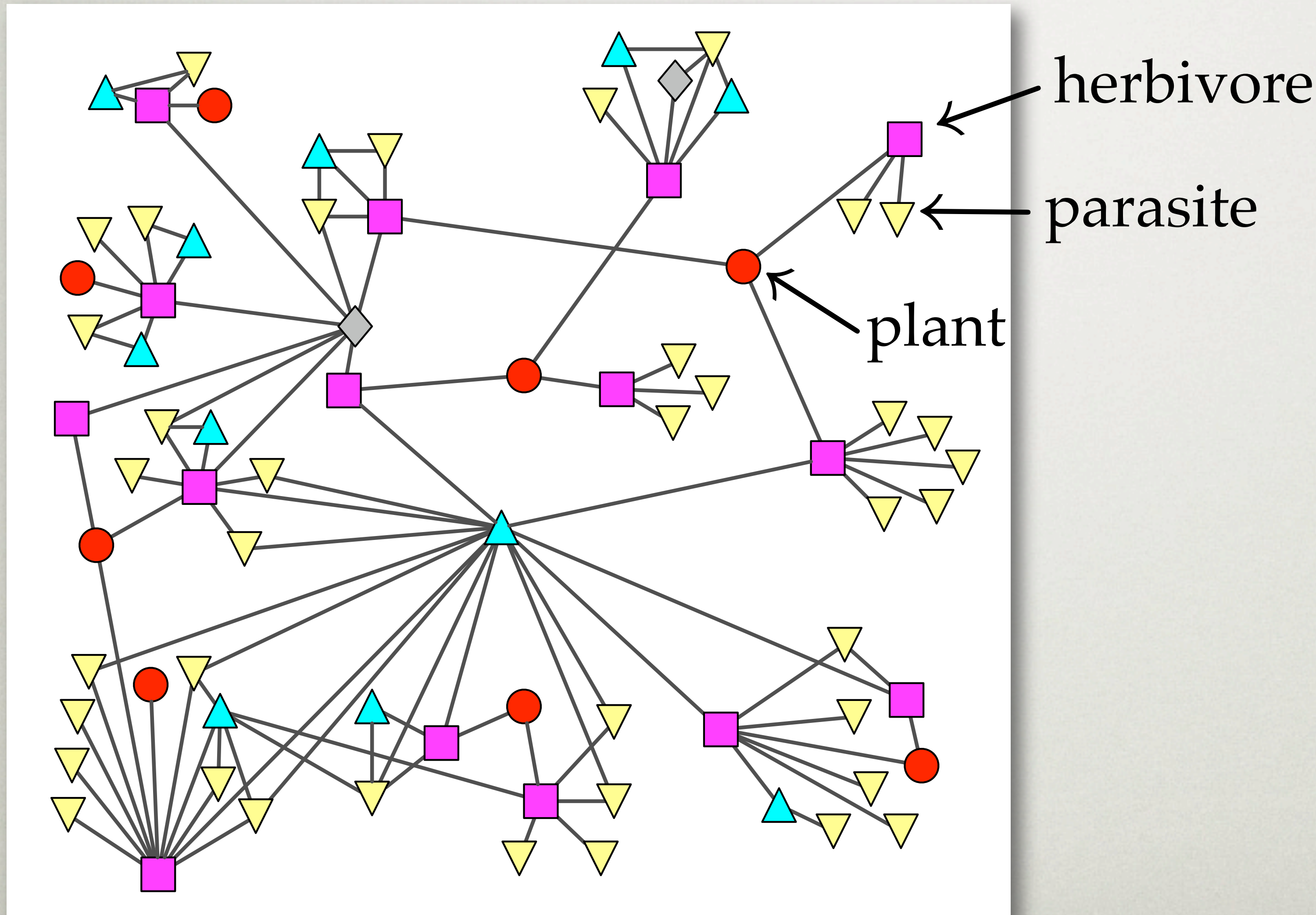
Hierarchically Modular Structure in Complex Networks

Aaron Clauset
Santa Fe Institute

3 December 2008
SFI Workshop
"Statistical Inference for Complex Networks"

joint with C. Moore, M.E.J. Newman, T.A.S. Pierce

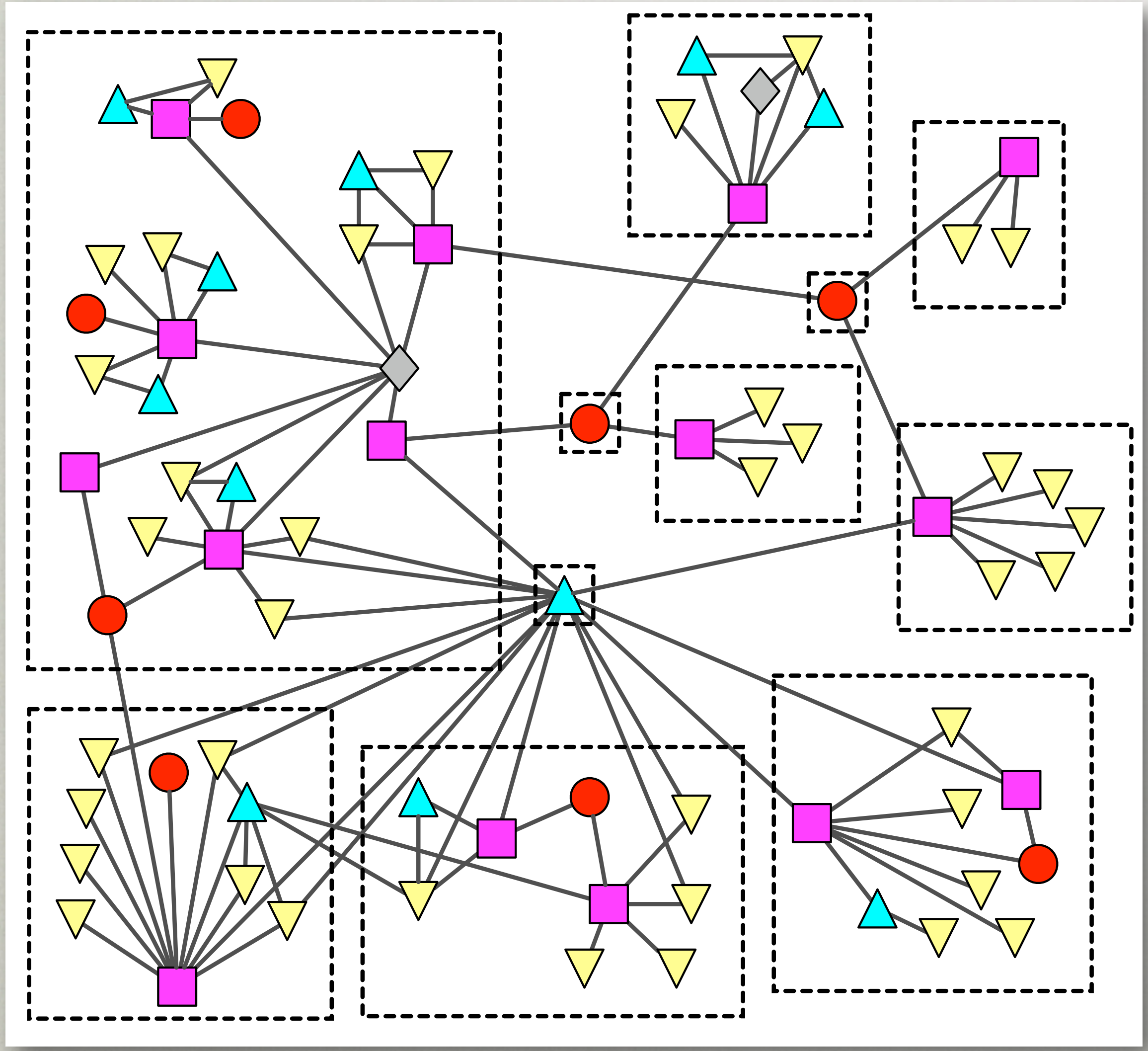
Hierarchies of Modules



Grassland species*

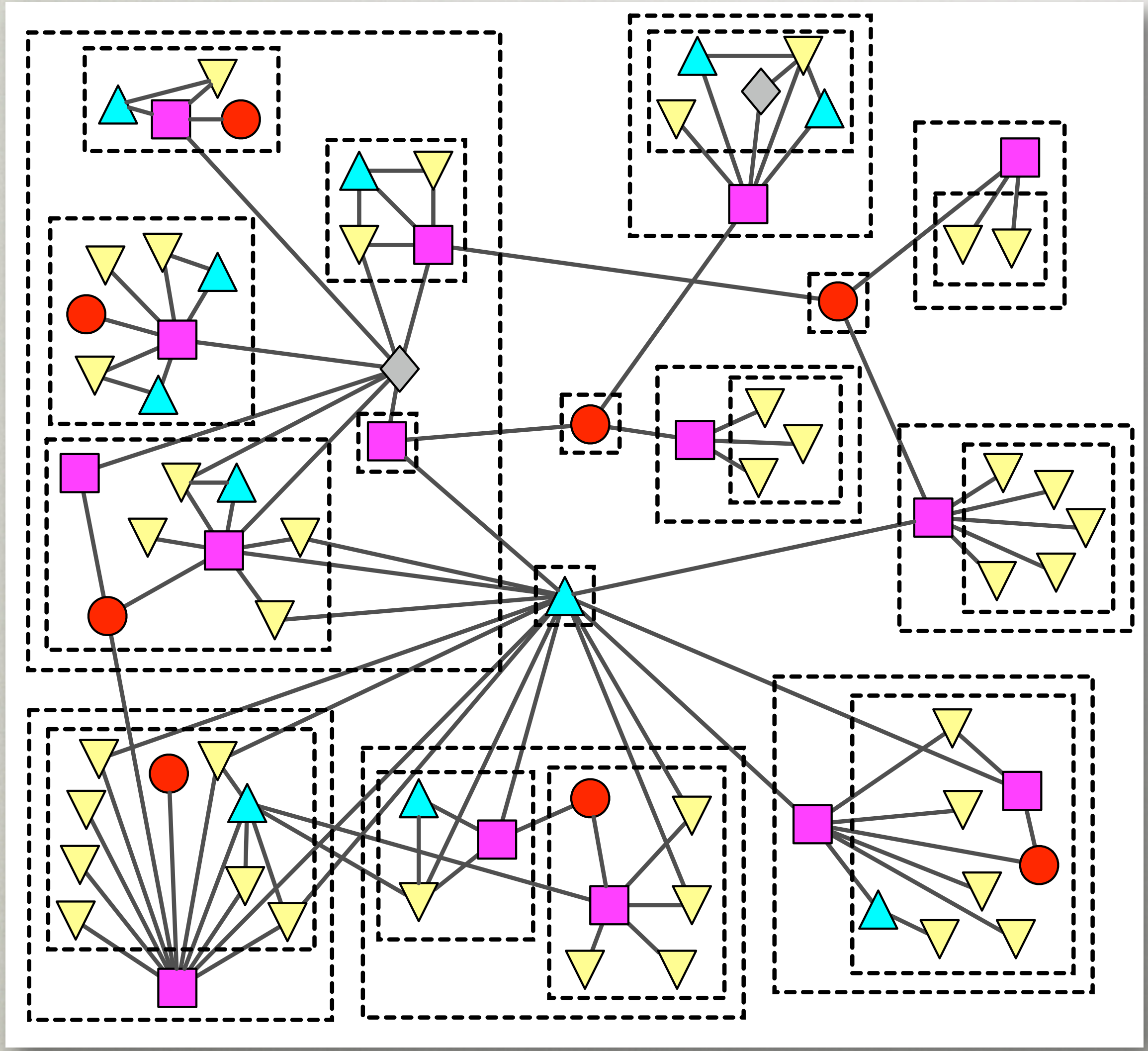
*thank you: Jennifer Dunne

Hierarchies of Modules



one level

Hierarchies of Modules



many levels

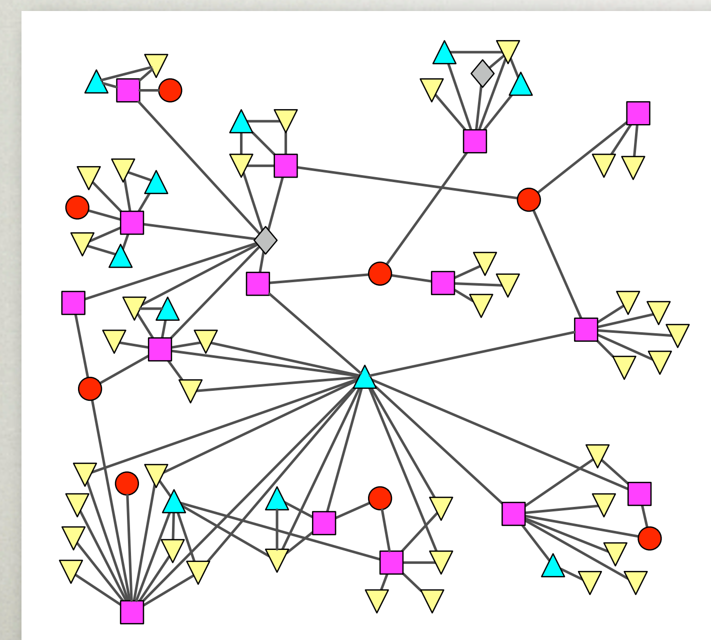
The Task

How can we extract

- **this hierarchical (multi-scale) structure**

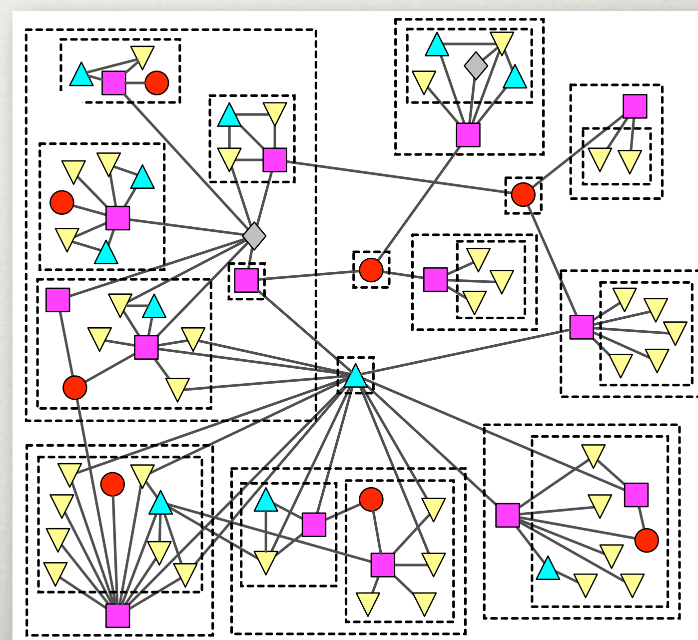
from complex networks?

network data



$\xrightarrow{?}$

hierarchical structure



One Approach

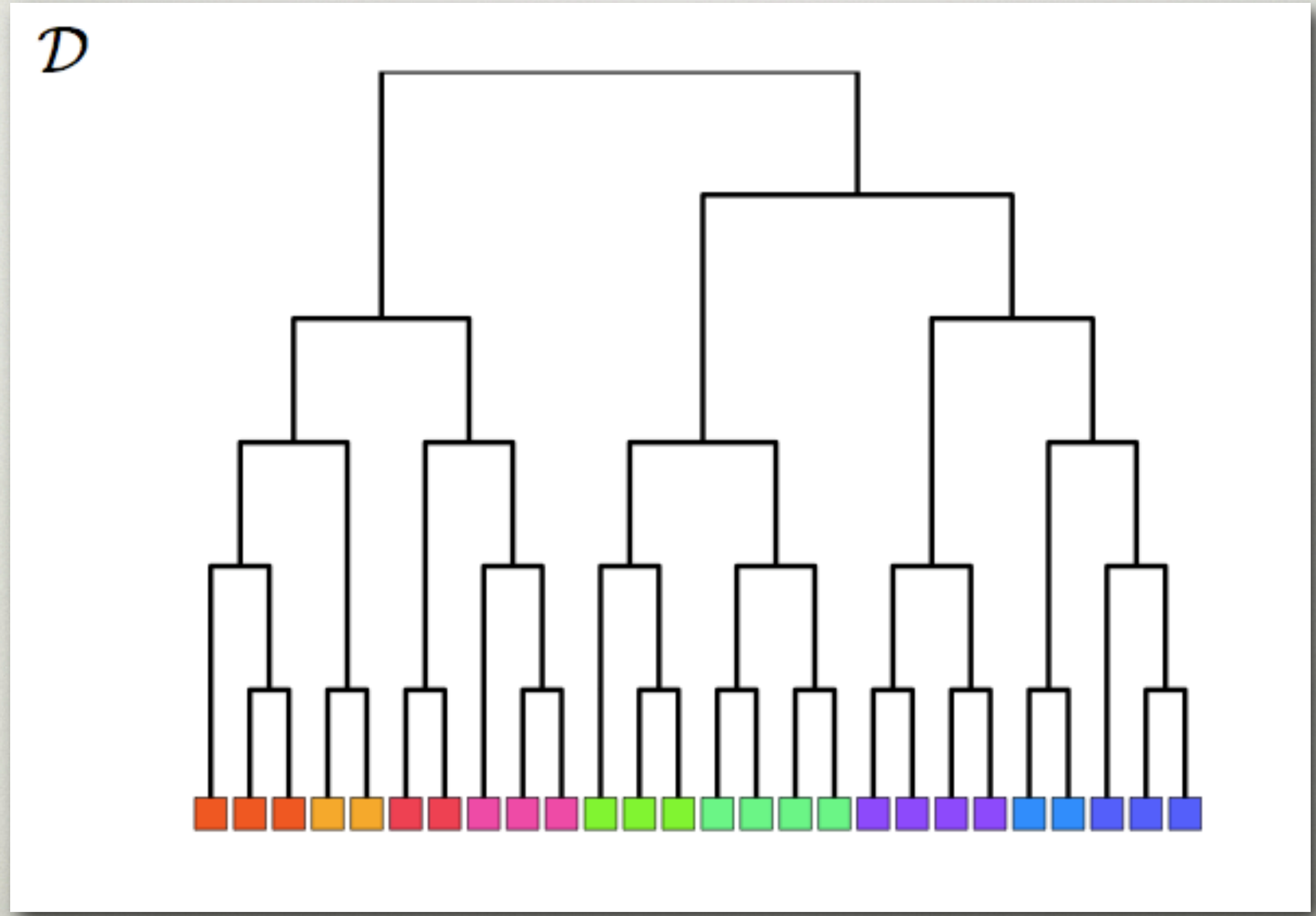
Model-based inference

1. generative model of hierarchies (a model)

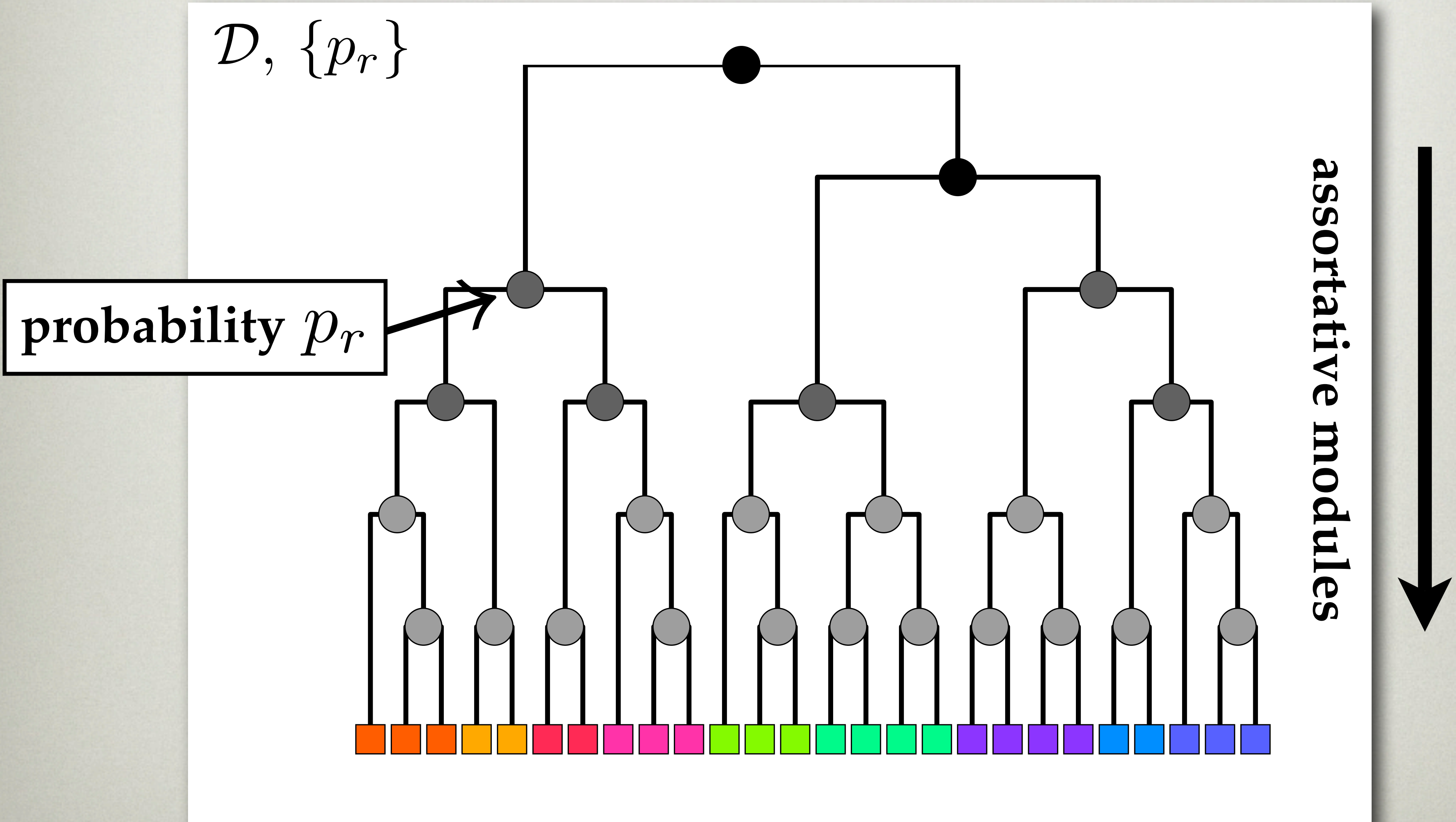
2. find “good” instances of this model

3. predict missing information

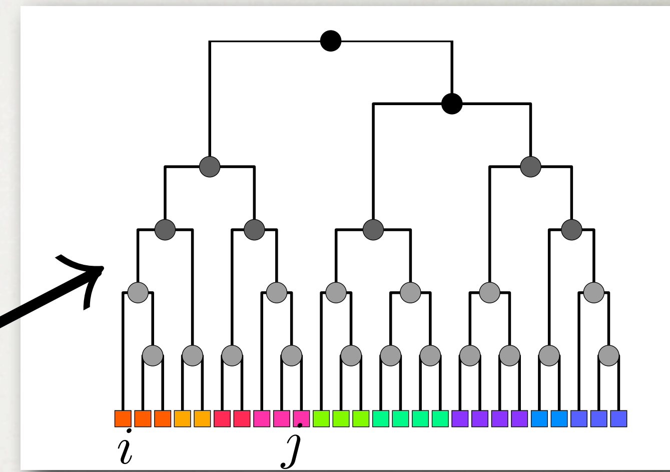
A Model of Hierarchy



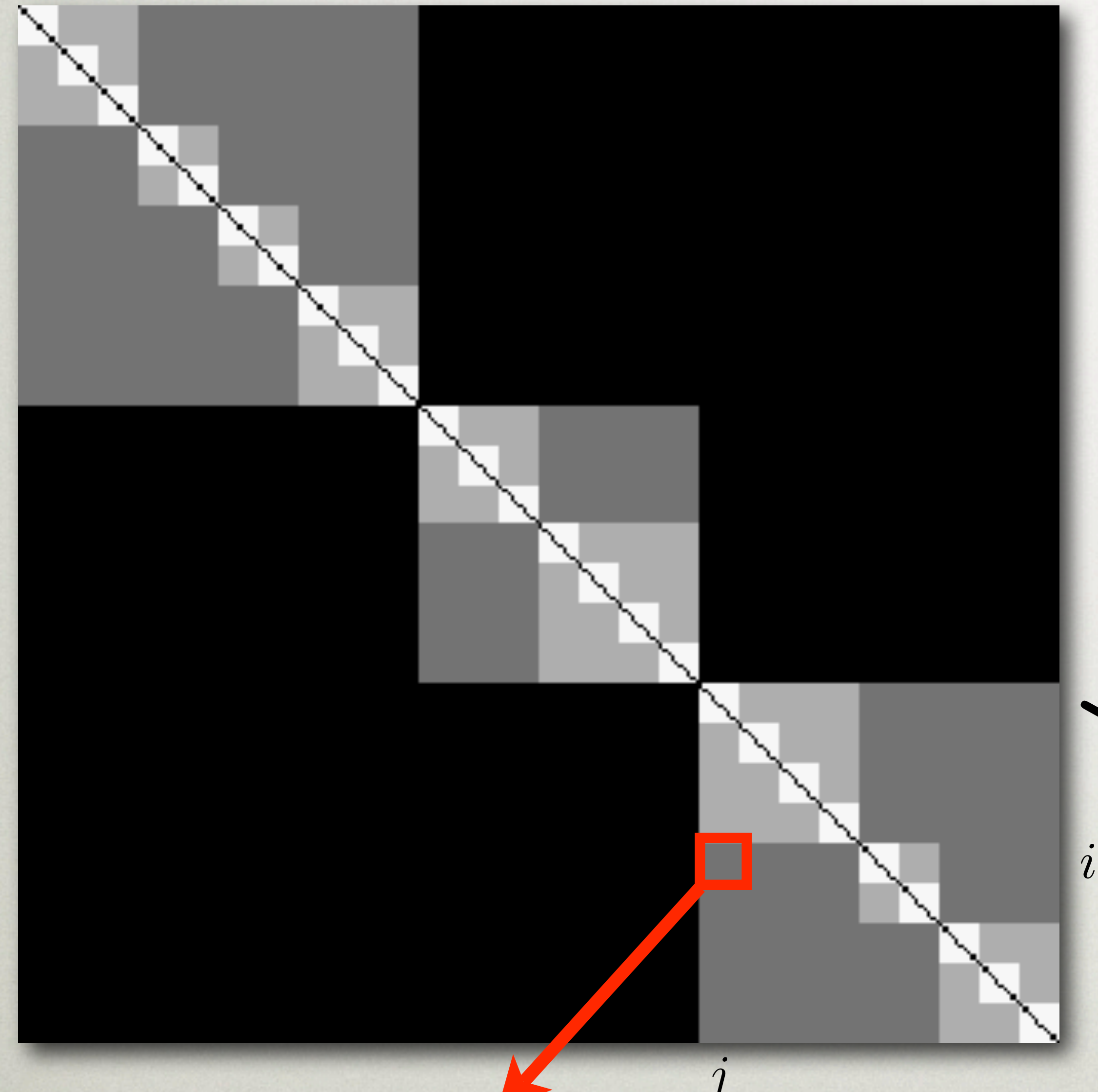
A Model of Hierarchy



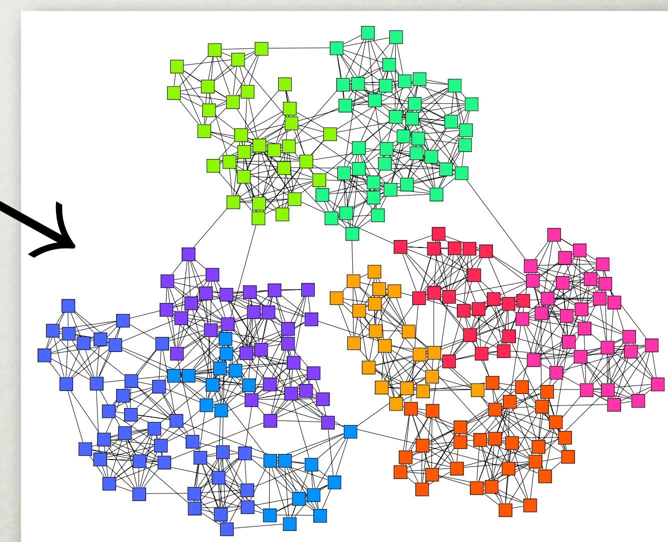
model



“inhomogeneous” random graph



instance



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

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

a direct approach

- **likelihood function** $\mathcal{L} = \Pr(\text{ data} \mid \text{model })$

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

- **sample** the **good** models

  via Markov chain Monte Carlo

- technical details in
  Nature **453** (2008) and *physics/0610051*

Missing Links

A test: can model predict missing links?

Guessing is Bad

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

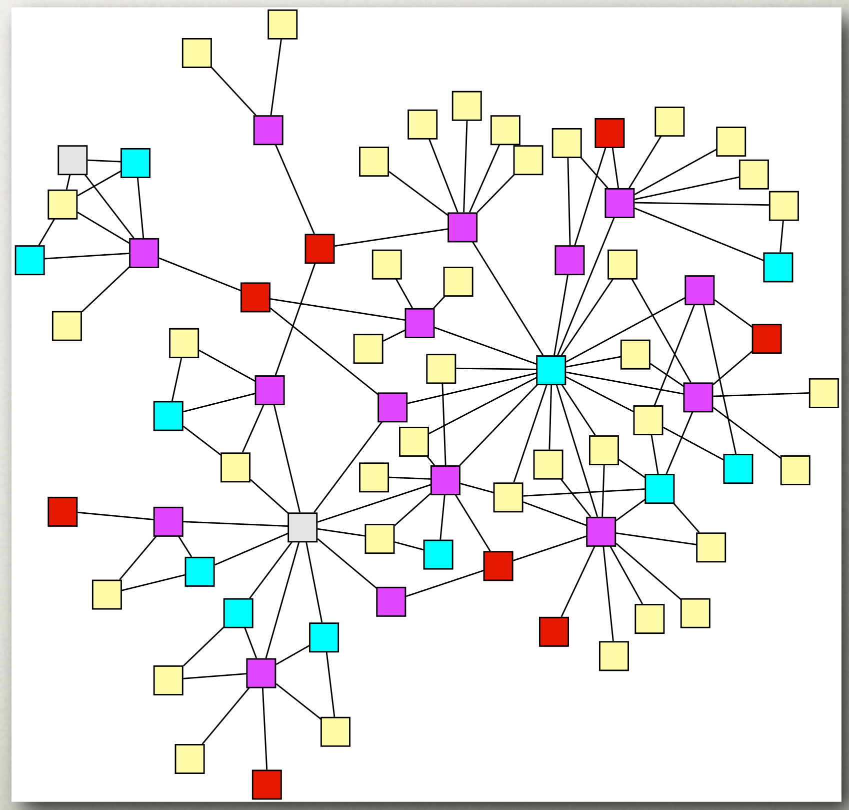
$$p_{\text{guess}} \approx \frac{k}{n^2 - m + k}$$

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

$$n = 75$$

$$m = 113$$

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



Predicting Missing Links

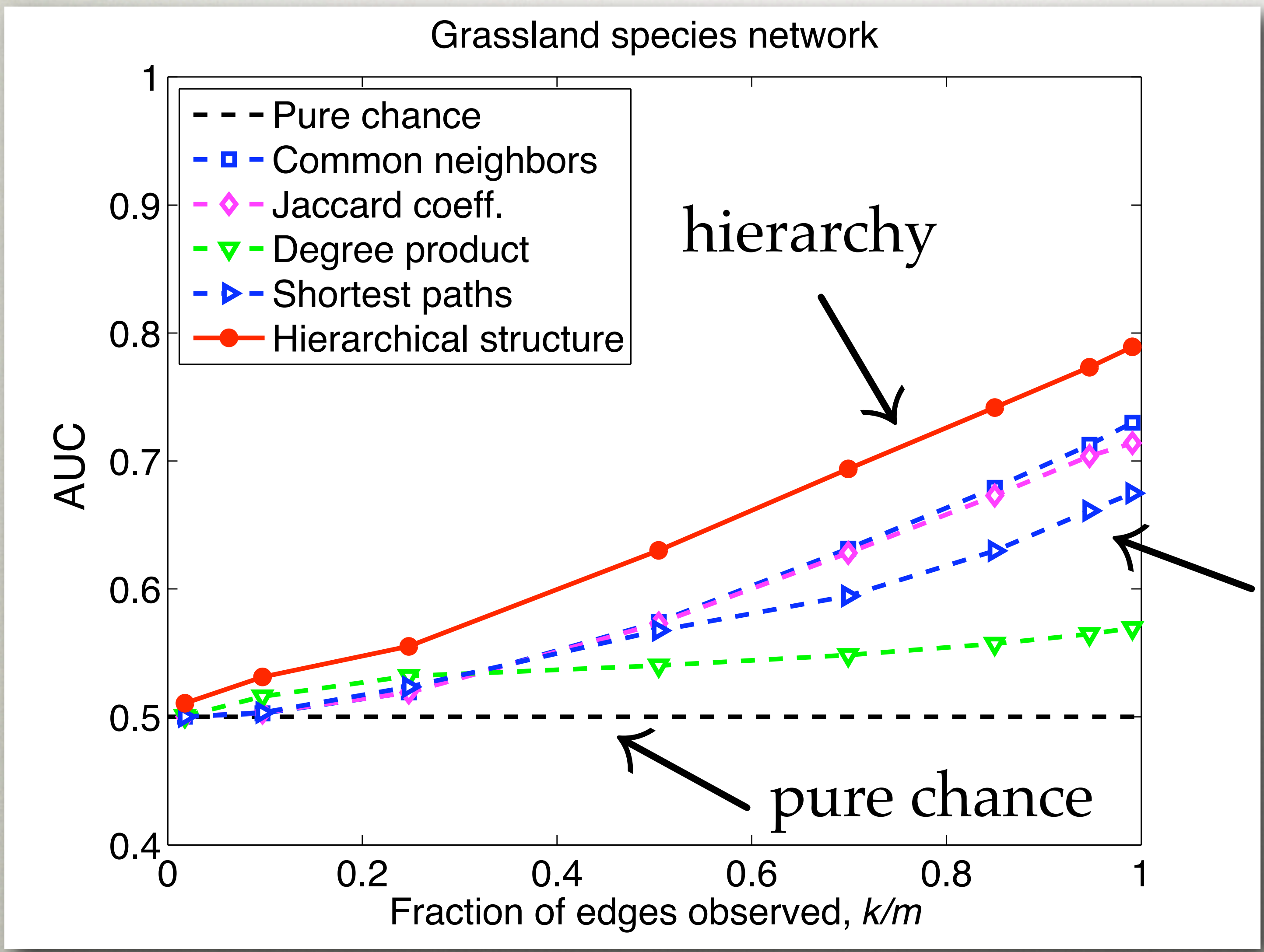
- Given incomplete graph G
- run MCMC to equilibrium
- then, over sampled $\mathcal{D}$, compute average $\langle p_r \rangle$ for links $(i, j) \notin G$
- predict links with high $\langle p_r \rangle$ values are missing

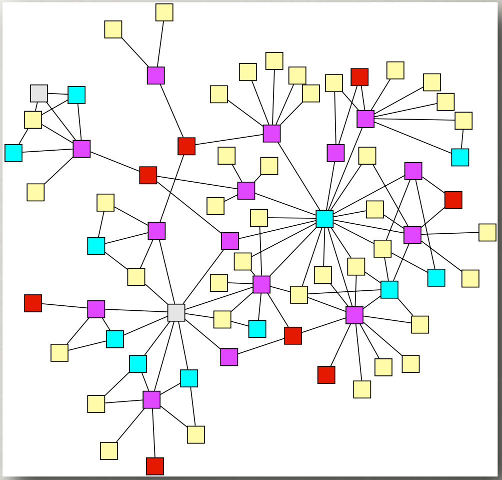
Test idea via leave-*k*-out cross-validation

perfect accuracy: AUC = 1

no better than chance: AUC = 1/2

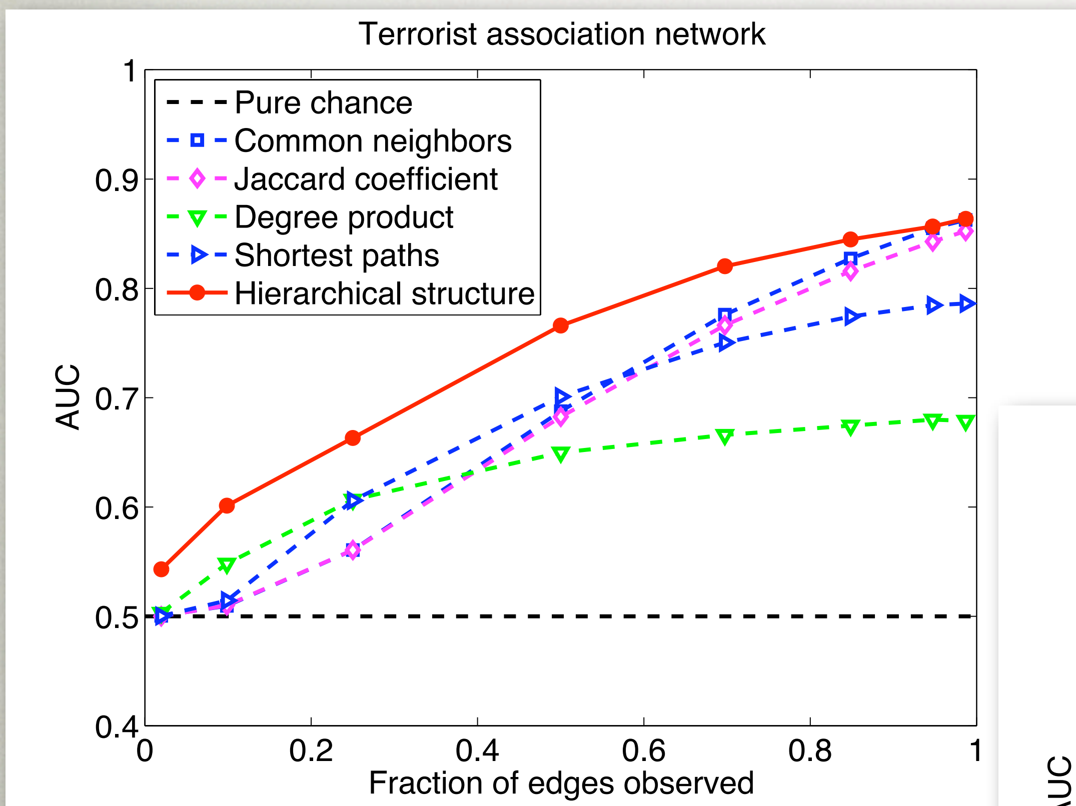
Missing Structure

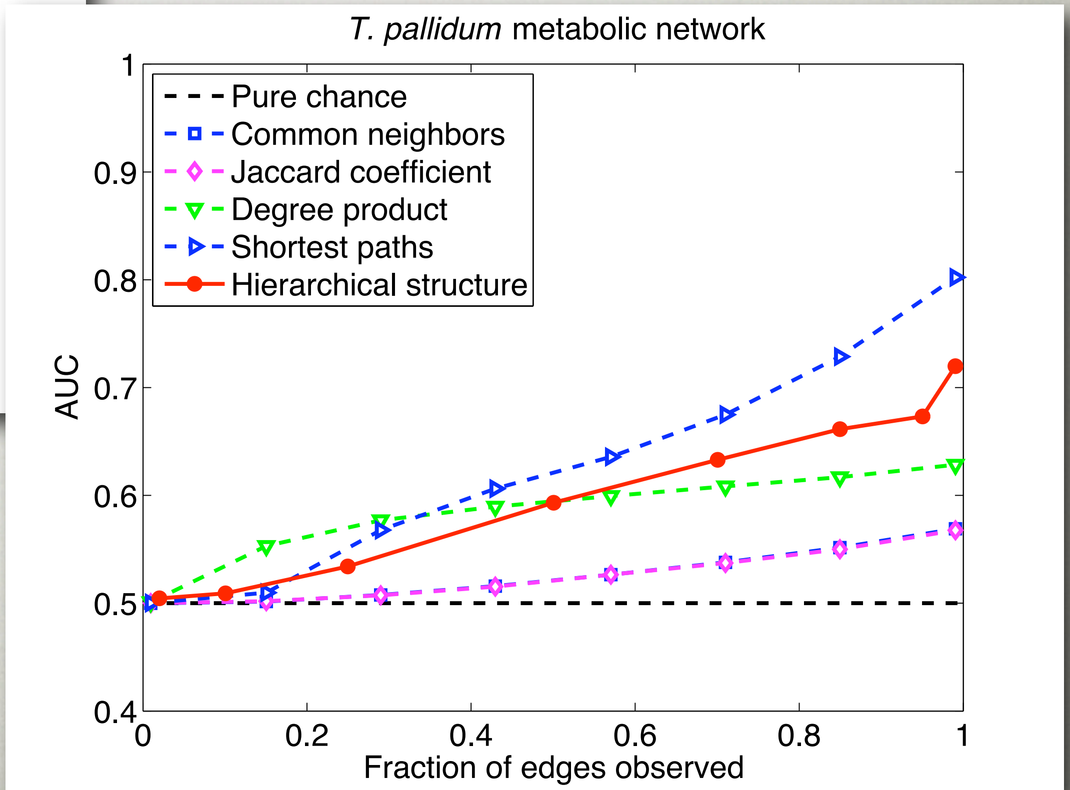




simple predictors

Other Networks





Other Data

- node labels, attributes, weights
- edge labels, attributes, weights
- geographical structure
- temporal information
- combinations of these

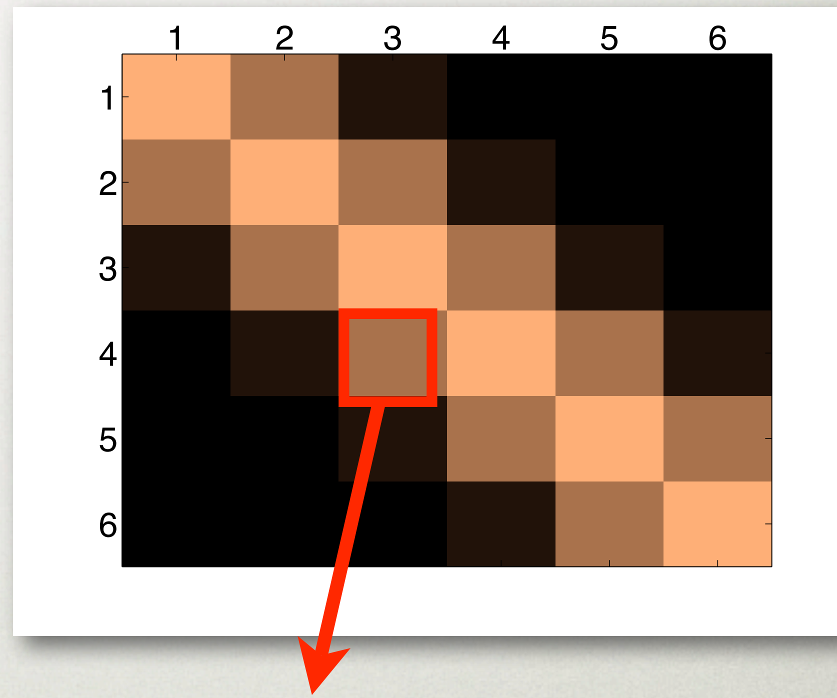
Node labels only:

a simple ("stochastic block") model can do well

Stochastic Block Model

- k types of nodes
- $k \times k$ matrix $p_{i,j}$ of module connectivity
- no assumptions on structure of $p_{i,j}$
- we use MCMC to search over all node labelings
- can we predict *missing* labels?

$k = 6$ types

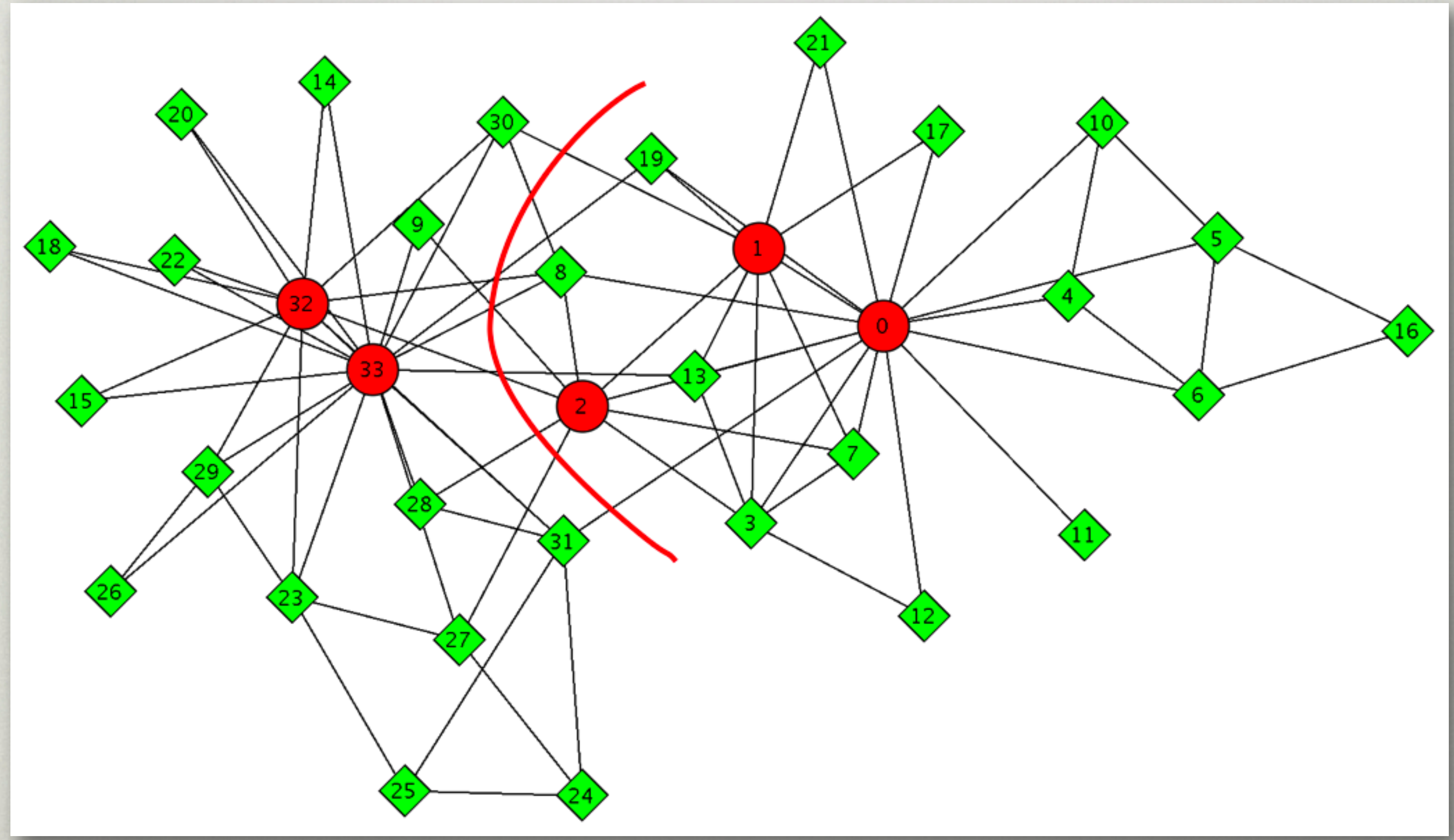


assortative modules

$\Pr(\text{node in } i \text{ connected to node in } j) = p_{i,j}$

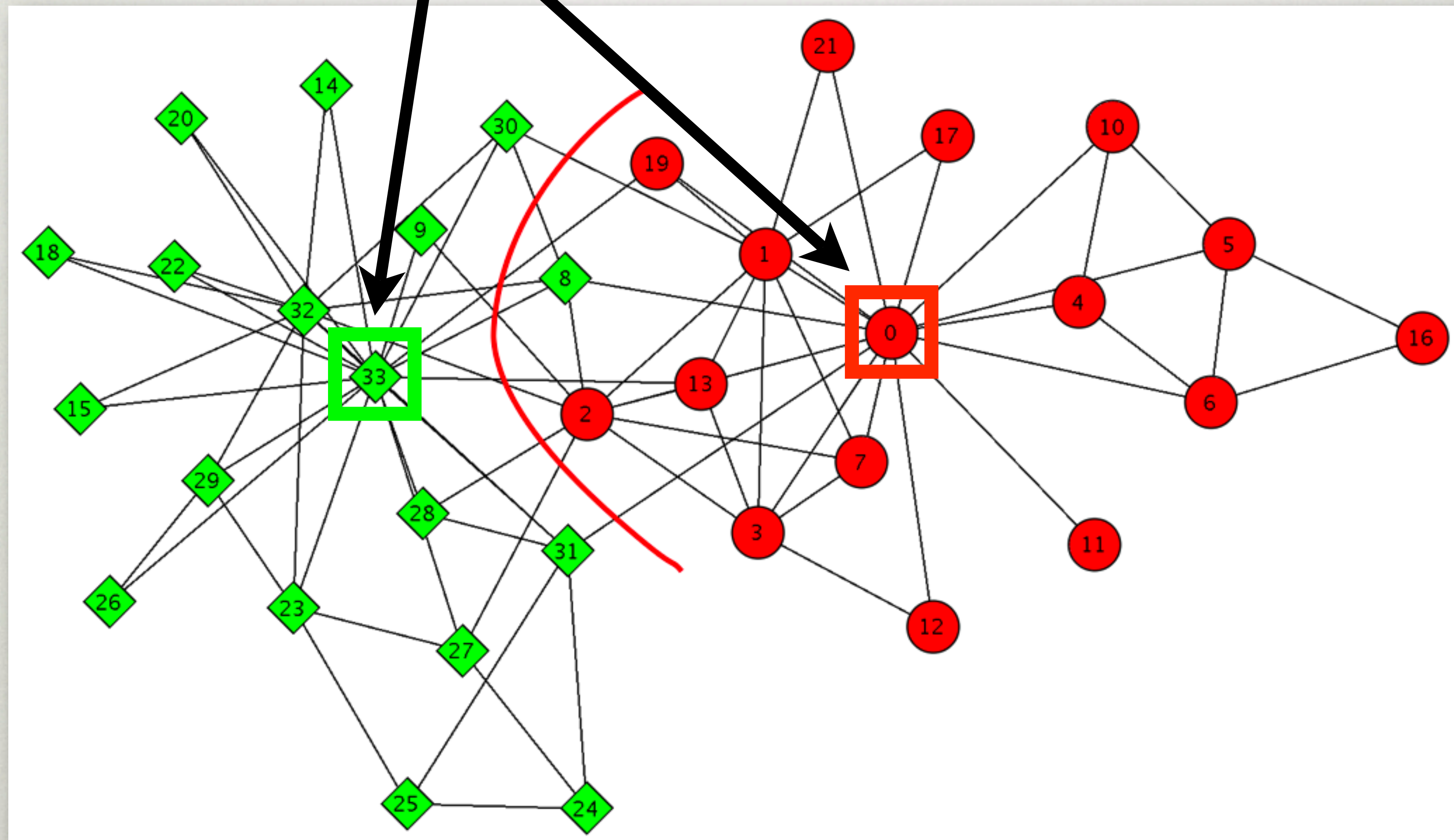
Unconstrained Inference

let all labels vary



Constrained Inference

fix these labels; let the others vary



Summary

- generative models of
  - hierarchical modules
  - simple modules
  - potentially many others
- predict missing information (edges, types)
- principled way to fit and test structural theories

Acknowledgments:

C. Moore, M.E.J. Newman, T.A.S. Pierce,
C.H. Wiggins, C.R. Shalizi

Fin