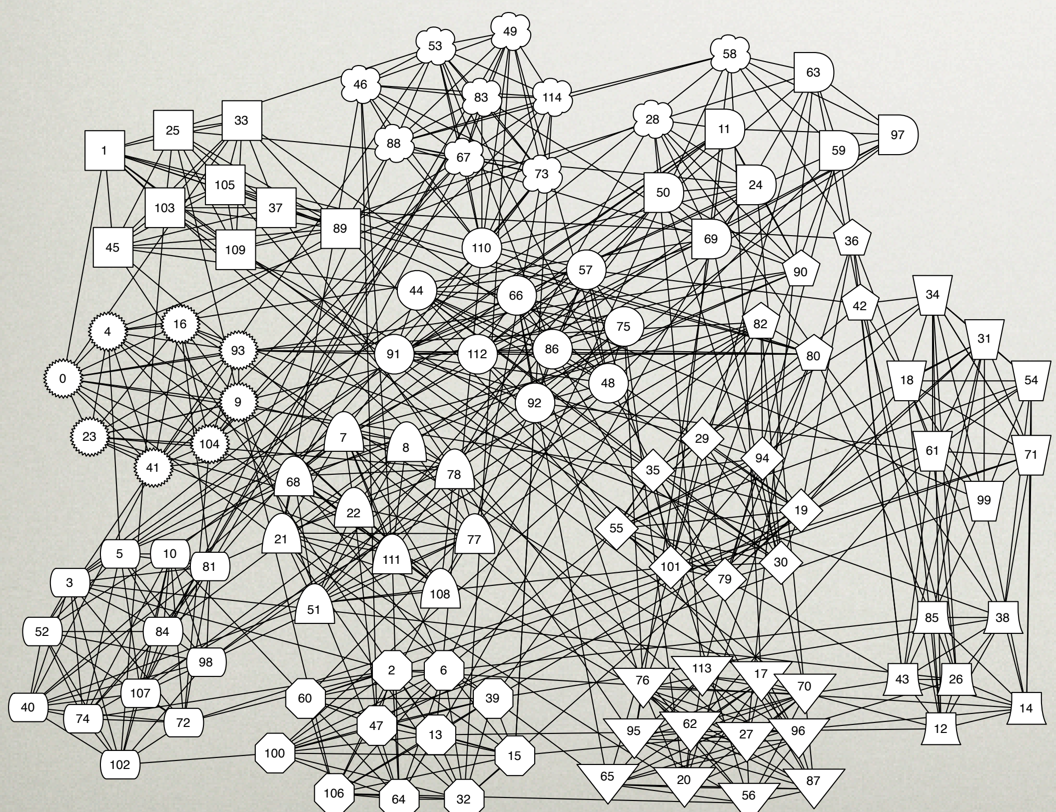
The bad

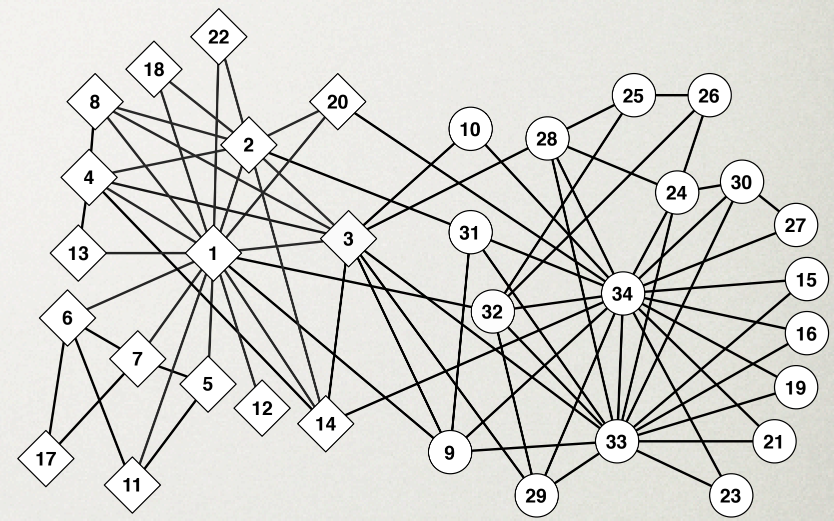
- models must be explicit, precise
- often hard computations
- data intensive

Two Case Studies

NCAA Schedule 2000

$n = 115 \quad m = 613$





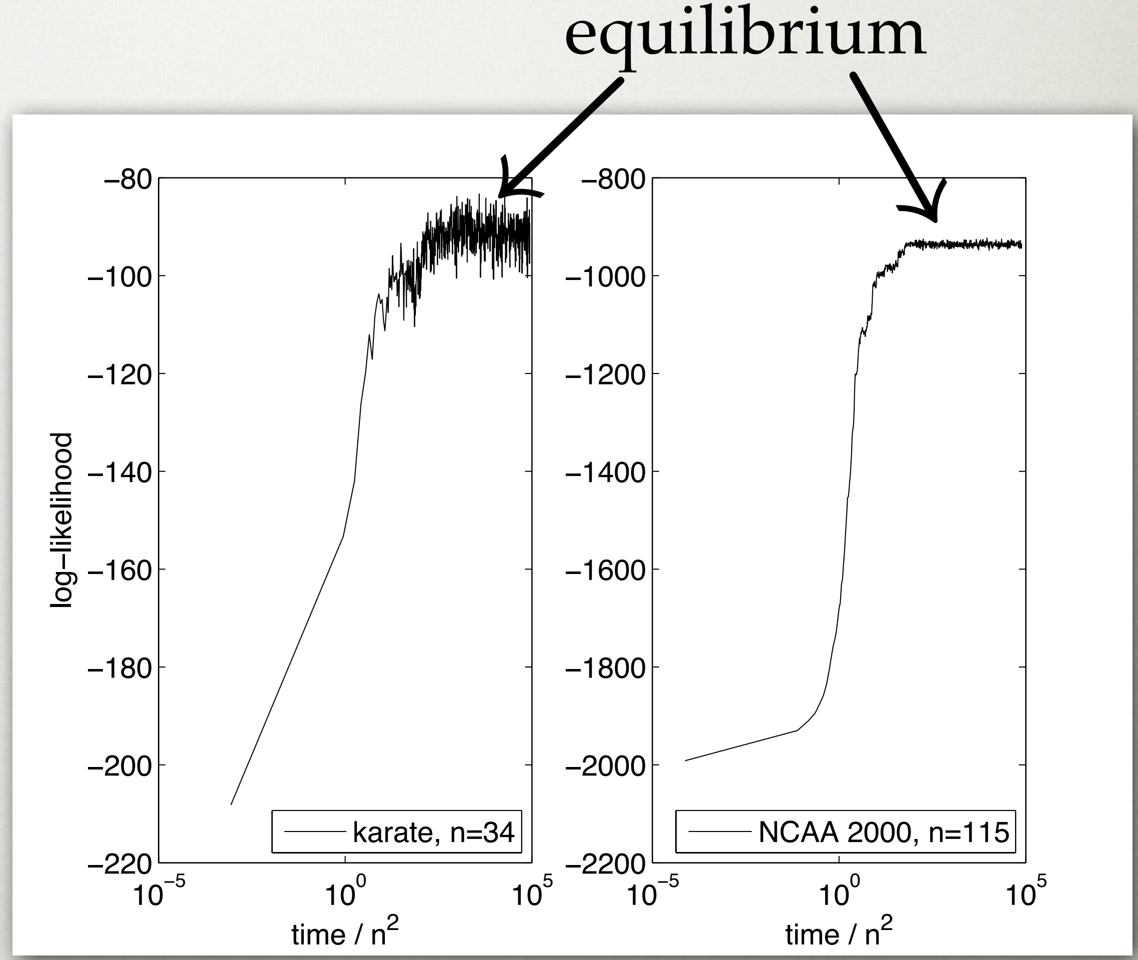
Zachary's Karate Club

$n = 34 \quad m = 78$

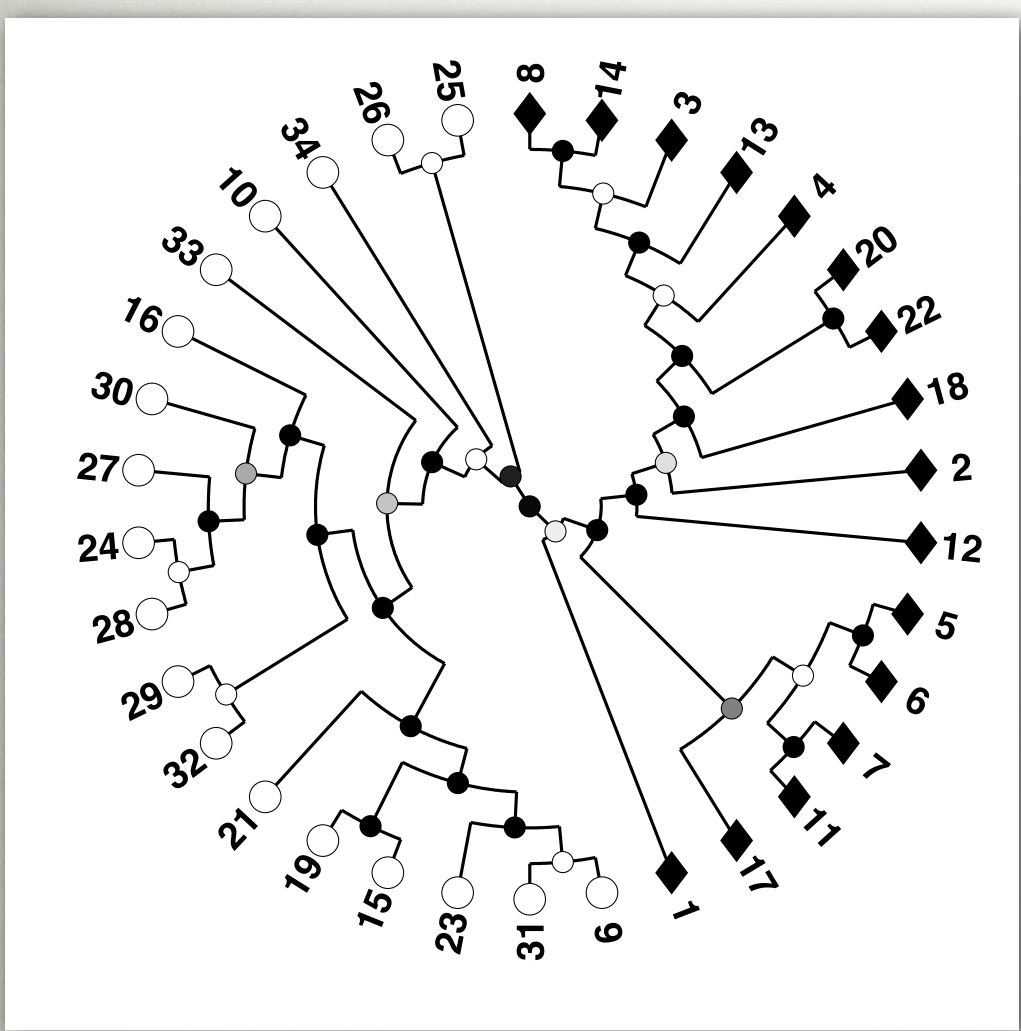
Mixing Times

MCMC mixes relatively quickly

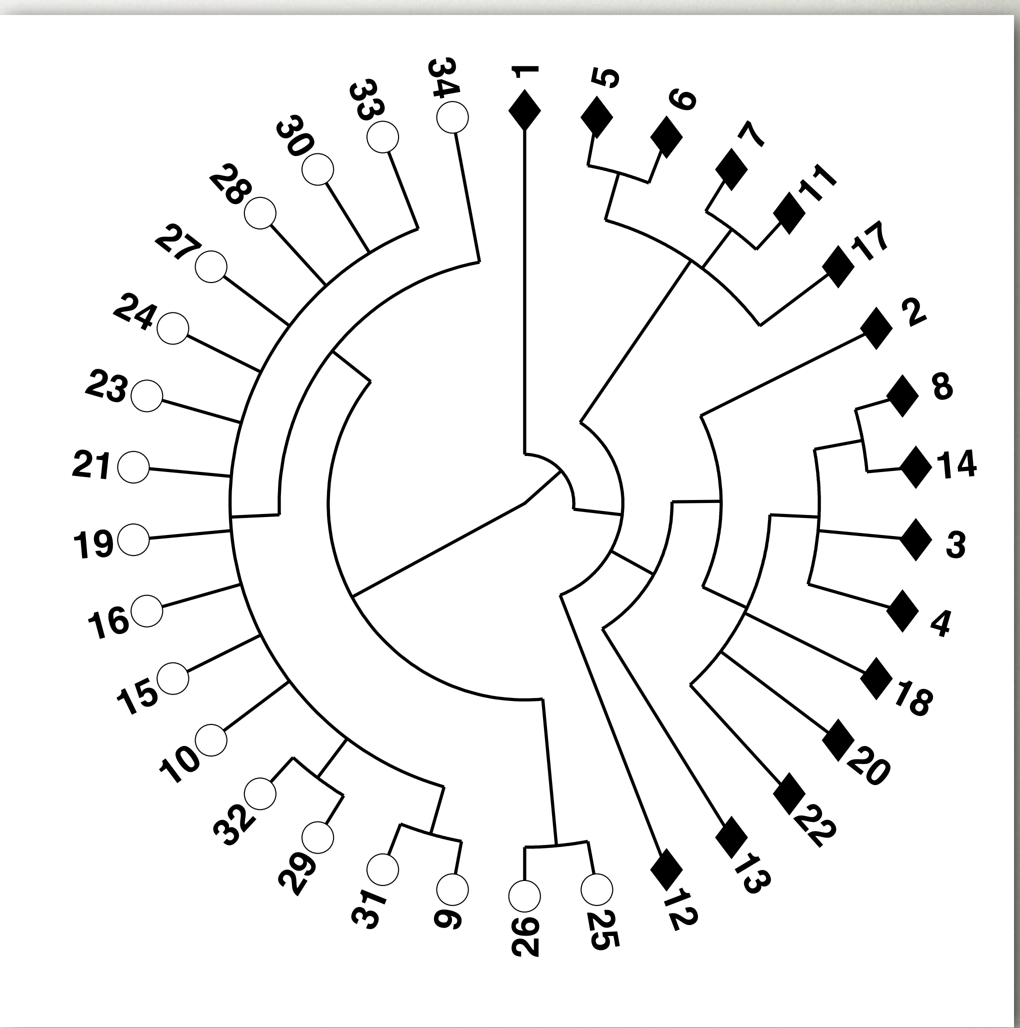
Equilibrium in $O(n^2)$ steps



Hierarchies

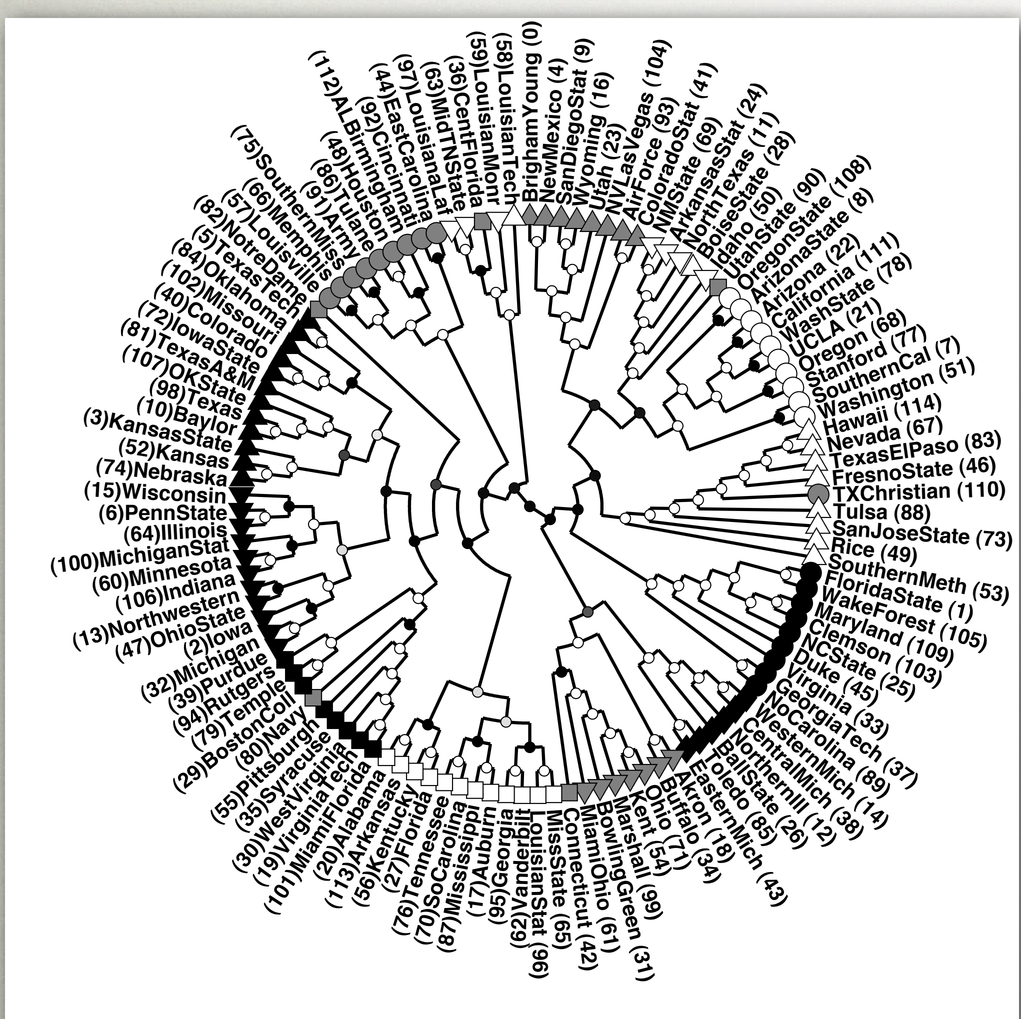


point estimate

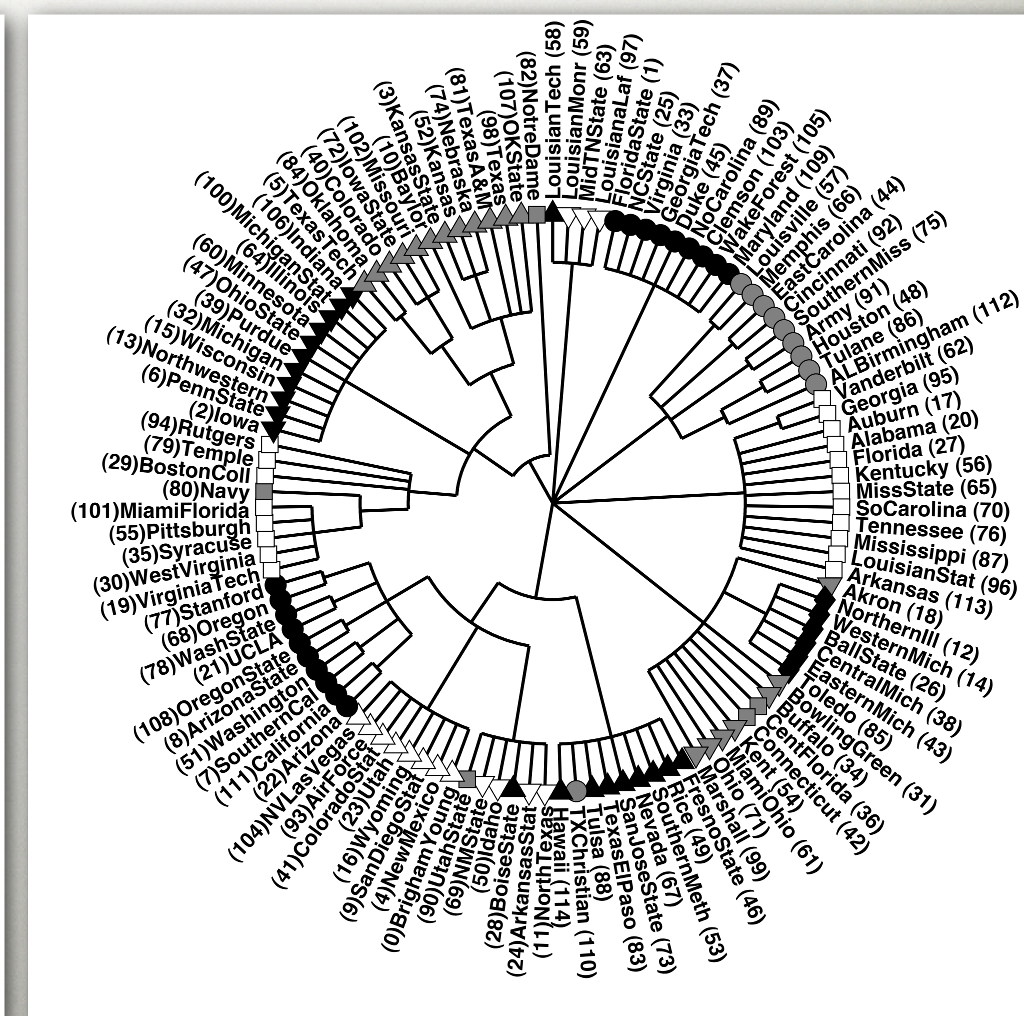


consensus hierarchy

Hierarchies



point estimate



consensus hierarchy