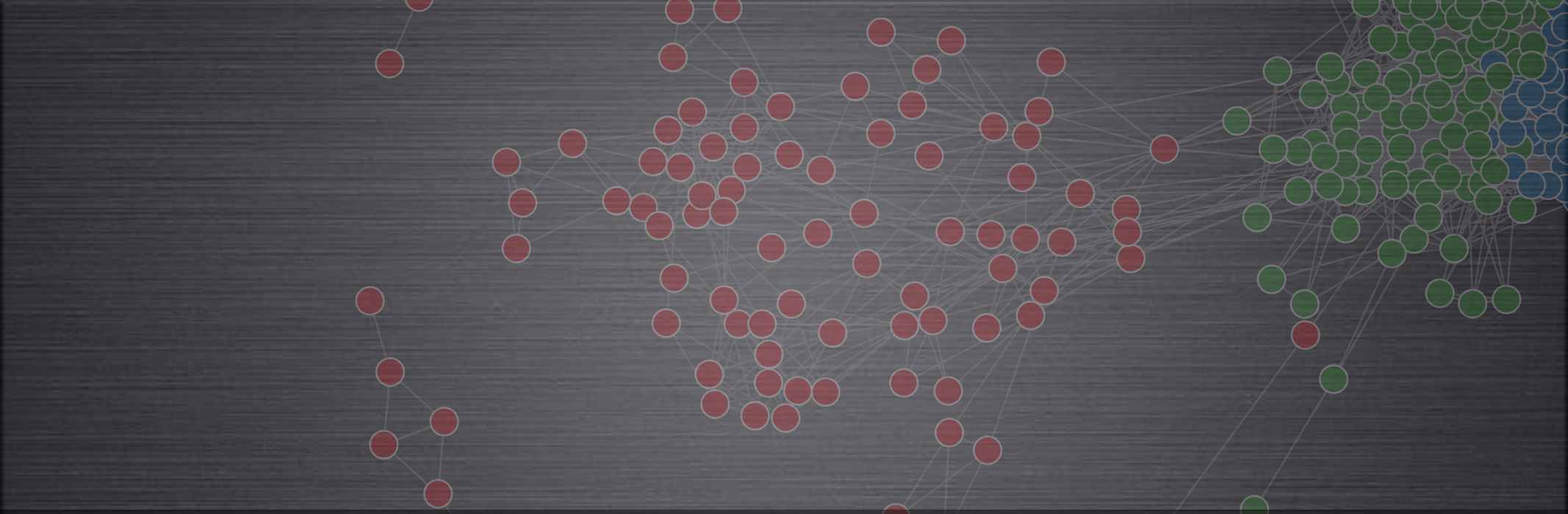


Lecture 8: Generalized large-scale structure

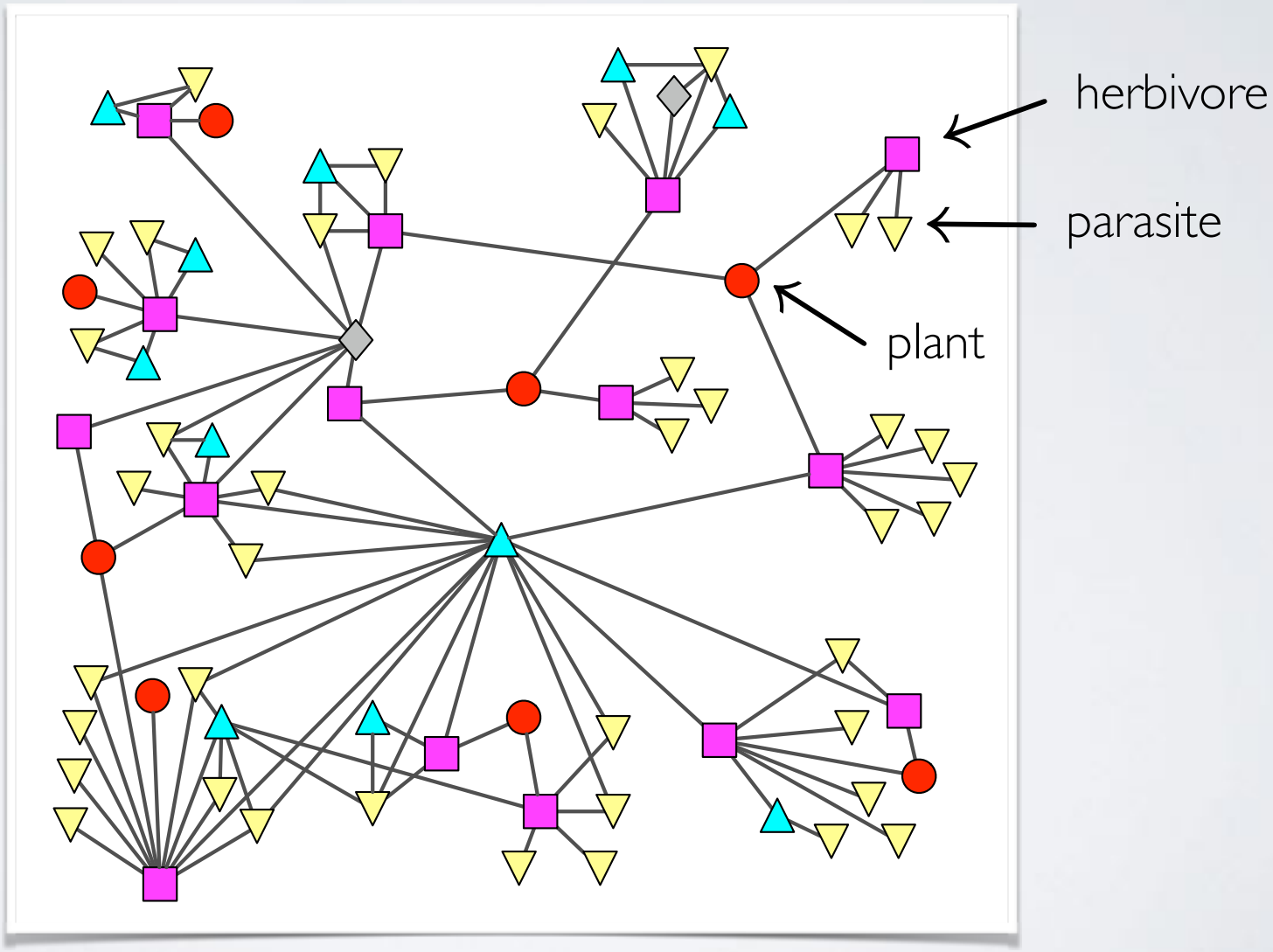
Aaron Clauset
@aaronclauset
Assistant Professor of Computer Science
University of Colorado Boulder
External Faculty, Santa Fe Institute

hierarchical communities

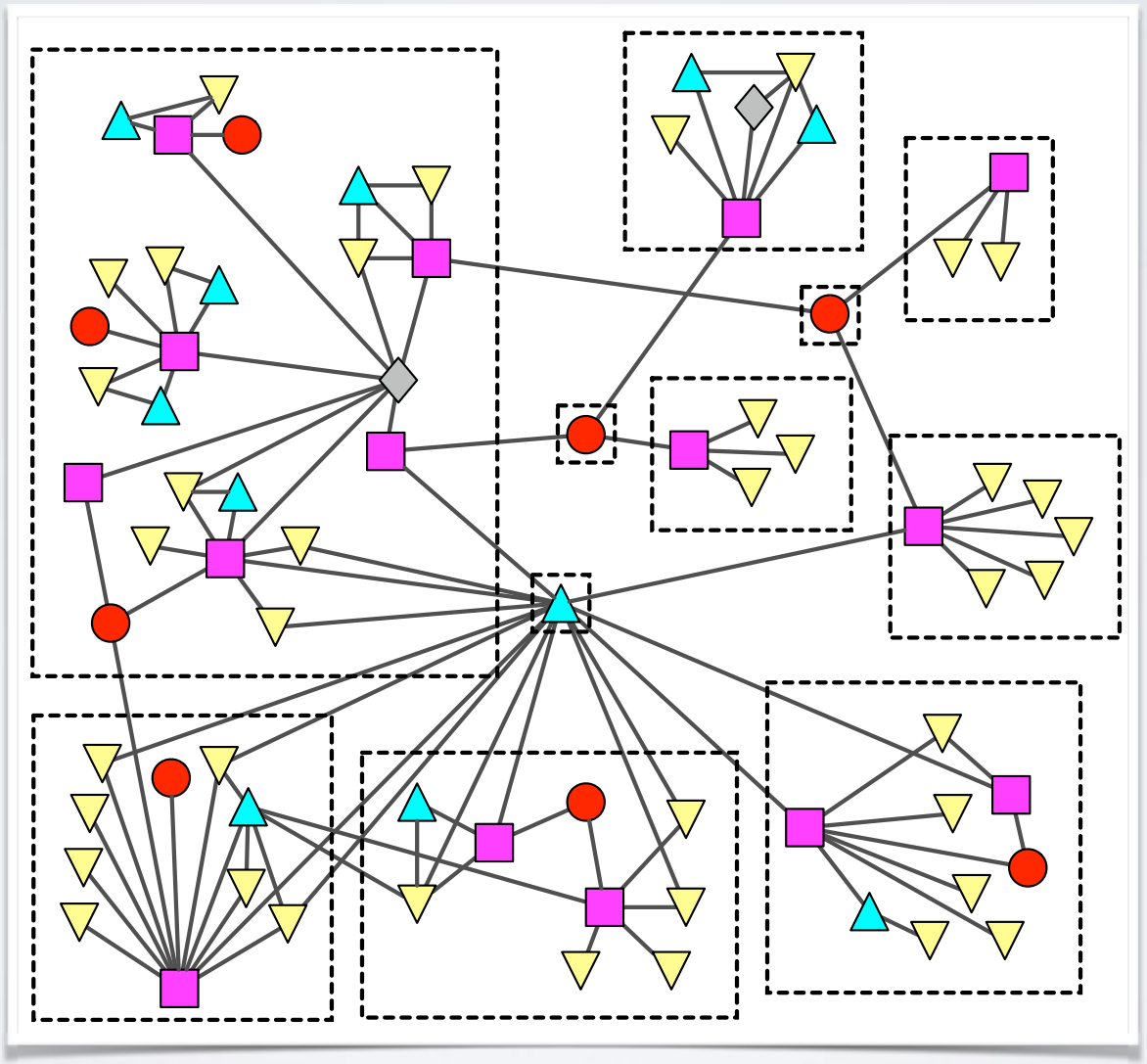
most communities are not random graphs

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

hierarchical communities

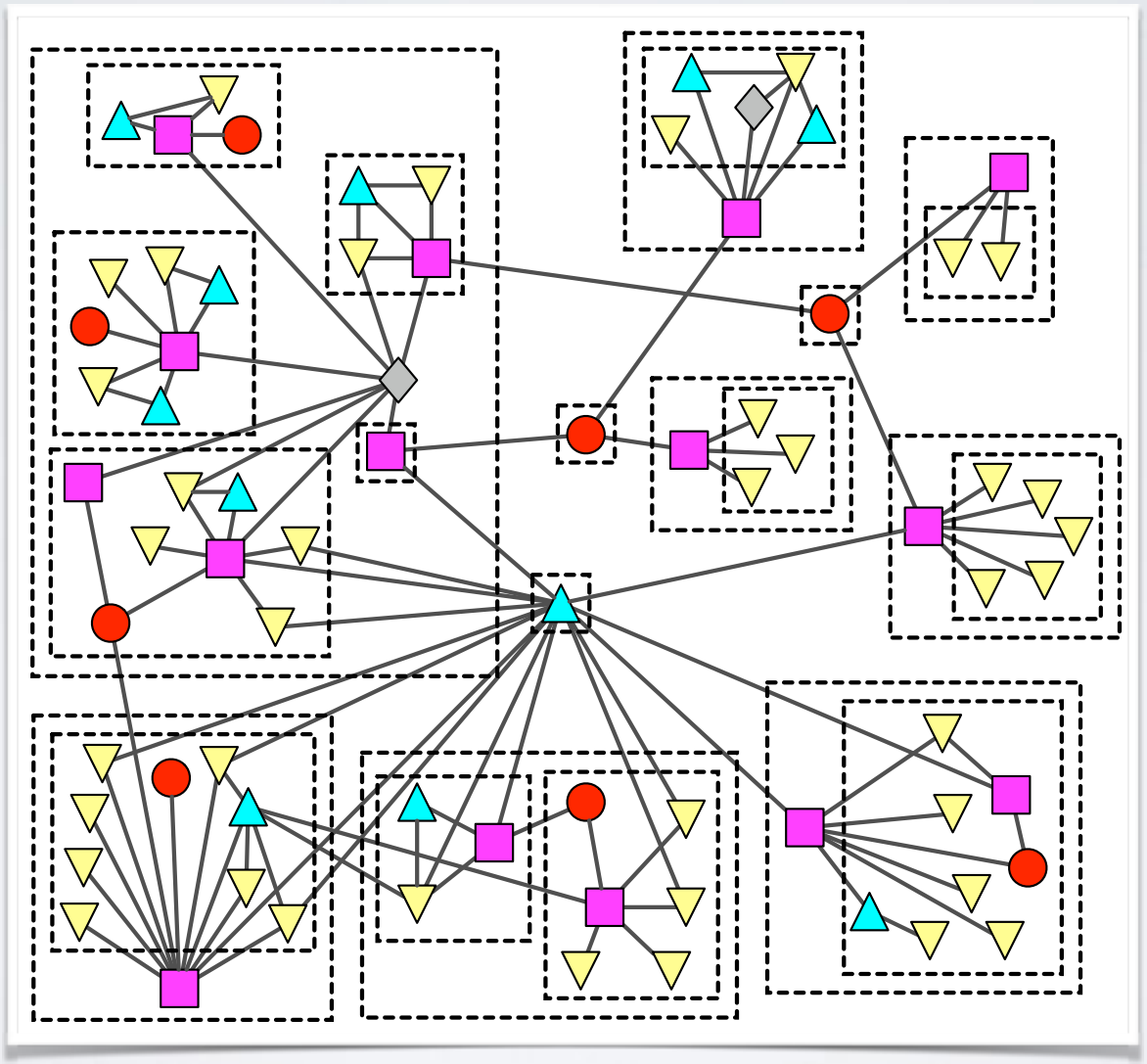


hierarchical communities



modules

hierarchical communities

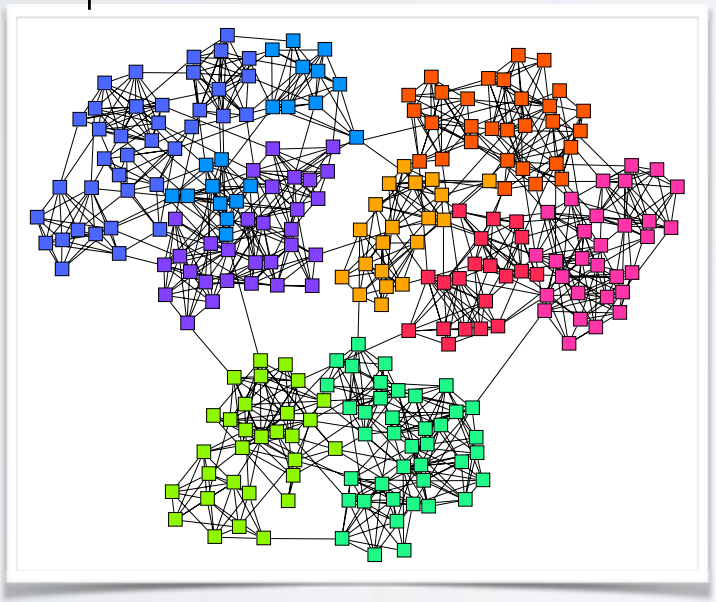


nested
modules

hierarchical communities

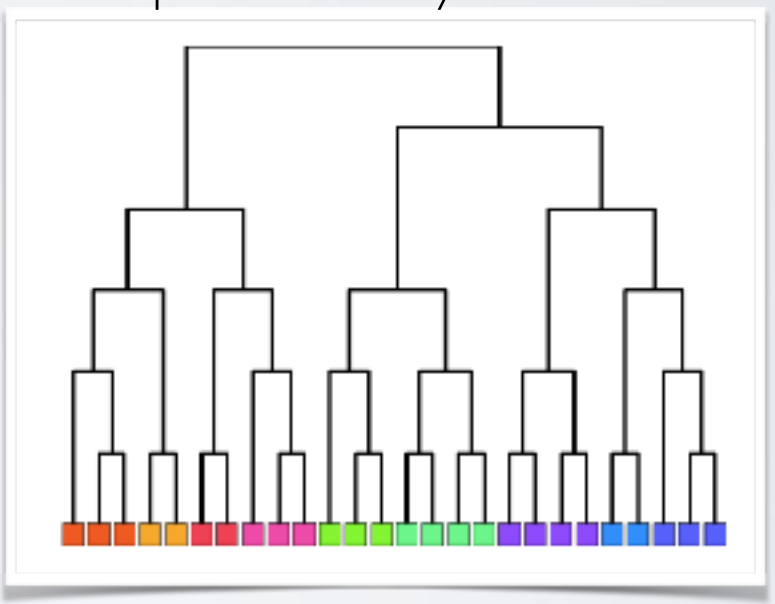
can we automatically extract such hierarchies?

step 1: network data



?

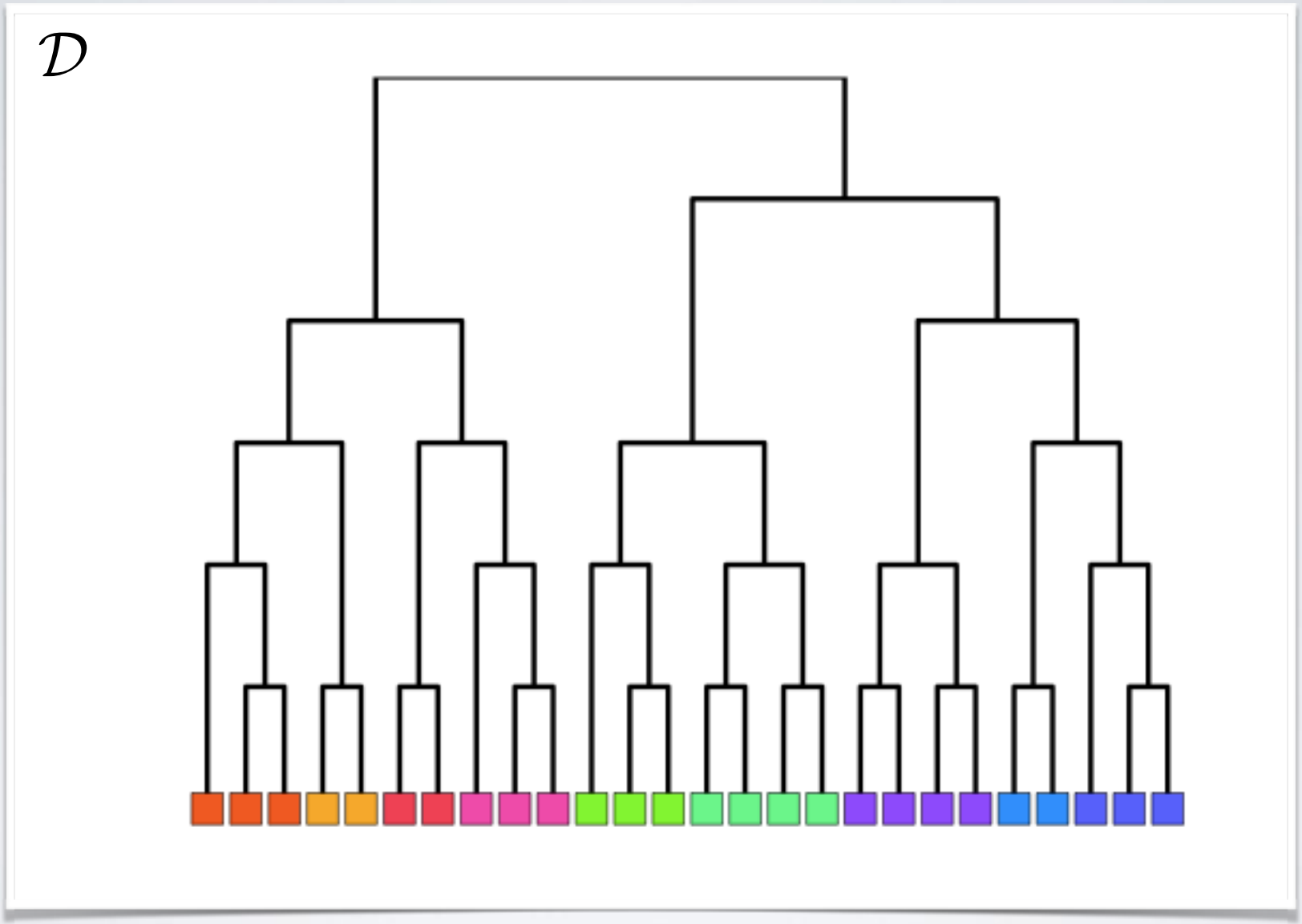
step 3: hierarchy



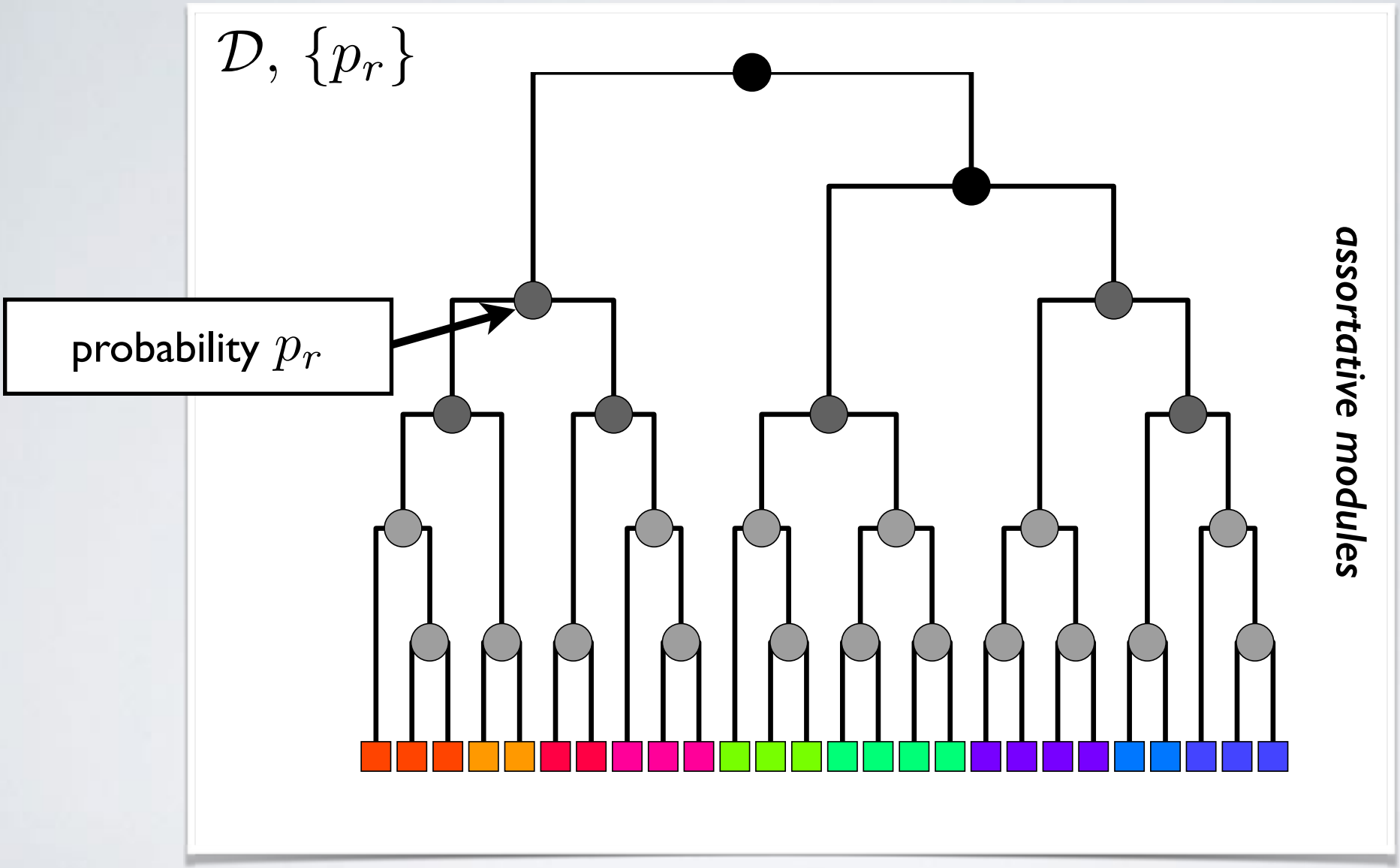
hierarchical communities

hierarchical random graph model

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

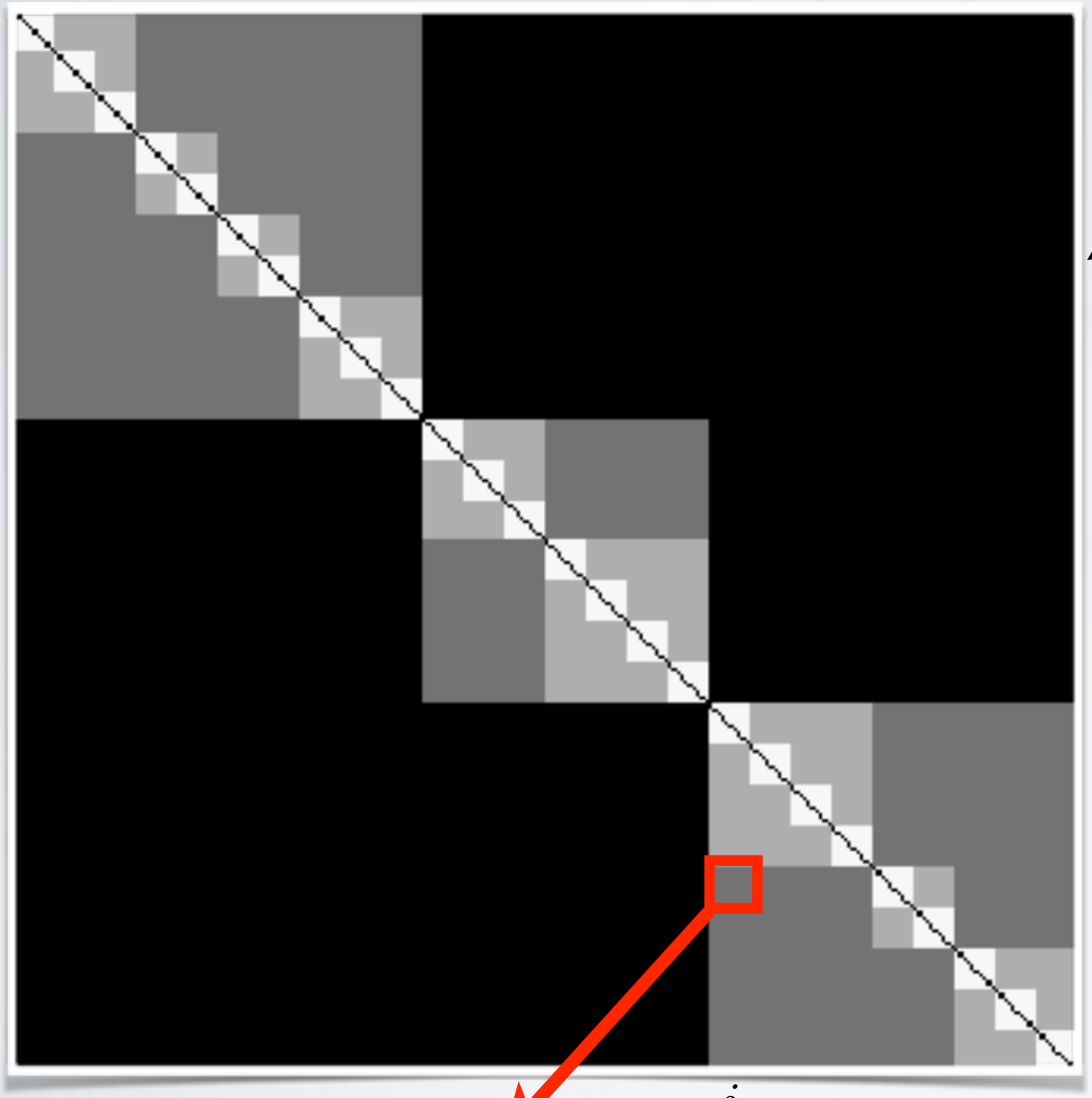


Clauset, Moore, Newman, *Nature* **453**, 98-101 (2008)
Clauset, Moore, Newman, *ICML* (2006)

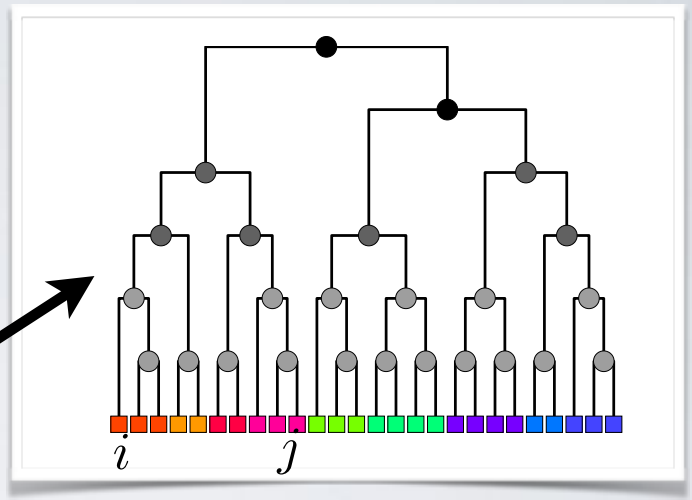


Clauset, Moore, Newman, *Nature* **453**, 98-101 (2008)
Clauset, Moore, Newman, *ICML* (2006)

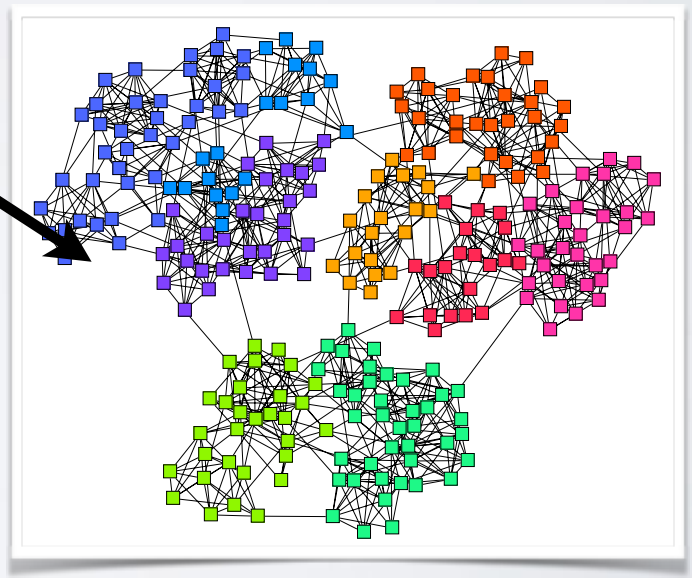
"inhomogeneous" random graph



model



instance



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

hierarchical communities

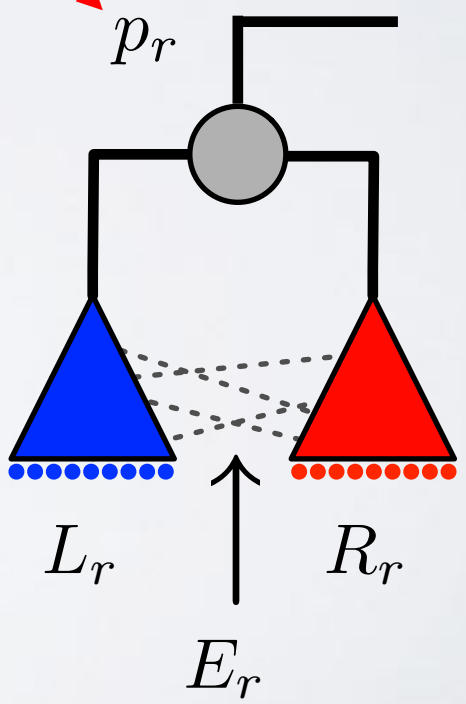
hierarchical random graph model

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

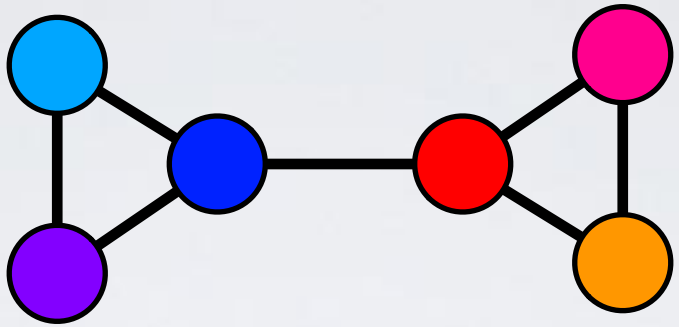
L_r = number nodes in left subtree

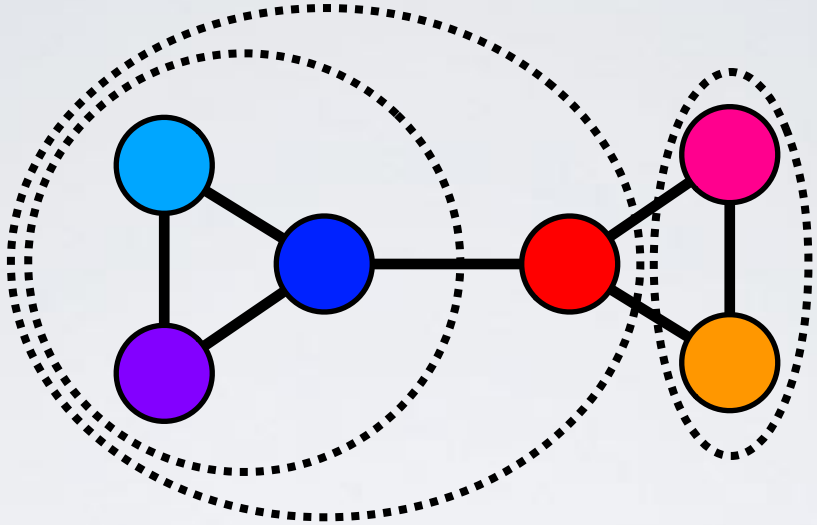
R_r = number nodes in right subtree

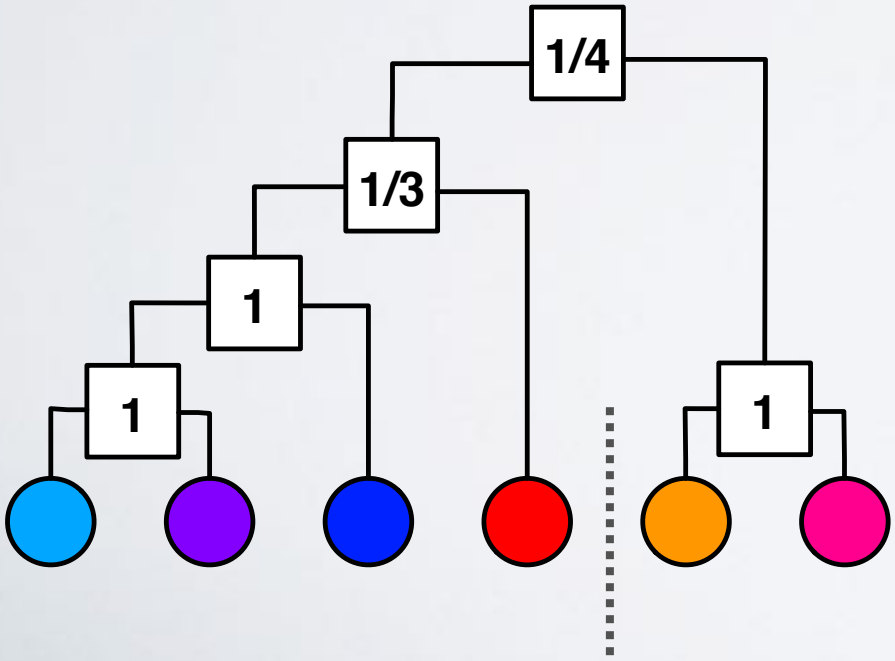
E_r = number edges with r as lowest common ancestor



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



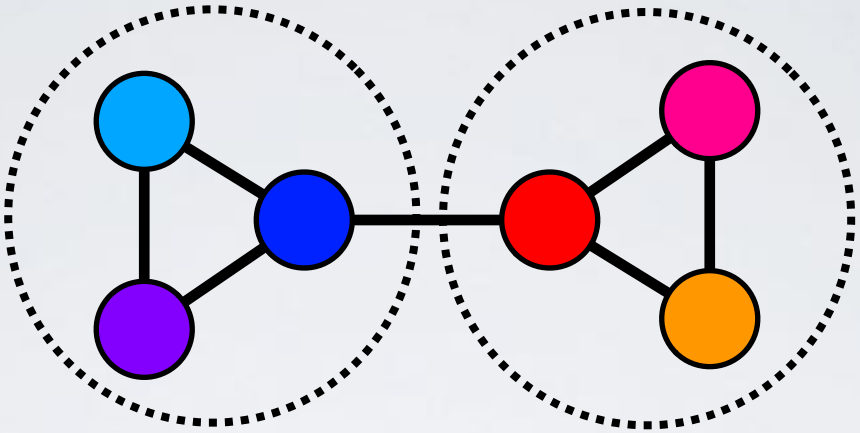




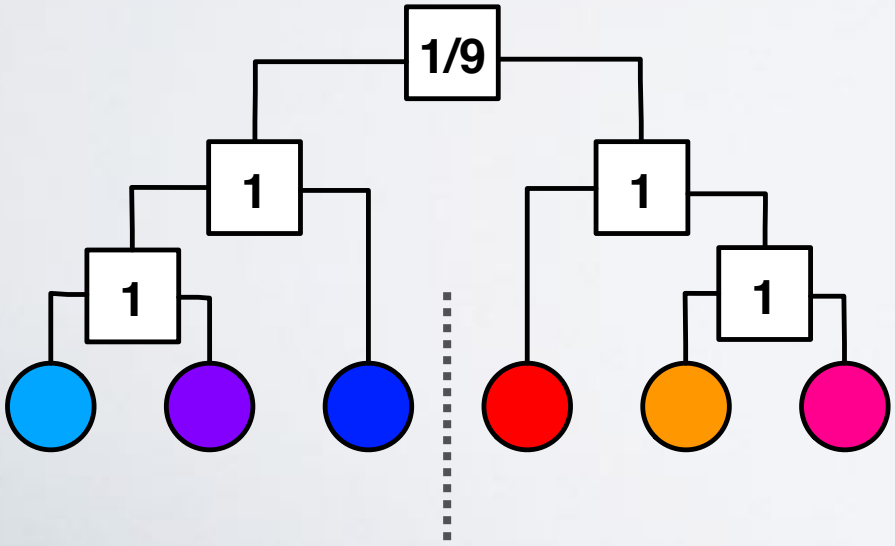
$$\mathcal{L}(\mathcal{D}, \{p_r\}) = \prod_r p_r^{E_r} (1 - p_r)^{L_r R_r - E_r}$$

$$\mathcal{L} = \left[\left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^2 \right] \cdot \left[\left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^6 \right]$$

$$\mathcal{L} = 0.0016$$



$$\mathcal{L}(\mathcal{D},\,\{p_r\}) = \prod_r p_r^{E_r}\,(1-p_r)^{L_r R_r - E_r}$$



$$\mathcal{L} = \left[\left(\frac{1}{9}\right)^1 \left(\frac{8}{9}\right)^8\right]$$

$$\mathcal{L} = 0.0433$$

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 $\Pr(G \mid \mathcal{D}, \{p_r\})$

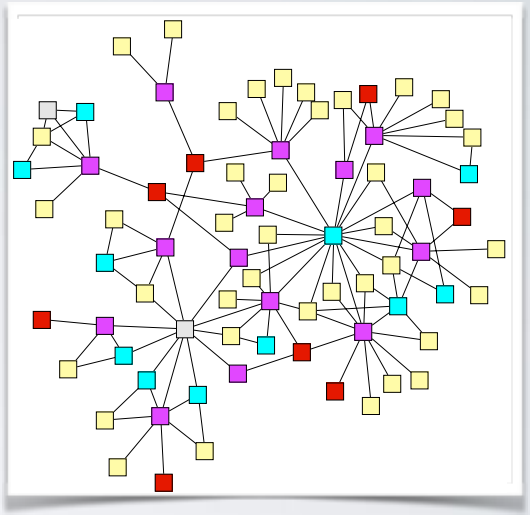
checking the models

compare resampled graphs with original data

check

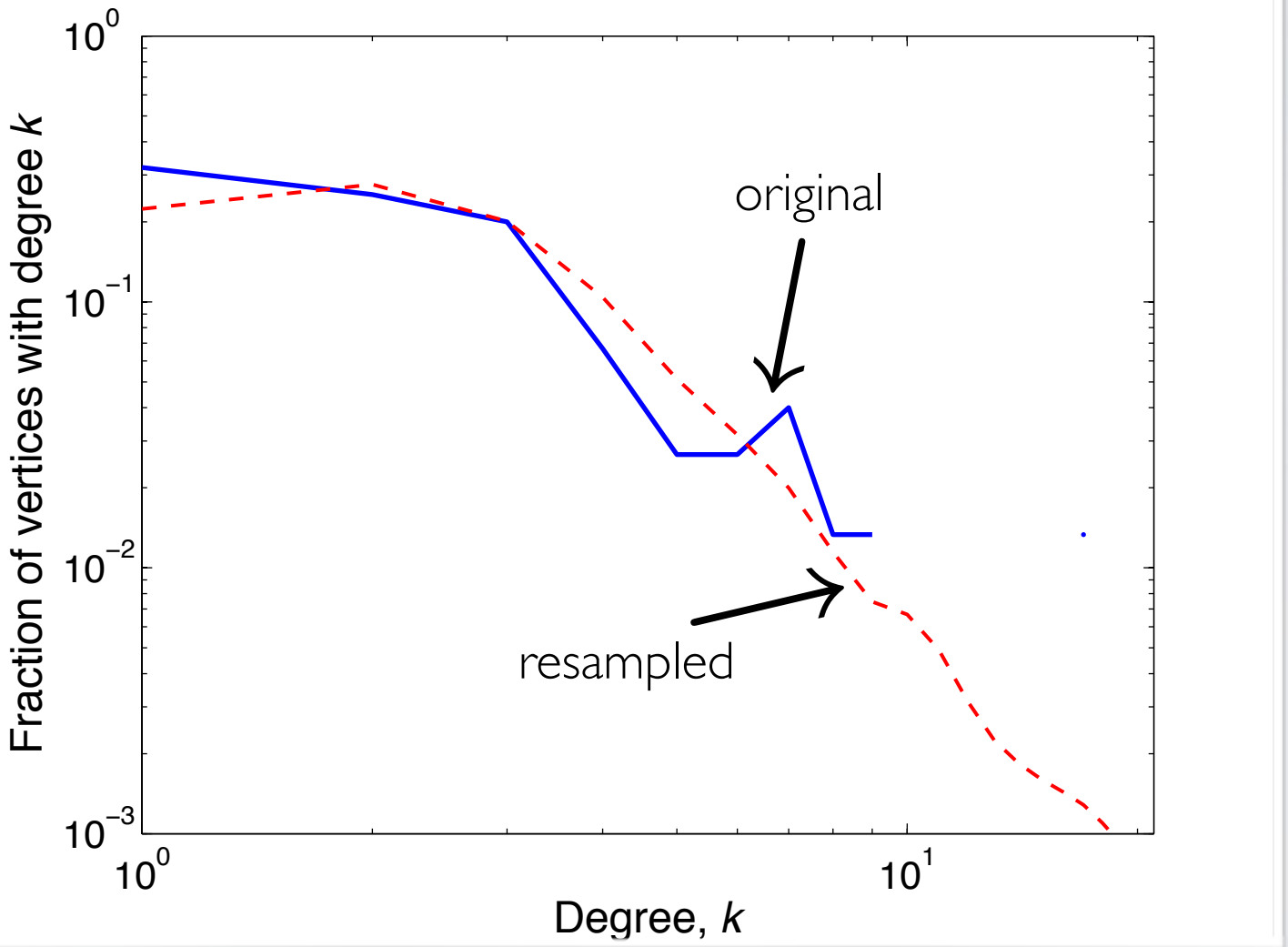
1. degree distribution
2. clustering coefficient
3. geodesic path lengths

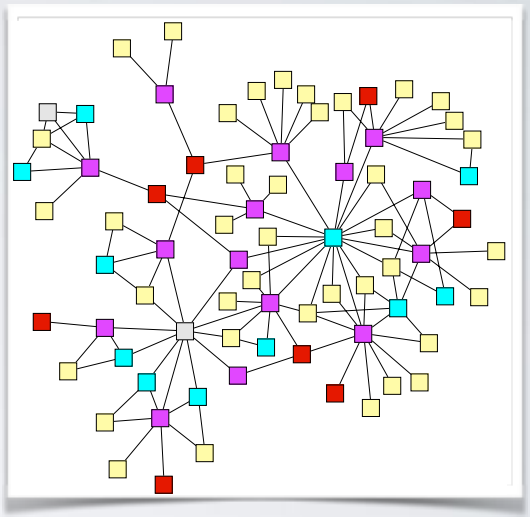
hierarchical communities



hierarchical communities

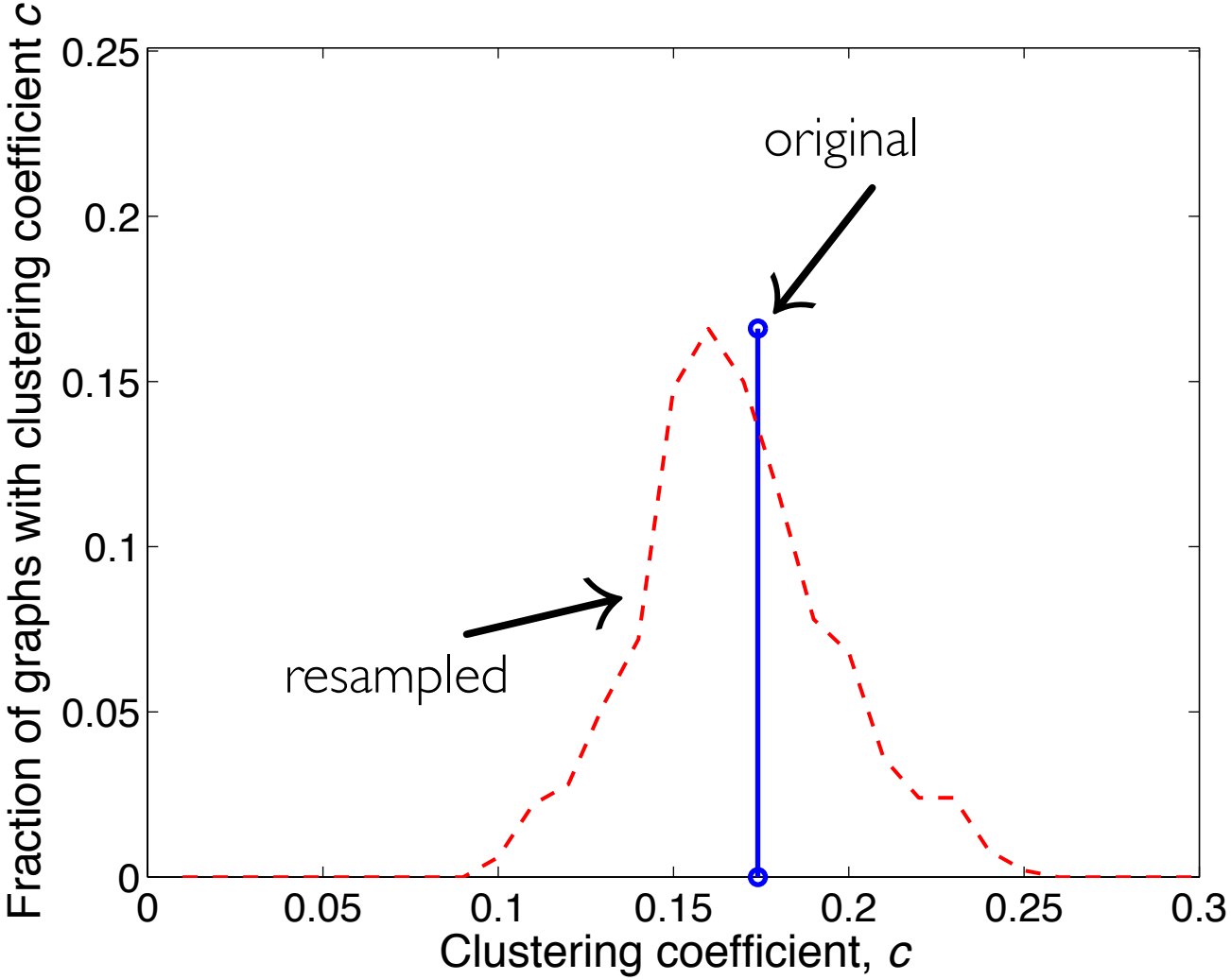
degree distribution

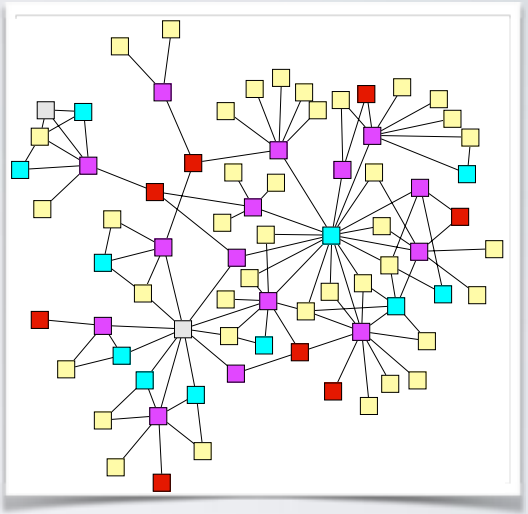




hierarchical communities

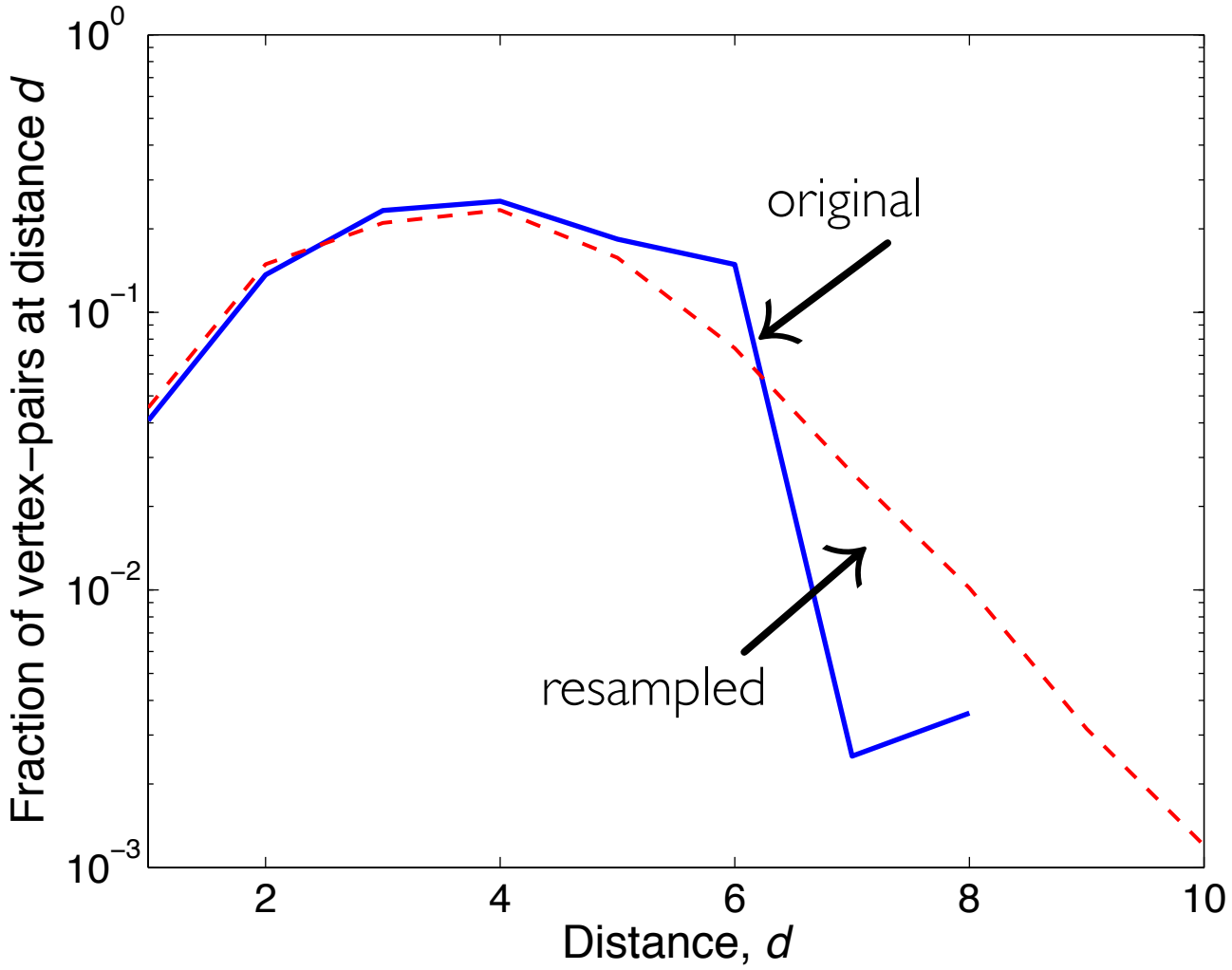
density of triangles

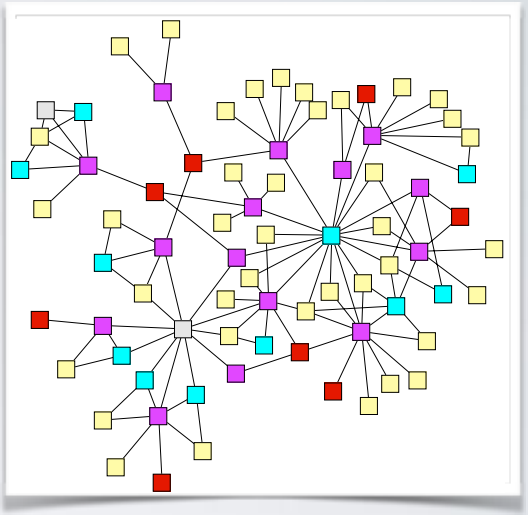




hierarchical communities

geodesic distances

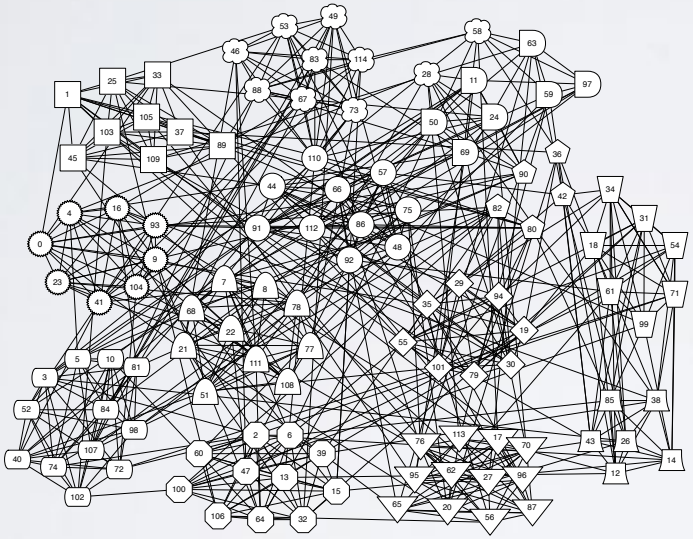




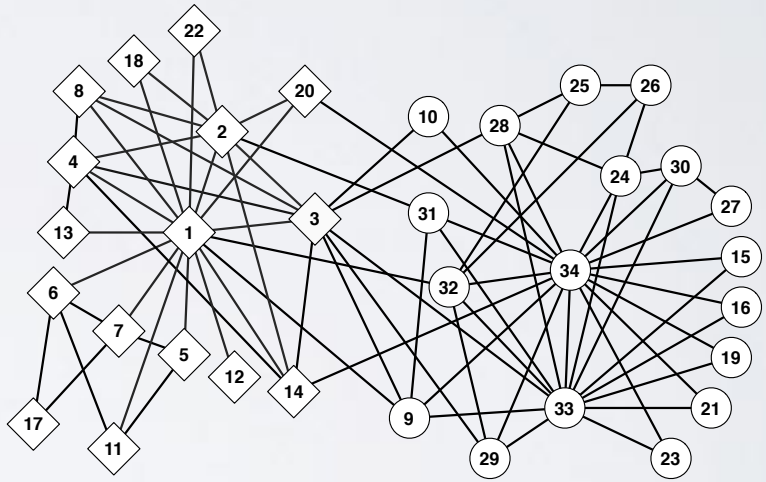
hierarchical communities

inspecting the dendrograms

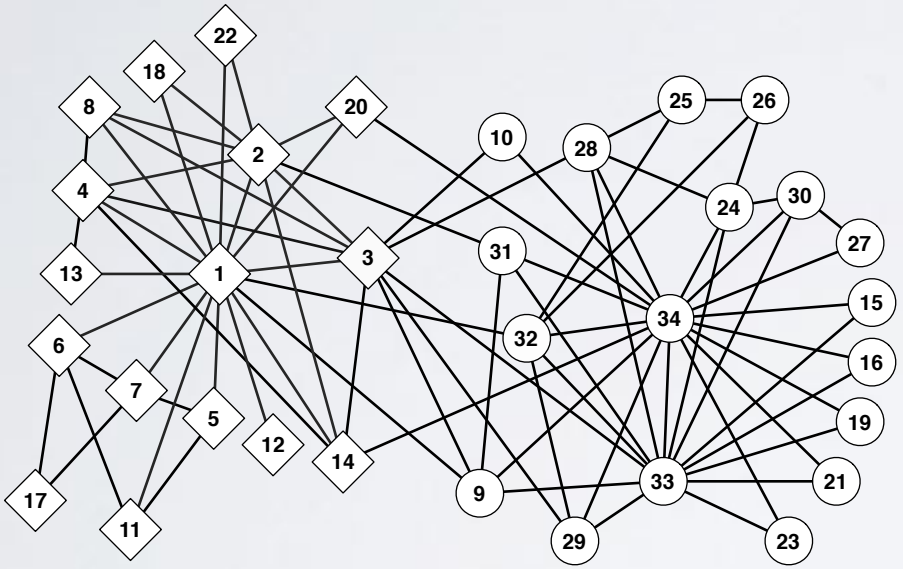
NCAA Schedule 2000

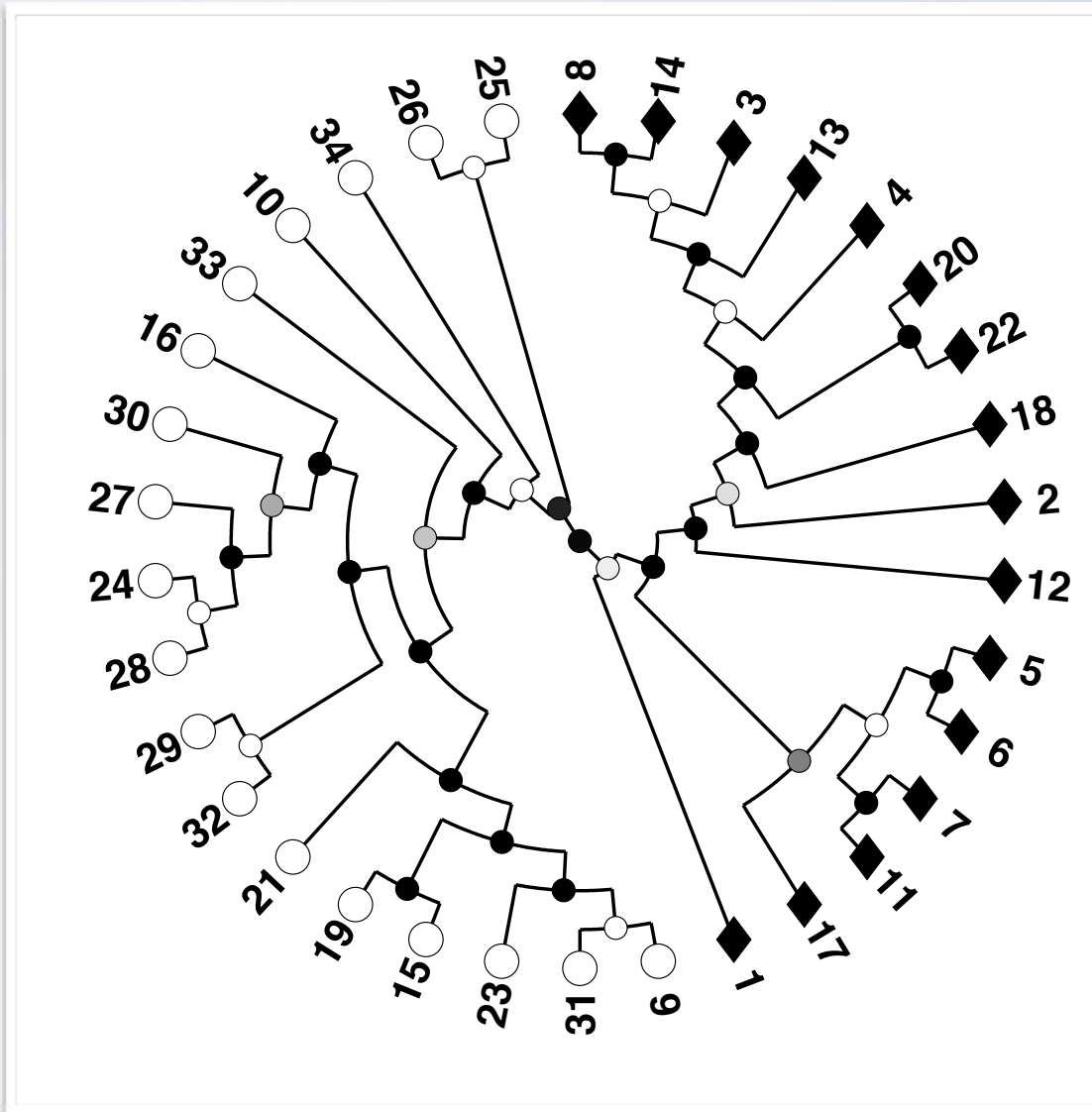


Zachary's Karate Club



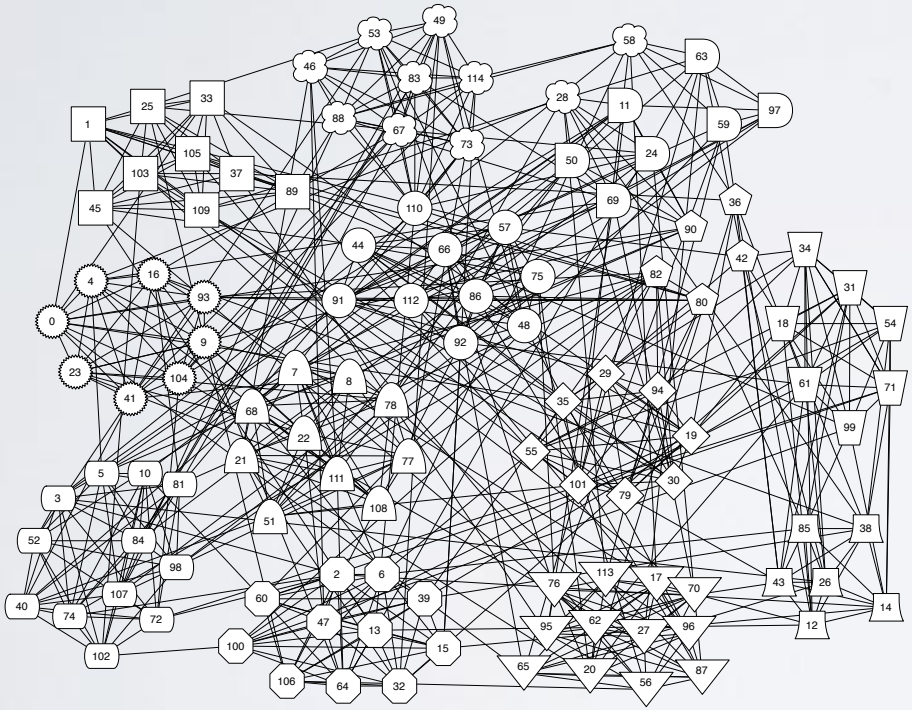
hierarchical communities

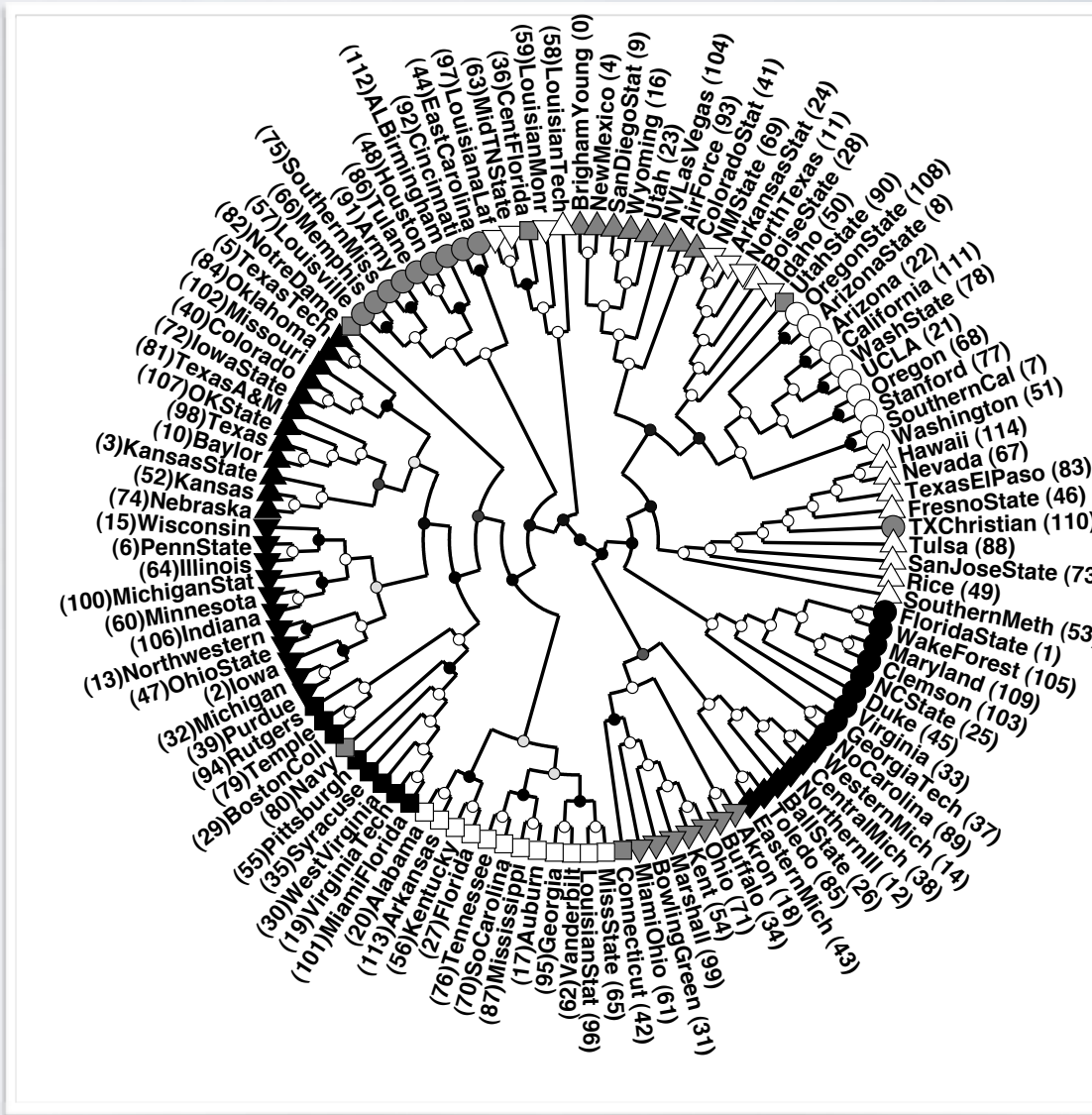




MAP

hierarchical communities





MAP

hierarchical communities

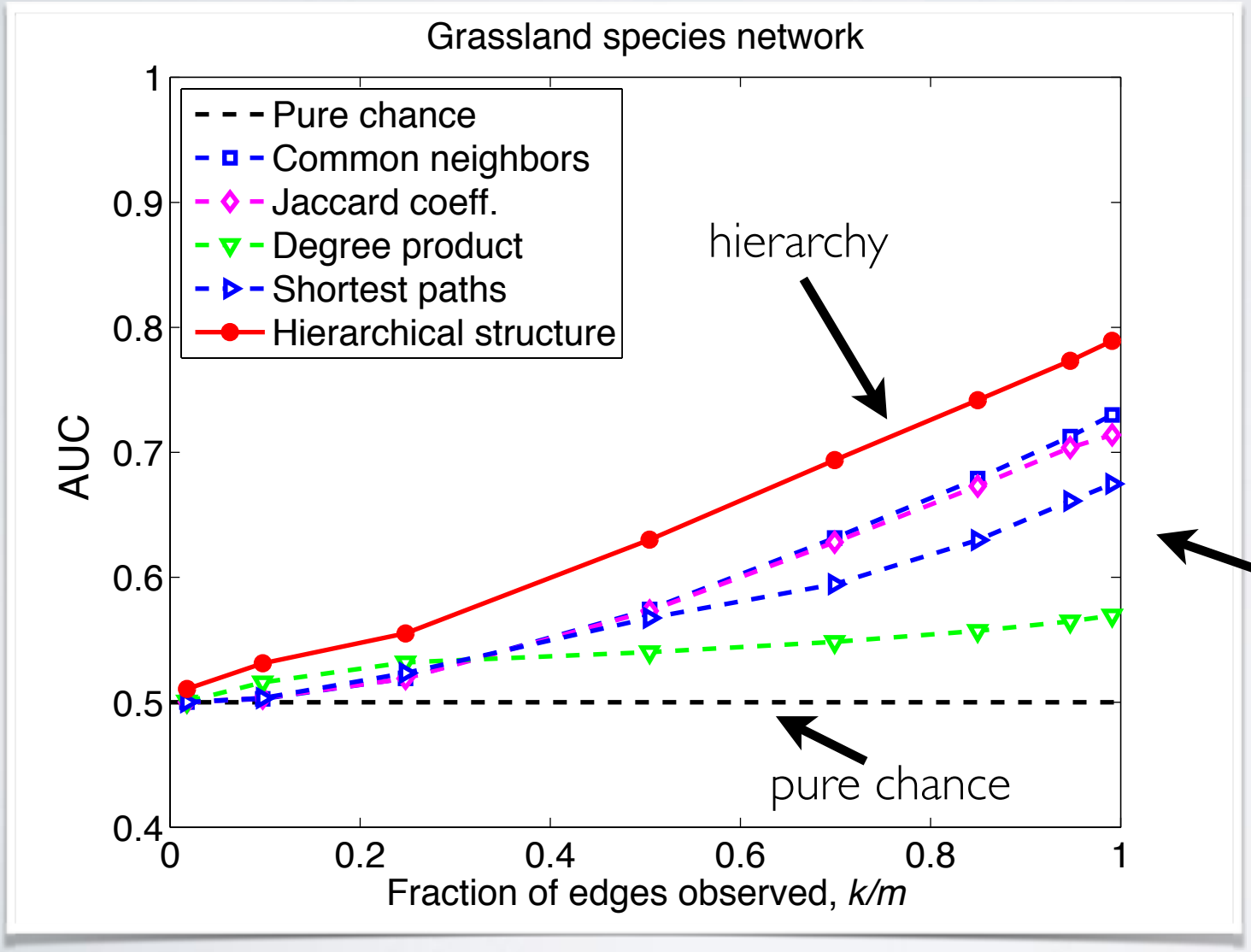
link prediction in networks

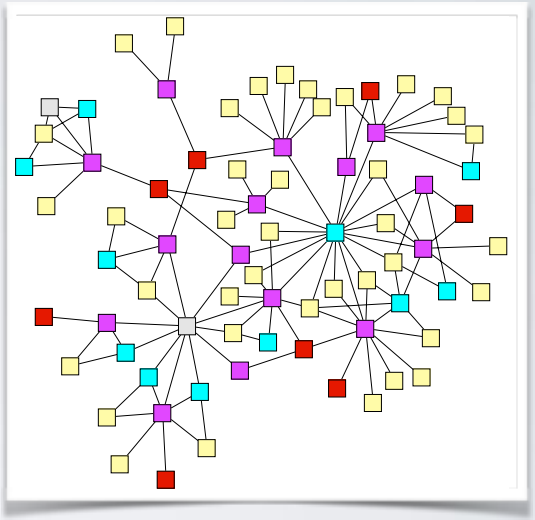
- many networks are sampled
- social nets, foodwebs, protein interactions, etc.
- generative models provide estimate of $\Pr(A_{ij} \mid \theta)$
  for either $A_{ij} = 0$ (missing links) or $A_{ij} = 1$ (spurious links)
- like cross-validation: hold out some adjacencies, $\{A_{ij}\}$
  measure accuracy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

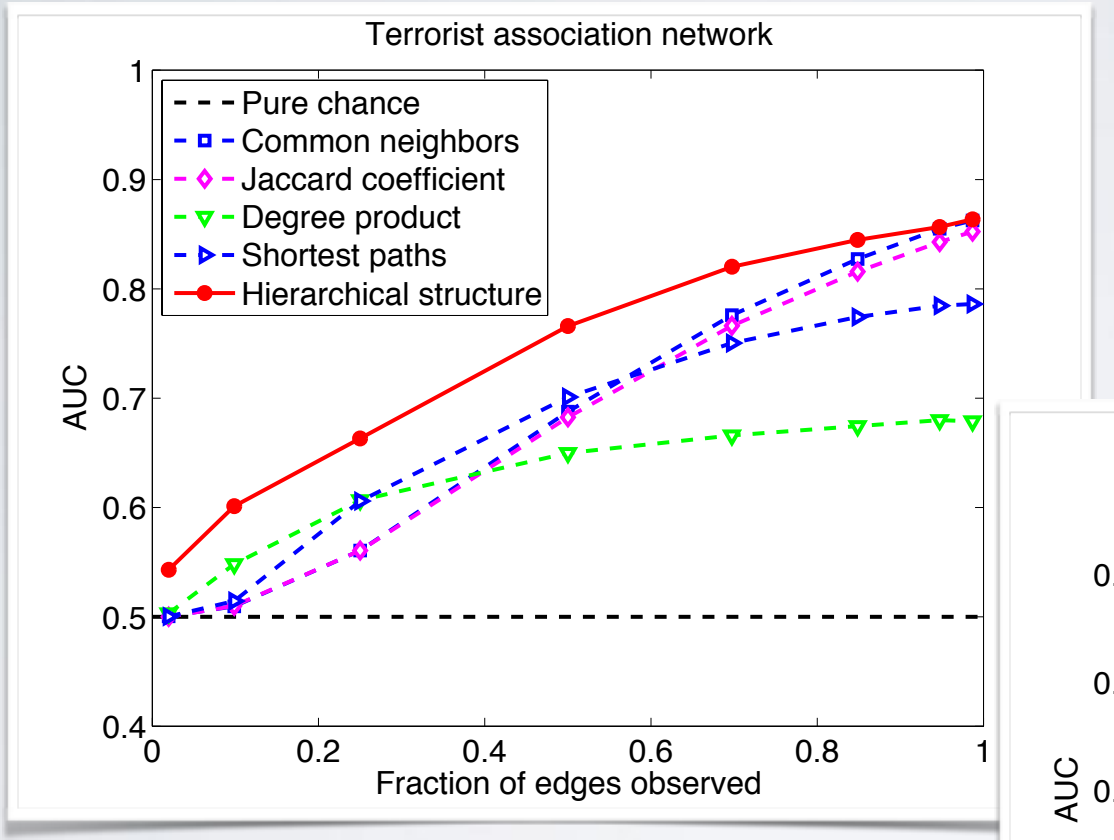
hierarchical communities

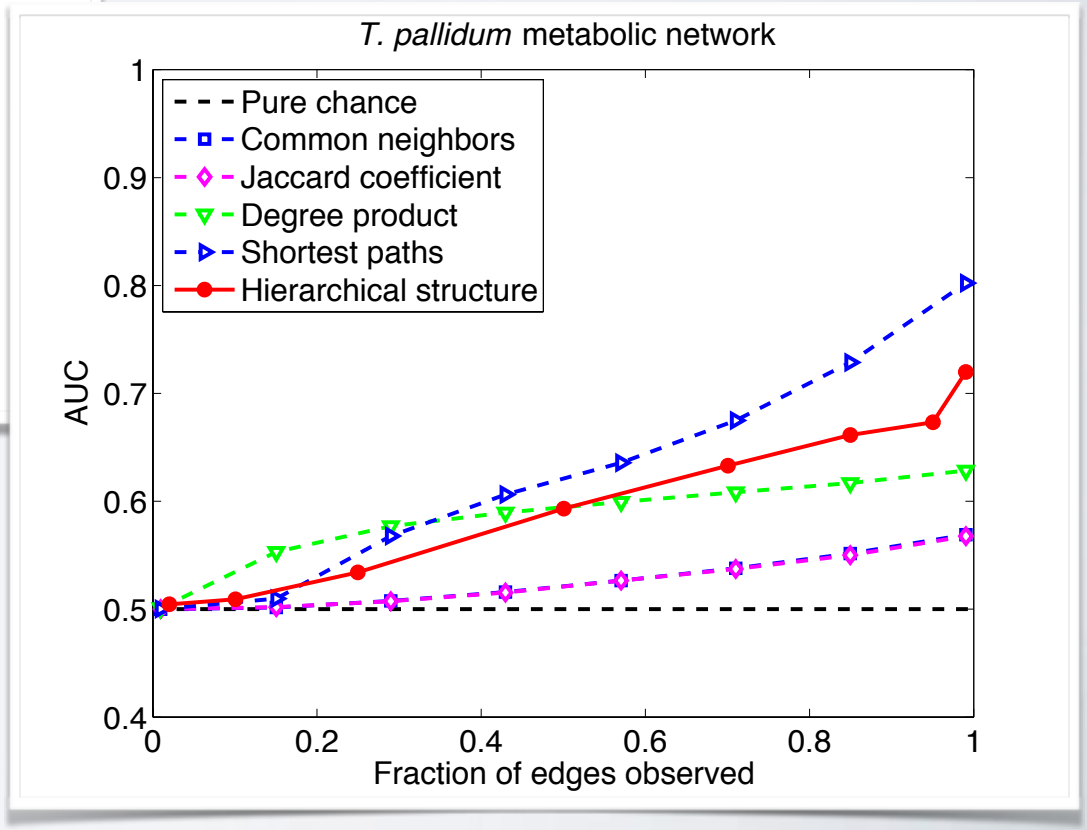




simple predictors

hierarchical communities





hierarchical communities

other approaches

hierarchical communities

other approaches

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

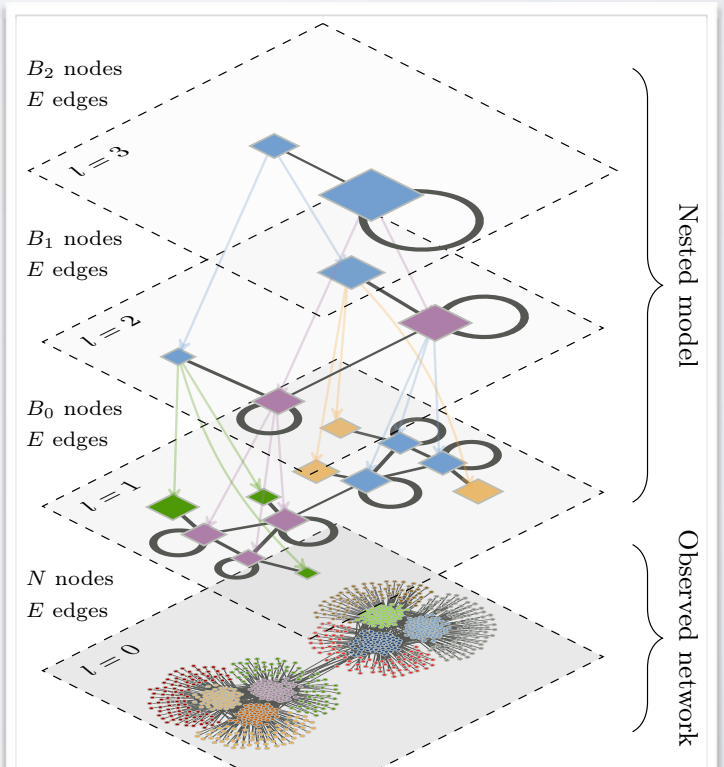
Hierarchical Block Structures and High-Resolution Model Selection in Large Networks

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

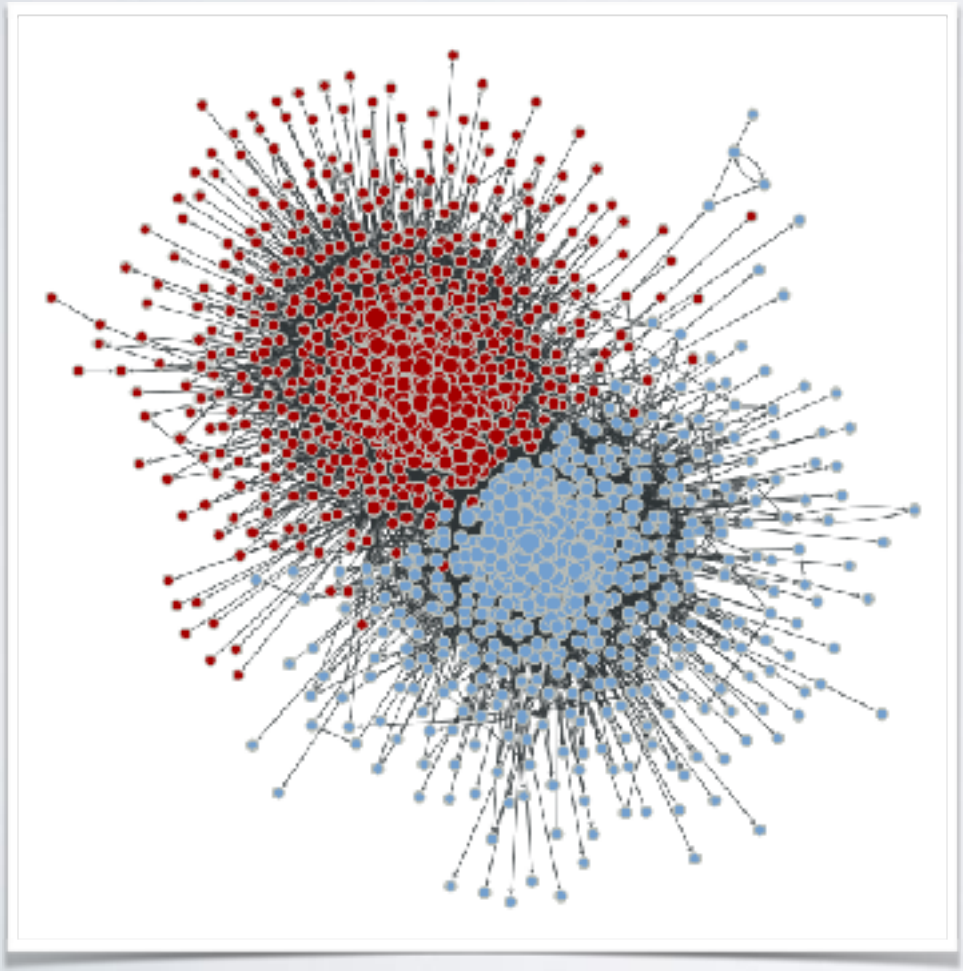


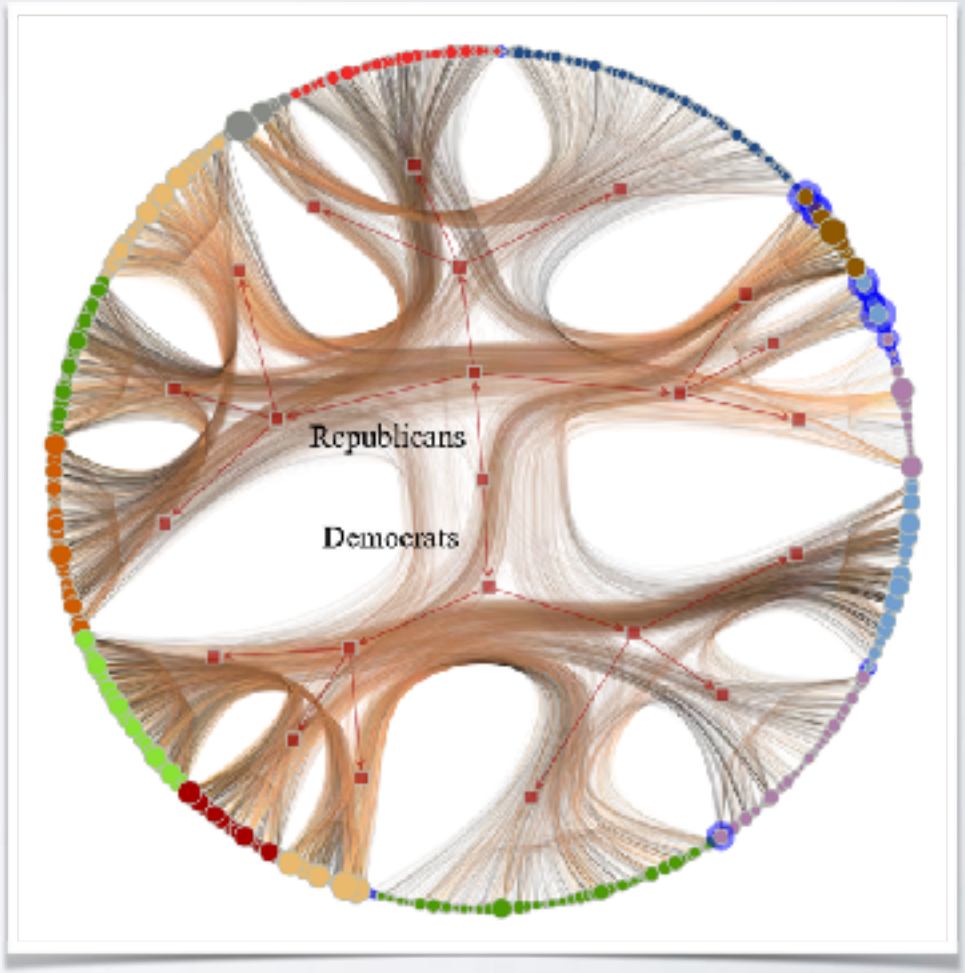
Peixoto, *Phys. Rev. X* **4**, *011047* (2014)

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

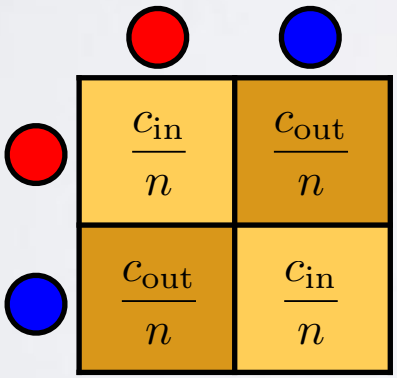
community structure in networks

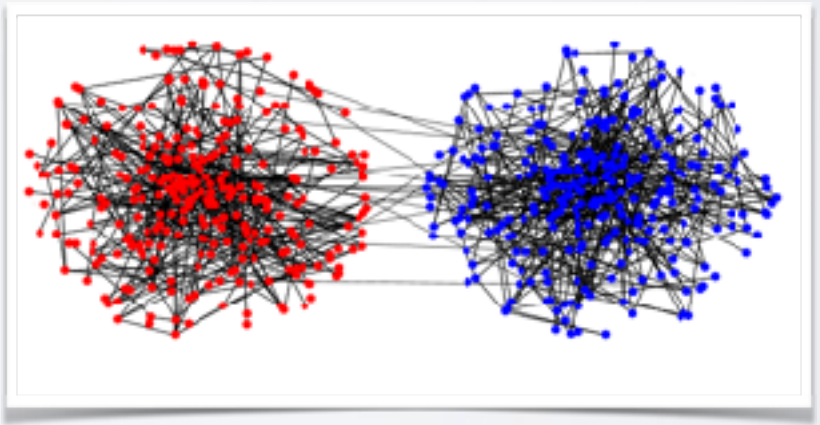
- dozens of algorithms for finding it
- generative models among the most powerful
- *how methods fail is as important as how they succeed*
- even if communities exist in a network, they may not be detectable

limits of statistical inference

planted partition problem

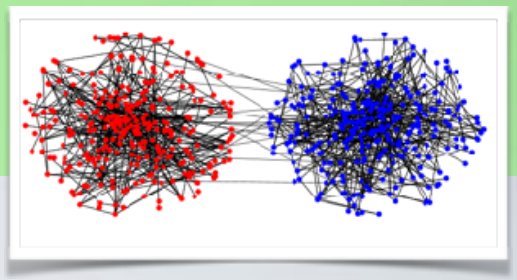
- synthetic data with known communities
- 2 groups, equal sized
- mean degree c
- parameterized strength of communities $\epsilon = c_{\text{out}}/c_{\text{in}}$





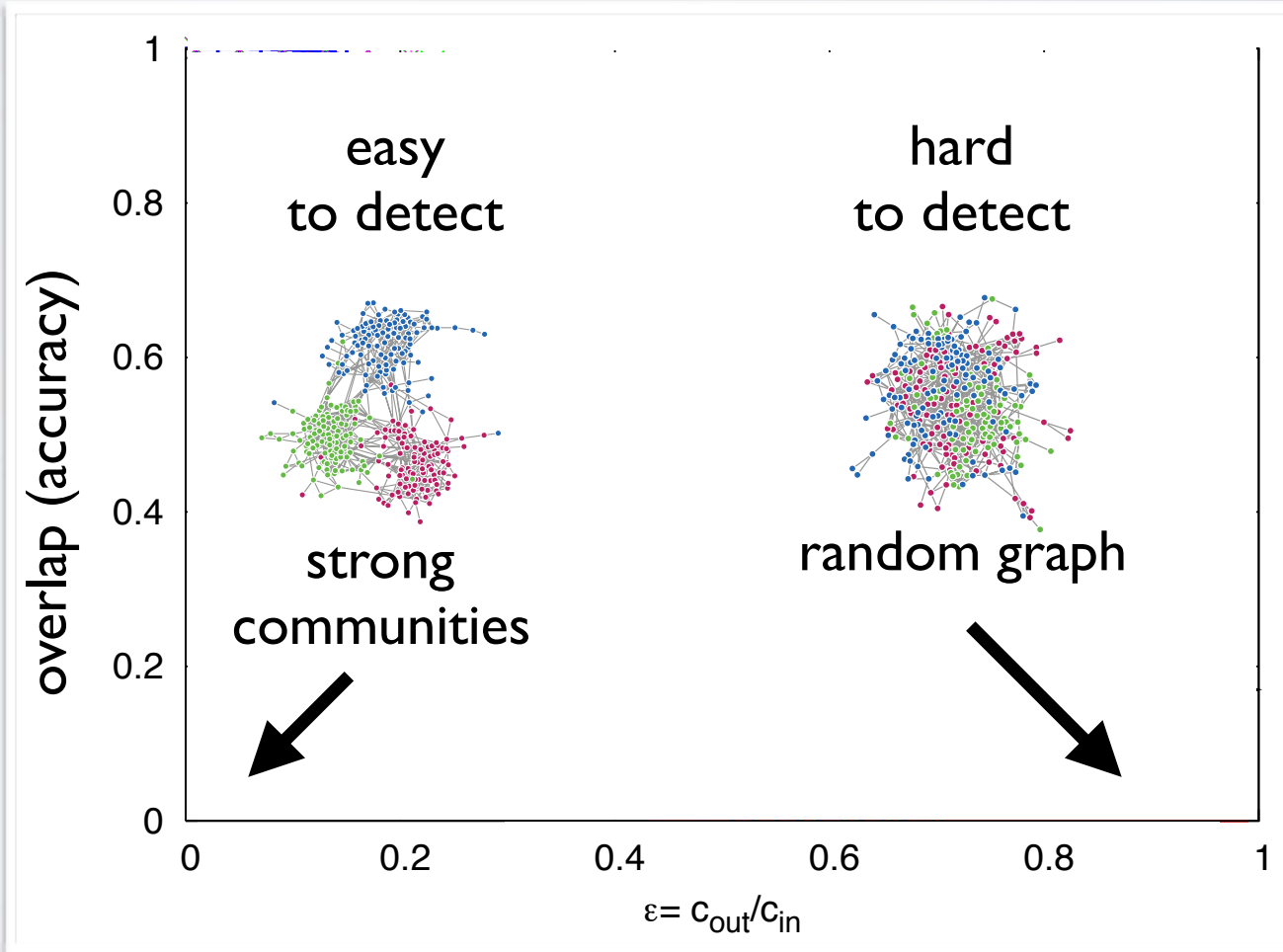
Decelle, Krzakala, Moore, & Zdeborová, *Phys. Rev. Lett.* **107**, 065701 (2011)

limits of statistical inference



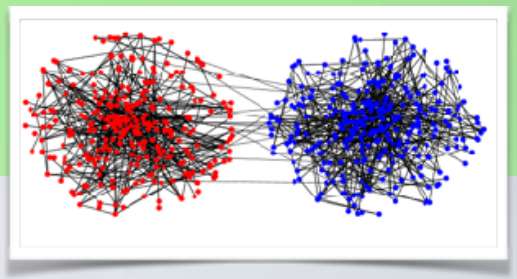
planted partition problem

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



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

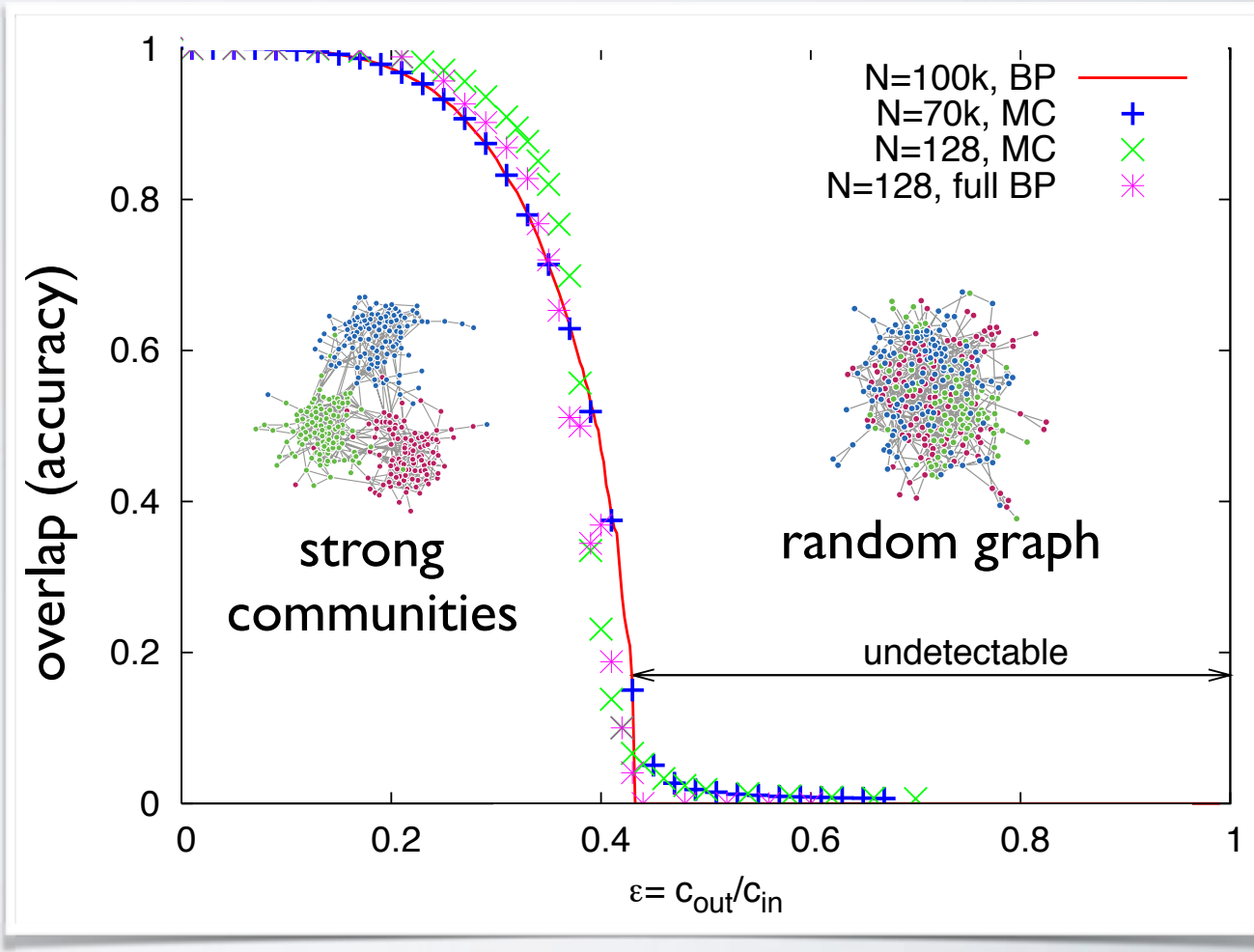
limits of statistical inference



planted partition problem

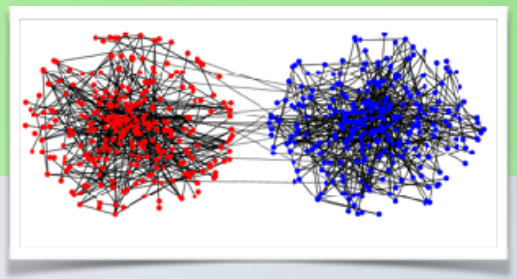
- synthetic data with known communities
- 2 groups, equal sized
- mean degree c
- 2nd order phase transition in detectability
- overlap goes to 0 for

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



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)

the trouble with community detection

the trouble with community detection

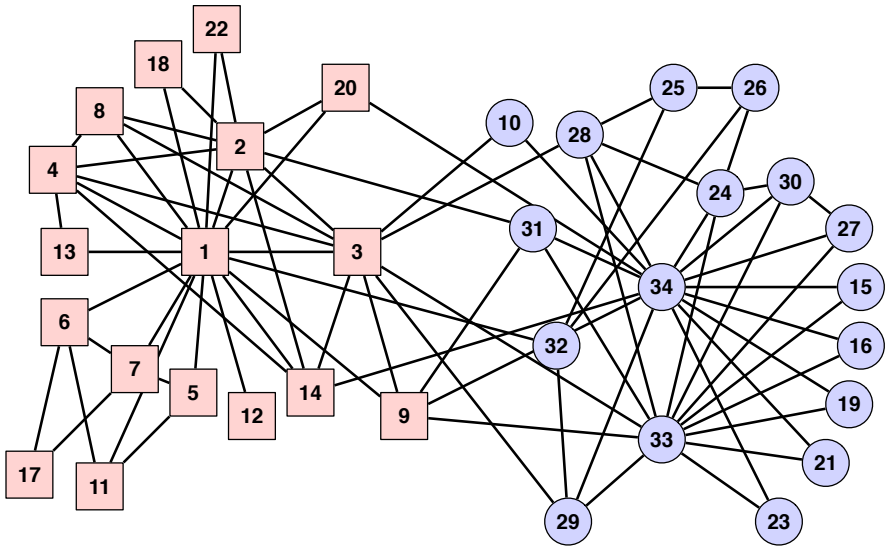
many networks include ***metadata*** on their nodes:

social networks	age, sex, ethnicity or race, etc.
food webs	feeding mode, species body mass, etc.
Internet	data capacity, physical location, etc.
protein interactions	molecular weight, association with cancer, etc.

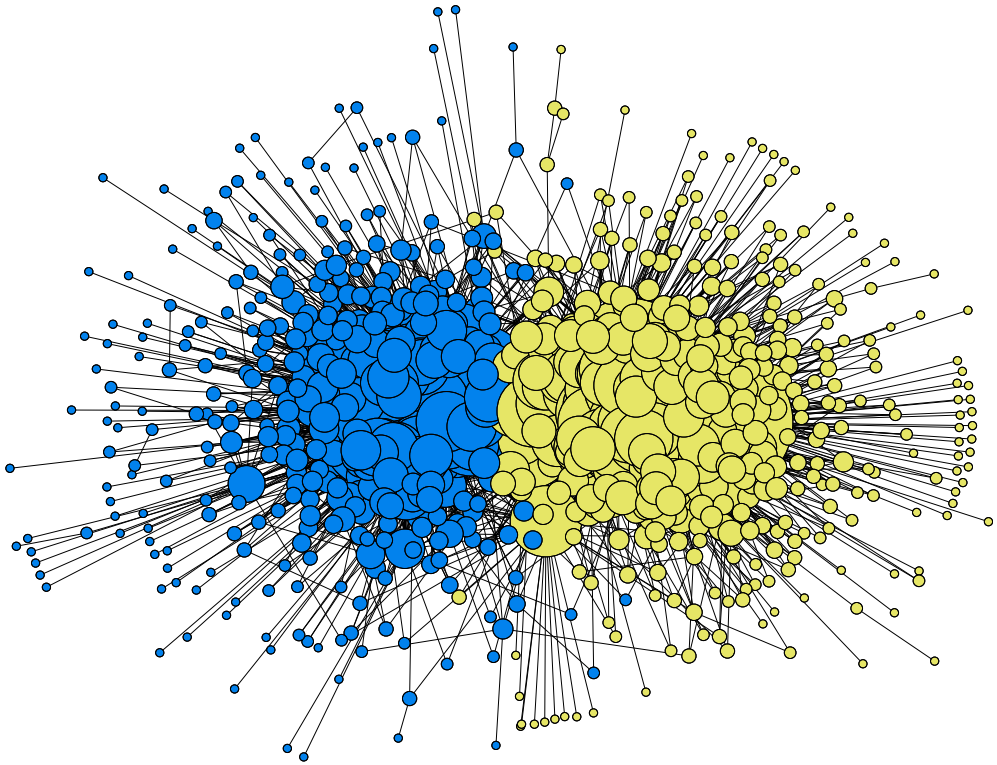
metadata $\mathbf{x}$ is often used to *evaluate the accuracy* of community detection algs.

if community detection method $\mathcal{A}$ finds a partition $\mathcal{P}$ that correlates with $\mathbf{x}$ then we say that $\mathcal{A}$ is good

the trouble with community detection

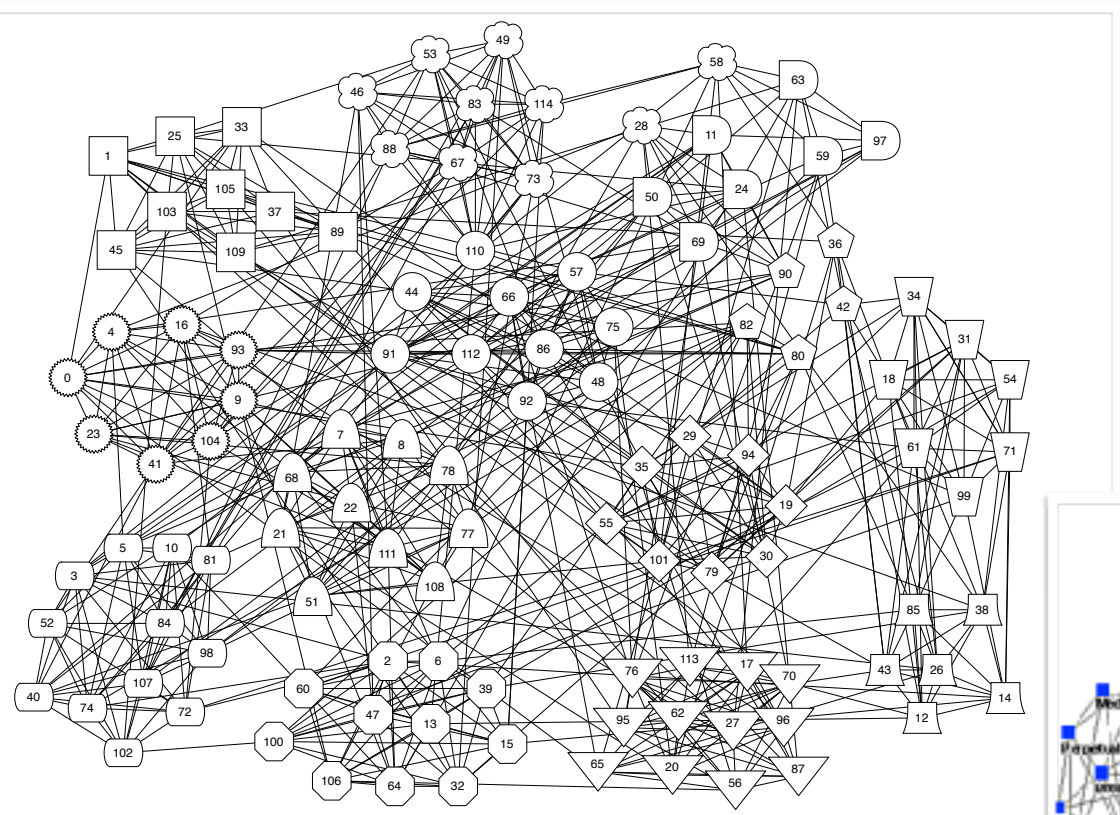


Zachary karate club



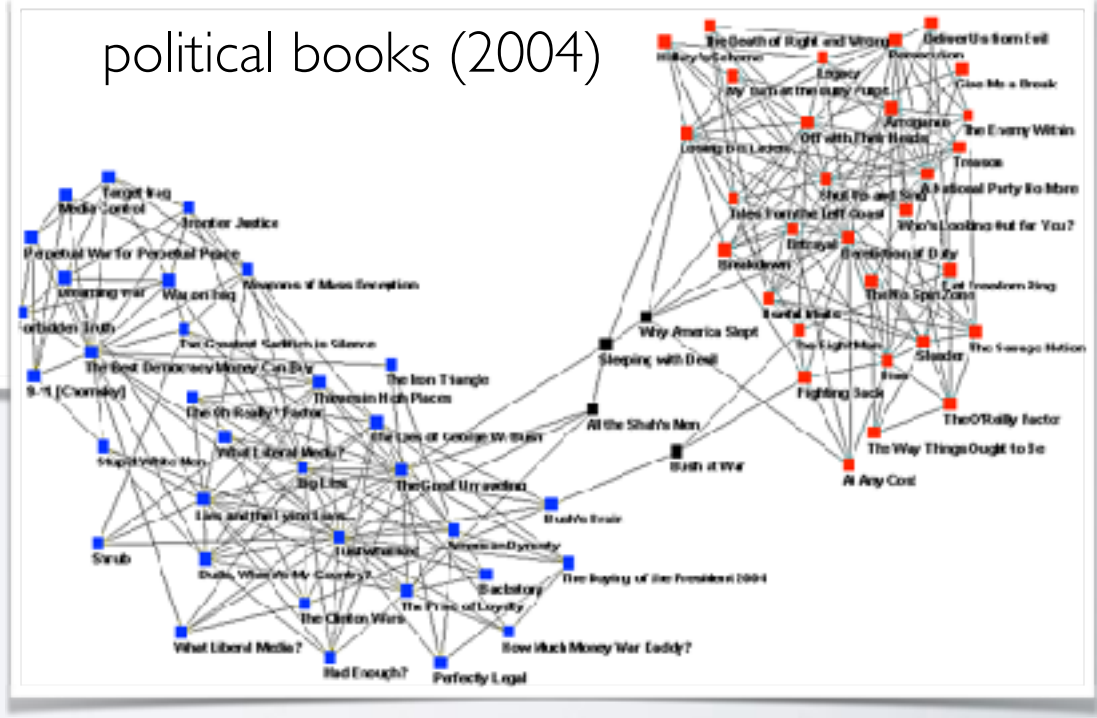
political blogs network

the trouble with community detection



NCAA 2000 Schedule

political books (2004)



the trouble with community detection

often, groups found by community detection are meaningful

- allegiances or personal interests in social networks [1]
- biological function in metabolic networks [2]

but

[1] see Fortunato (2010), and Adamic & Glance (2005)
[2] see Holme, Huss & Jeong (2003), and Guimera & Amaral (2005)

the trouble with community detection

often, groups found by community detection are meaningful

- allegiances or personal interests in social networks [1]
- biological function in metabolic networks [2]

but some recent studies claim these are the exception

- real networks ***either*** do not contain structural communities ***or*** communities exist but they do not correlate with metadata groups [3]

[1] see Fortunato (2010), and Adamic & Glance (2005)
[2] see Holme, Huss & Jeong (2003), and Guimera & Amaral (2005)
[3] see Leskovec et al. (2009), and Yang & Leskovec (2012), and Hric, Darst & Fortunato (2014)

the trouble with community detection

Hric, Darst & Fortunato (2014)

- 115 networks with metadata & 12 community detection methods
- compare extracted $\mathcal{P}$ with observed $\mathbf{x}$ for each $\mathcal{A}$

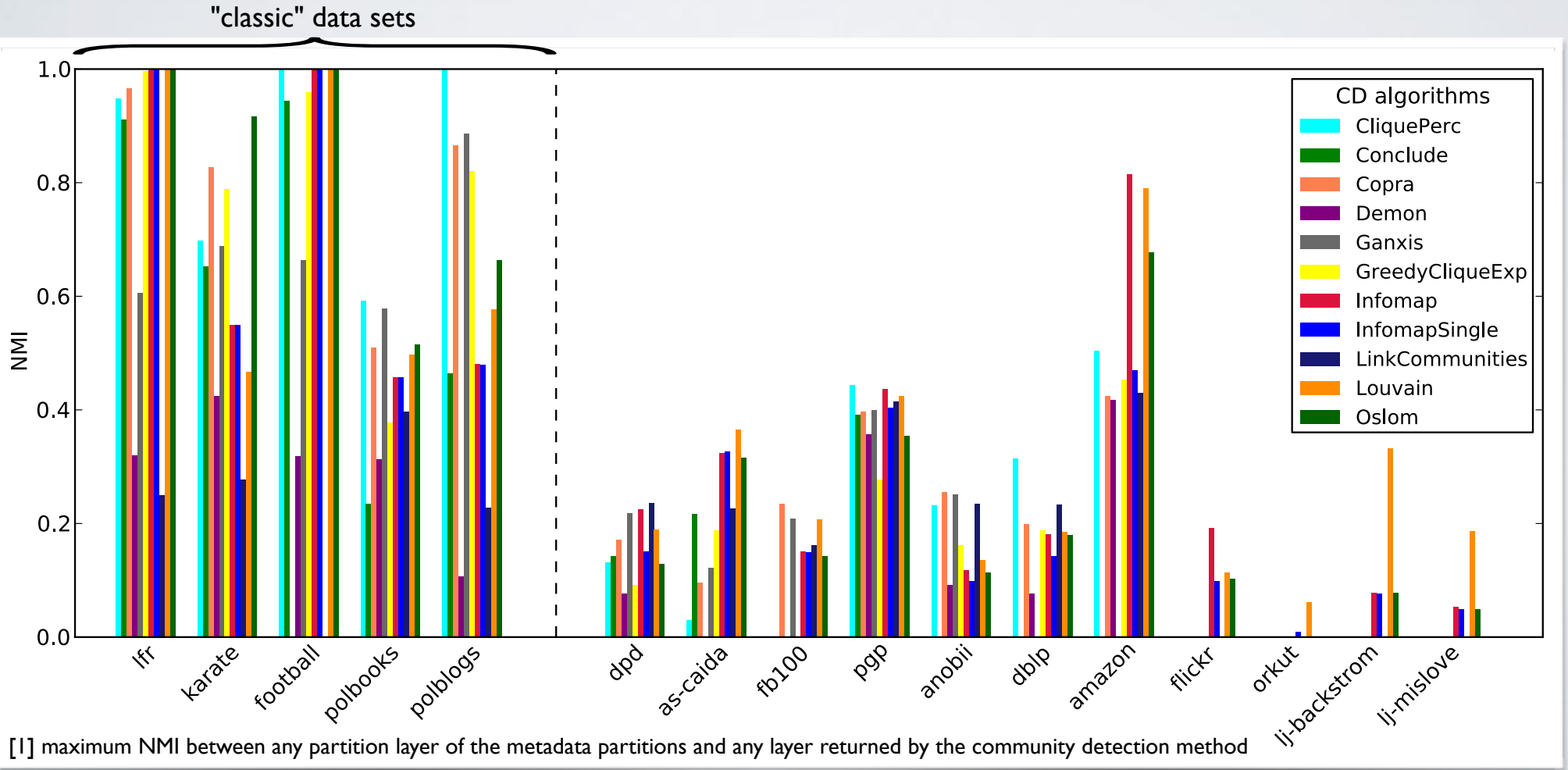
Name	No. Nodes	No. Edges	No. Groups	Description of group nature
lfr	1000	9839	40	artificial network (lfr, 1000S, $\mu = 0.5$)
karate	34	78	2	membership after the split
football	115	615	12	team scheduling groups
polbooks	105	441	2	political alignment
polblogs	1222	16782	3	political alignment
dpd	35029	161313	580	software package categories
as-caida	46676	262953	225	countries
fb100	762–41536	16651–1465654	2–2597	common students' traits
pgp	81036	190143	17824	email domains
anobii	136547	892377	25992	declared group membership
dblp	317080	1049866	13472	publication venues
amazon	366997	1231439	14–29432	product categories
flickr	1715255	22613981	101192	declared group membership
orkut	3072441	117185083	8730807	declared group membership
lj-backstrom	4843953	43362750	292222	declared group membership
lj-mislove	5189809	49151786	2183754	declared group membership

[1] fb100 is 100 networks

the trouble with community detection

Hric, Darst & Fortunato (2014)

- evaluate by normalized mutual information $\mathrm{NMI}(\mathcal{P}, \mathbf{x})$



[1] maximum NMI between any partition layer of the metadata partitions and any layer returned by the community detection method

but wait!

[1] image copyright BostonGazette or maybe 20th Century Fox? gah

a solution

idea:

use metadata $\mathbf{x}$ to help select a partition $\mathcal{P}^* \in \{\mathcal{P}\}$ that correlates with $\mathbf{x}$, from among the exponential number of *plausible* partitions

[1] image copyright BostonGazette or maybe 20th Century Fox? gah

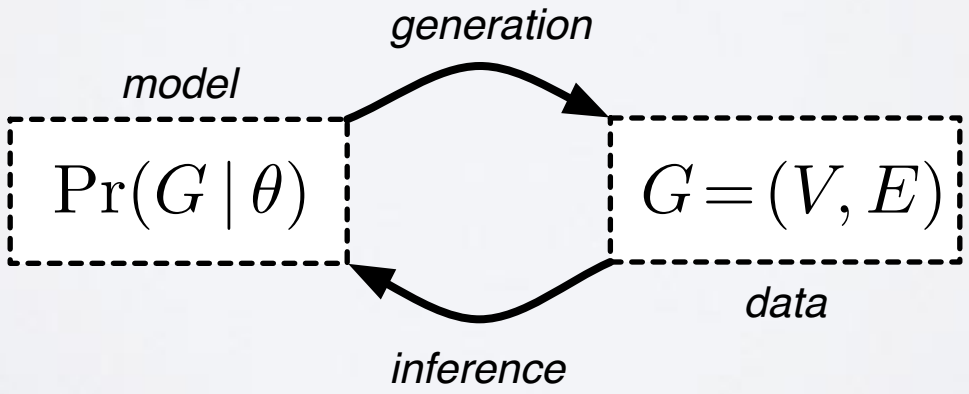
a solution

idea:

use metadata $\mathbf{x}$ to help select a partition $\mathcal{P}^* \in \{\mathcal{P}\}$ that correlates with $\mathbf{x}$, from among the exponential number of *plausible* partitions

use a generative model to guide the selection:

- **define a parametric probability distribution over networks** $\Pr(G \mid \theta)$
- ***generation*** **: given** θ**, draw** G **from this distribution**
- ***inference*** **: given** G**, choose** θ **that makes** G **likely**



a metadata-aware stochastic block model

generation

given metadata $\mathbf{x} = \{x_u\}$ and degree $\mathbf{d} = \{d_u\}$ for each node u

- each node u is assigned a community s with probability γ_{sx}
- thus, prior on community assignments is $P(s \,|\, \Gamma, \mathbf{x}) = \prod_i \gamma_{s_i, x_i}$
- given assignments, place edges independently, each with probability:

$$p_{uv} = d_u d_v \theta_{s_u, s_v}$$

- where the θ_{st} are the stochastic block matrix parameters

this is a degree-corrected stochastic block model (DC-SBM)
with a metadata-based prior on community labels

[1] Γ is the $k \times K$ matrix of parameters γ_{sx}
[2] Karrer & Newman (2011)

a metadata-aware stochastic block model

inference

given observed network $\mathbf{A}$ (adjacency matrix)

- the model likelihood is

$$P(\mathbf{A} \,|\, \Theta, \Gamma, \mathbf{x}) = \sum_{\mathbf{s}} \overbrace{P(\mathbf{A} \,|\, \Theta, \mathbf{s})}^{\text{network}} \overbrace{P(\mathbf{s} \,|\, \Gamma, \mathbf{x})}^{\text{metadata}}$$

$$= \sum_{\mathbf{s}} \prod_{u<v} p_{uv}^{A_{uv}} (1 - p_{uv})^{1 - A_{uv}} \prod_{u} \gamma_{s_u, x_u}$$

- where Θ is a $k \times k$ matrix of community interaction parameters θ_{st}, and the sum is over all possible assignments $\mathbf{s}$
- we fit this model to data using expectation-maximization (EM) to maximize $P(\mathbf{A} \,|\, \Theta, \Gamma, \mathbf{x})$ w.r.t. Θ and Γ

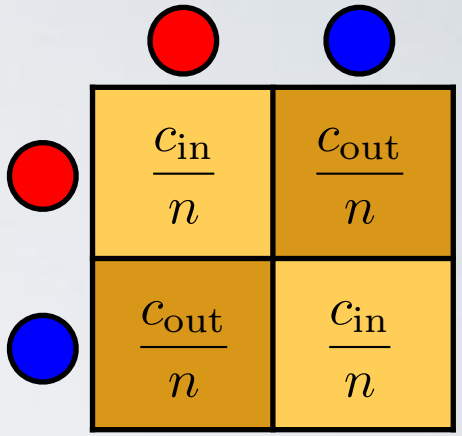
[1] technical details in Newman & Clauset (2015) arxiv:1507.04001

networks with planted structure

does this method recover known structure in synthetic data?

networks with planted structure

does this method recover known structure in synthetic data?



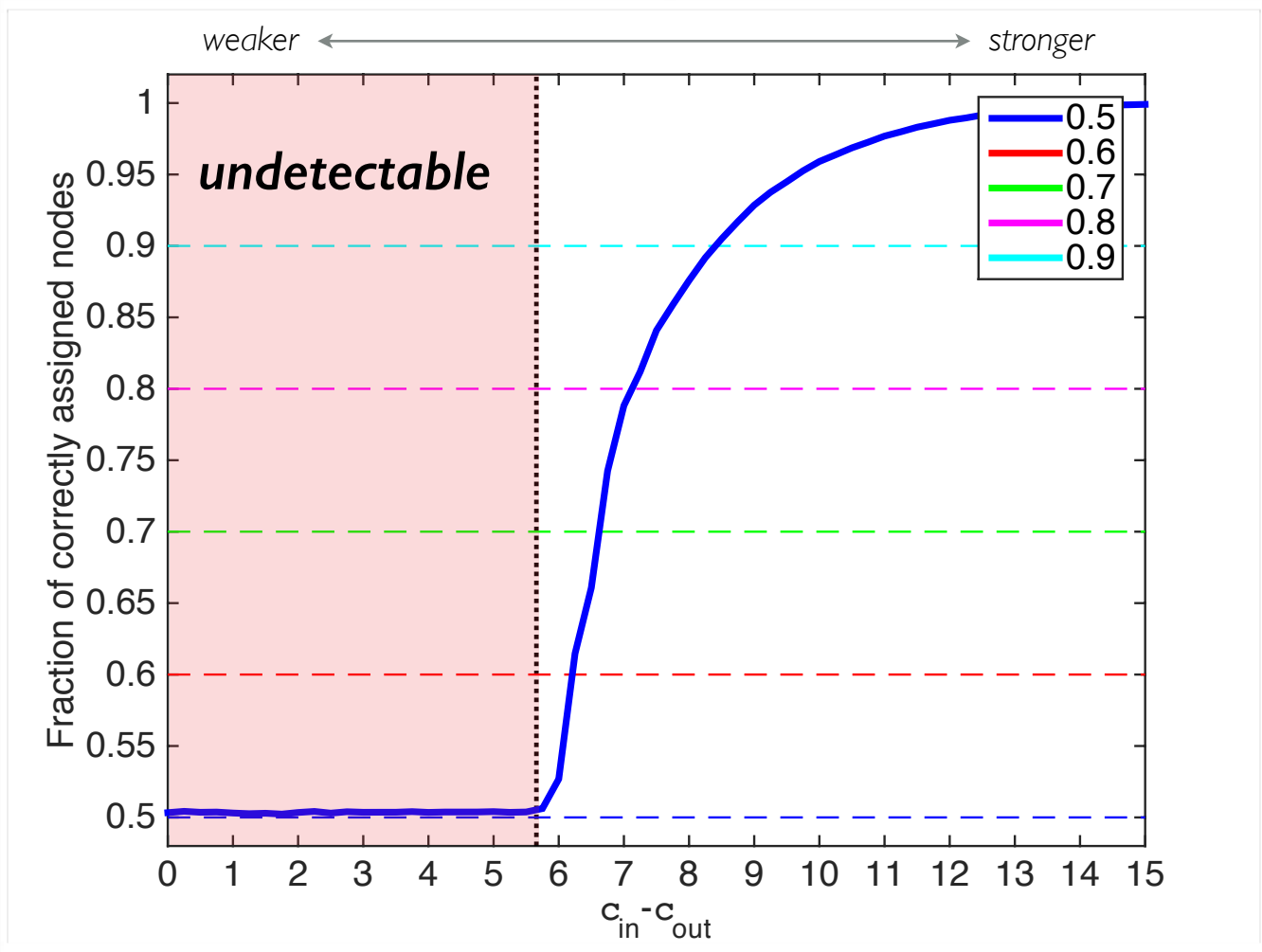
- use SBM to generate *planted partition* networks, with $k = 2$ equal-sized groups and mean degree $c = (c_{\text{in}} + c_{\text{out}})/2$
- assign metadata with variable correlation $\rho \in [0.5, 0.9]$ to true group labels
- vary strength of partition $c_{\text{in}} - c_{\text{out}}$
- when $c_{\text{in}} - c_{\text{out}} \leq \sqrt{2(c_{\text{in}} + c_{\text{out}})}$, *no structure-only* algorithm can recover the planted communities better than chance (the *detectability threshold*, which is a phase transition)

[1] Decelle, Krzakala, Moore & Zdeborova (2011)

networks with planted structure

let mean degree $c = 8$

- when $\rho = 0.5$, metadata isn't useful and we recover regular SBM behavior



[1] $n = 10\,000$

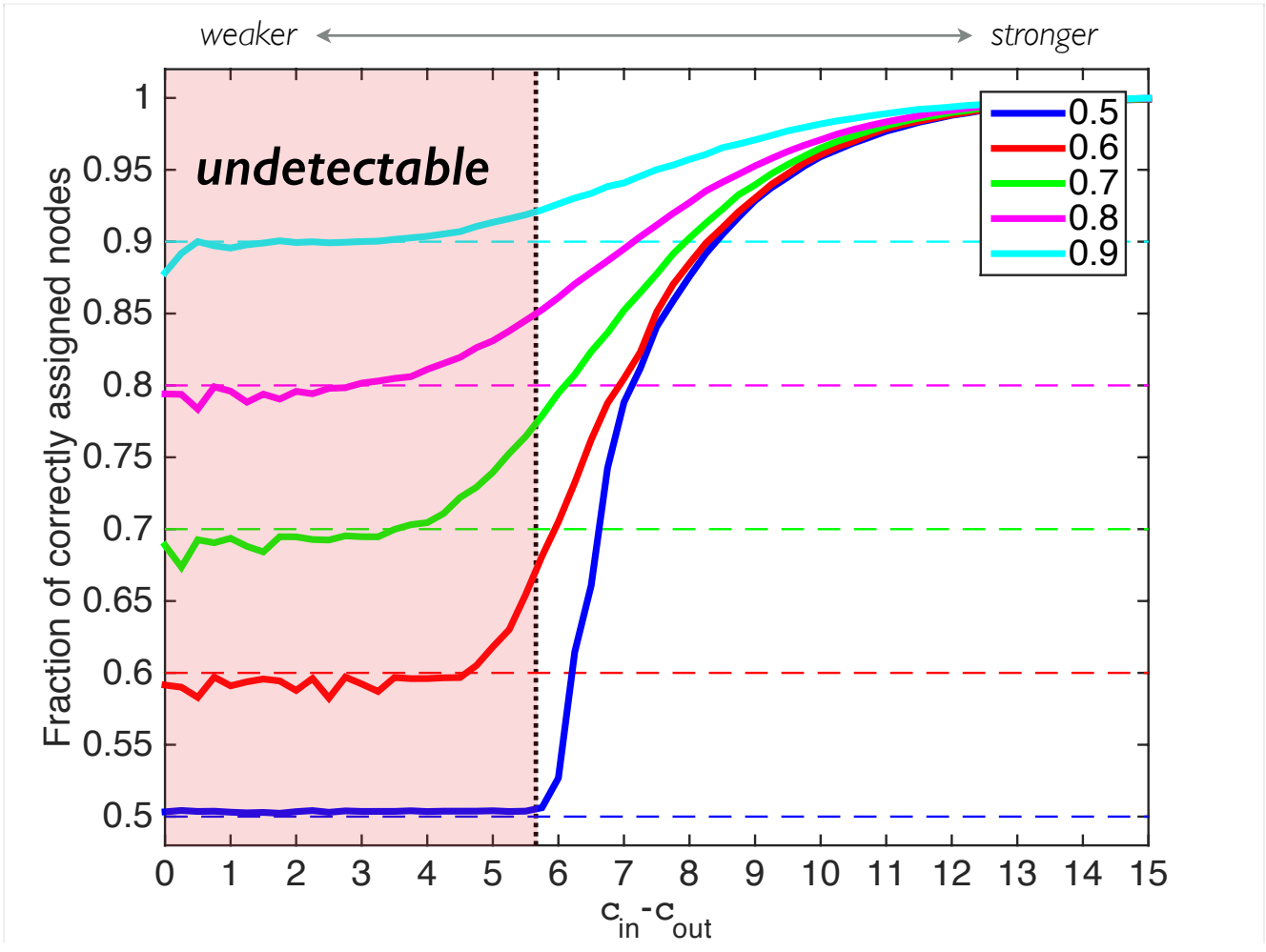
networks with planted structure

let mean degree $c = 8$

- when $\rho = 0.5$, metadata isn't useful and we recover regular SBM behavior
- when metadata correlates with true groups, $\rho > 0.5$ accuracy is better than either metadata or SBM alone

metadata + SBM performs better than either

- ***any algorithm without metadata, or***
- ***metadata alone.***



real-world networks

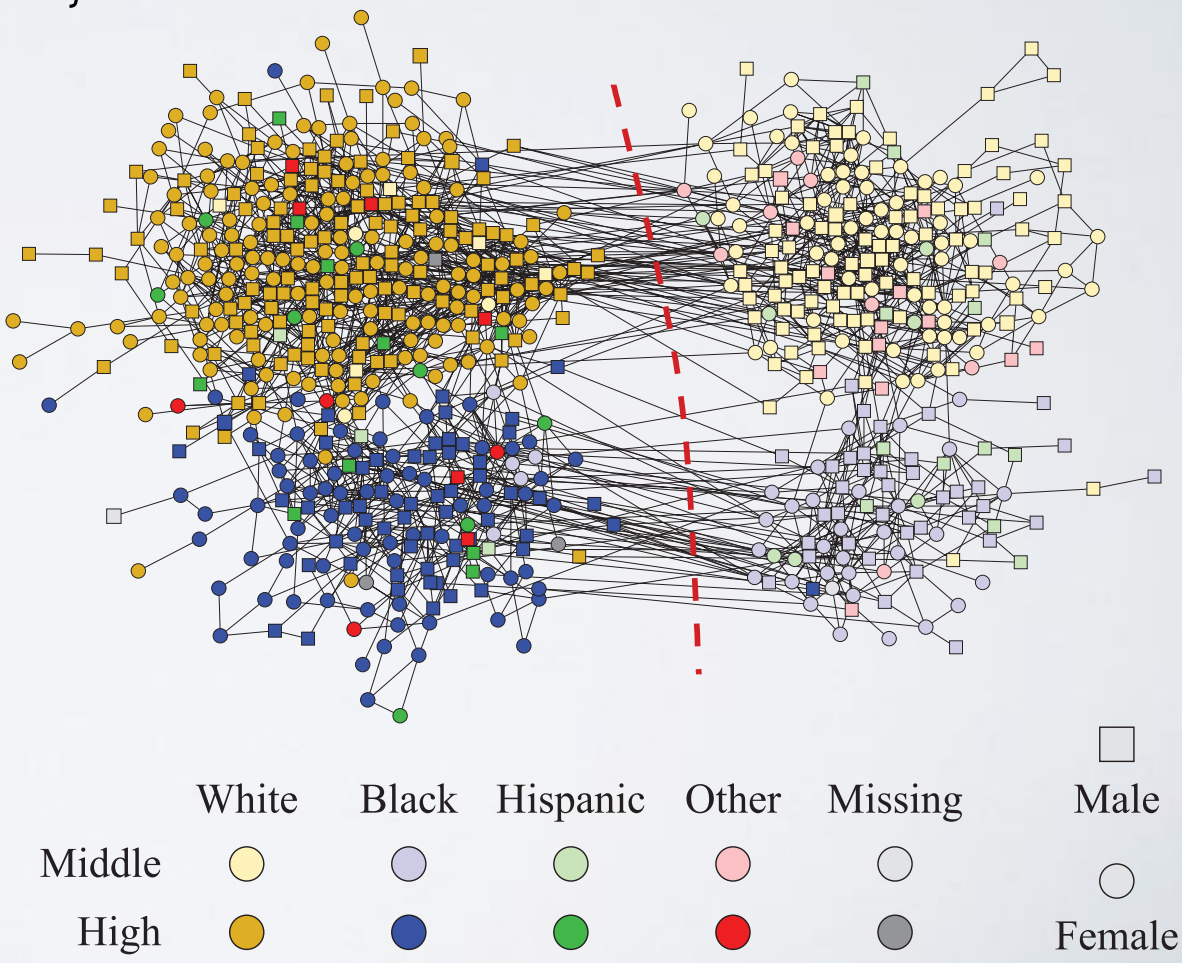
real-world networks

1. **high school social network**: 795 students in a medium-sized American high school and its feeder middle school
2. **marine food web**: predator-prey interactions among 488 species in Weddell Sea in Antarctica
3. **Malaria gene recombinations**: recombination events among 297 *var* genes
4. **Facebook friendships**: online friendships among 15,126 Harvard students and alumni
5. **Internet graph**: peering relations among 46,676 Autonomous Systems

real-world networks

1. **high school social network**: 795 students in a medium-sized American high school and its feeder middle school

- $\mathbf{x} = \{$grade 7-12, ethnicity, gender$\}$



[1] Add Health network data, designed by Udry, Bearman & Harris

real-world networks

1. **high school social network**: 795 students in a medium-sized American high school and its feeder middle school

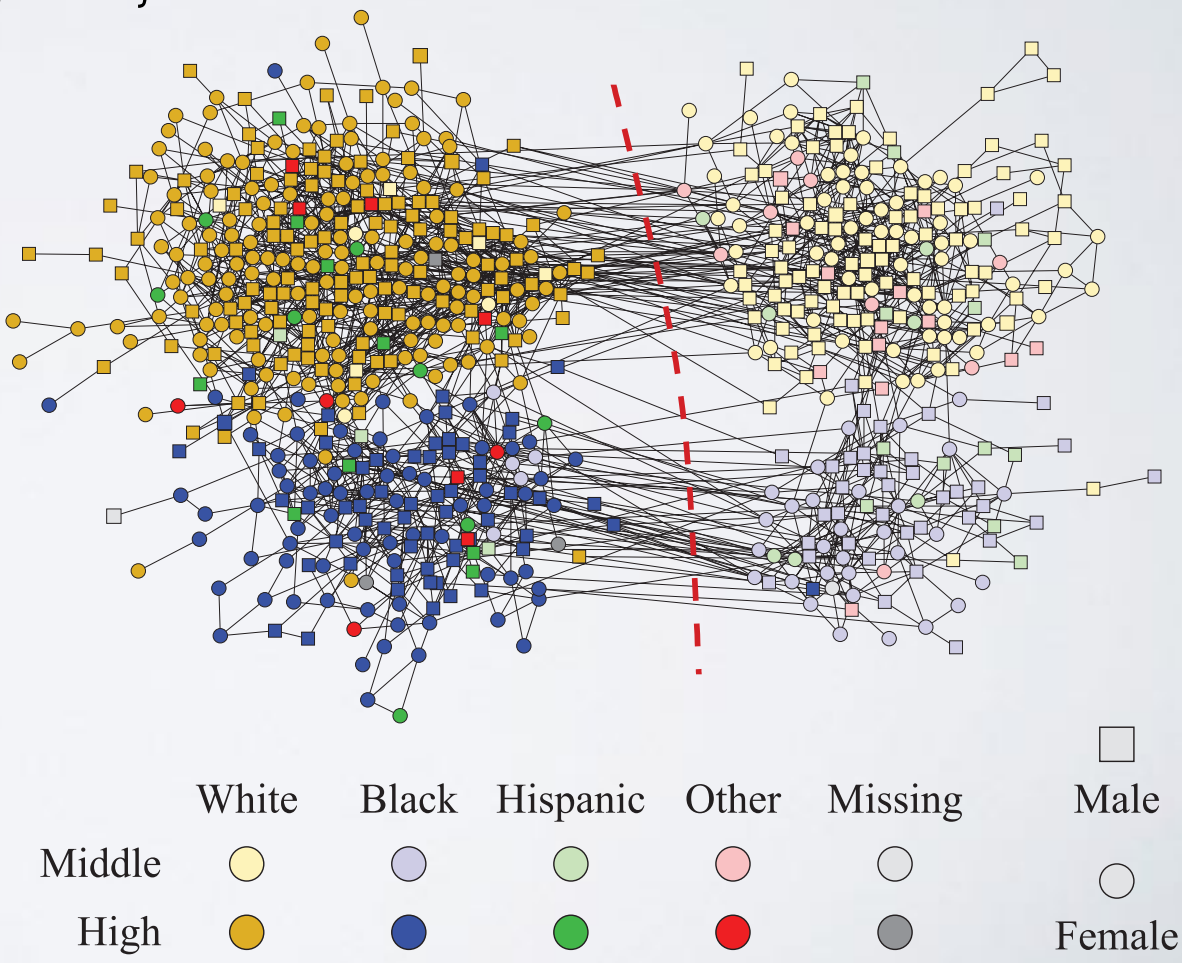
- $\mathbf{x} = \{$**grade 7-12**, ethnicity, gender$\}$

- method finds a good partition between high-school and middle-school

  $\mathrm{NMI} = 0.881$

- without metadata:

  $\mathrm{NMI} \in [0.105, 0.384]$



[1] Add Health network data, designed by Udry, Bearman & Harris

real-world networks

1. **high school social network**: 795 students in a medium-sized American high school and its feeder middle school

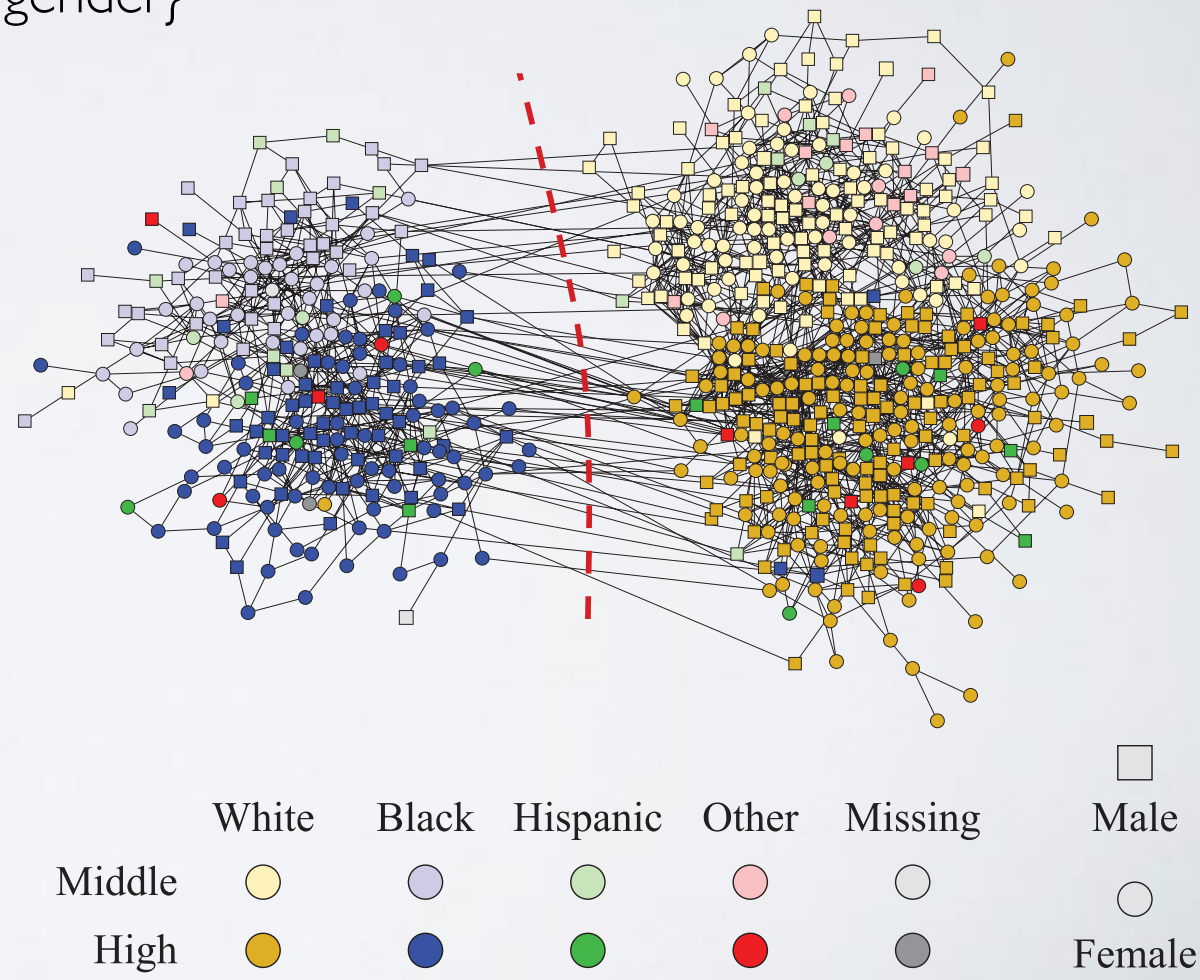
- $\mathbf{x} = \{$grade 7-12, **ethnicity**, gender$\}$

- method finds a good partition between blacks and whites (with others scattered among)

  $\mathrm{NMI} = 0.820$

- without metadata:

  $\mathrm{NMI} \in [0.120, 0.239]$



[1] Add Health network data, designed by Udry, Bearman & Harris

real-world networks

1. **high school social network**: 795 students in a medium-sized American high school and its feeder middle school

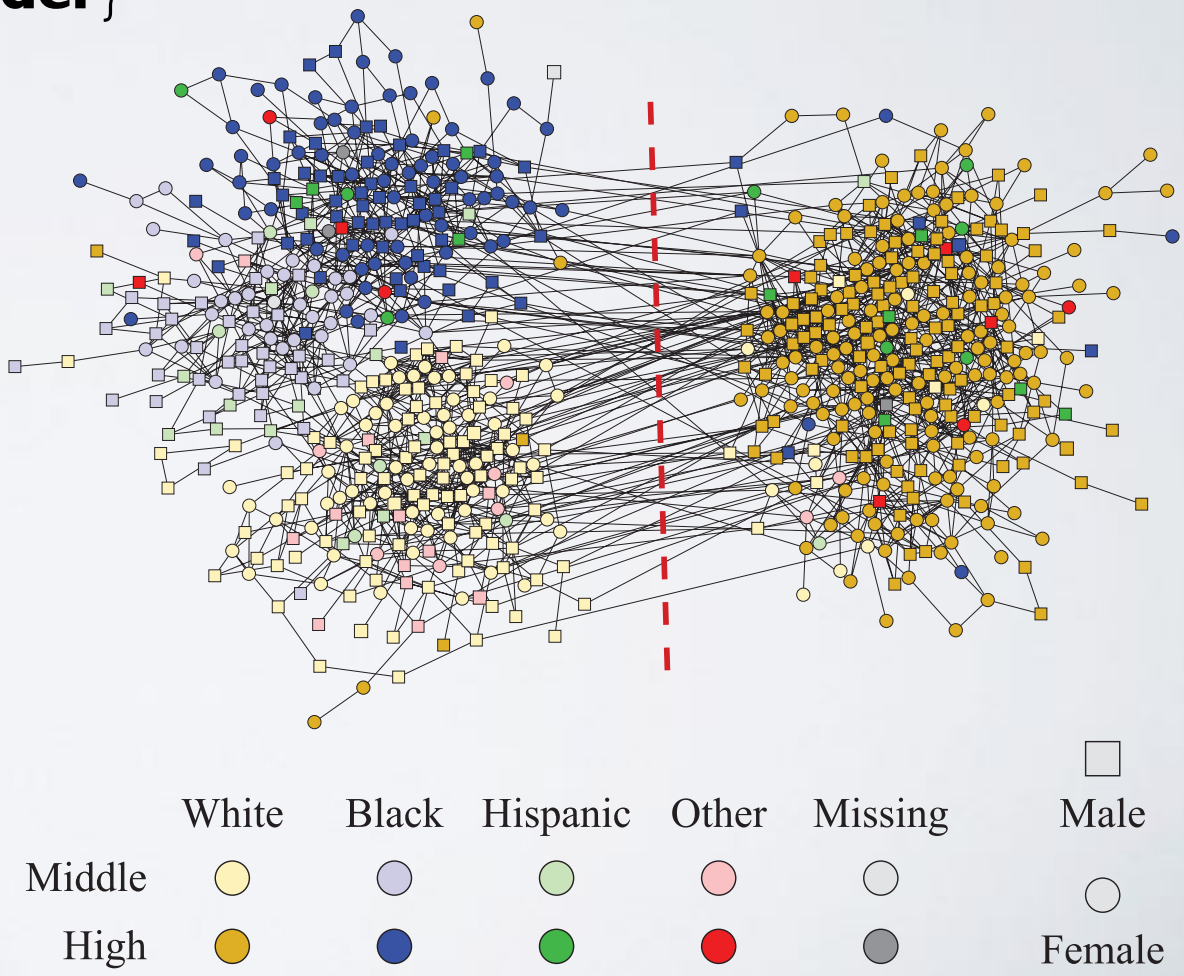
- $\mathbf{x} = \{\text{grade 7-12, ethnicity, } \mathbf{gender}\}$

- method finds no good partition between males/females.
  instead, chooses a mixture of grade/ethnicity partitions

  $$\text{NMI} = 0.003$$

- without metadata:

  $$\text{NMI} \in [0.000, 0.010]$$

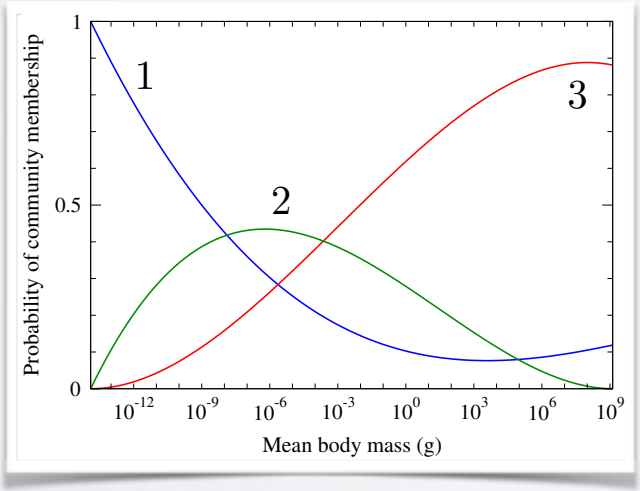


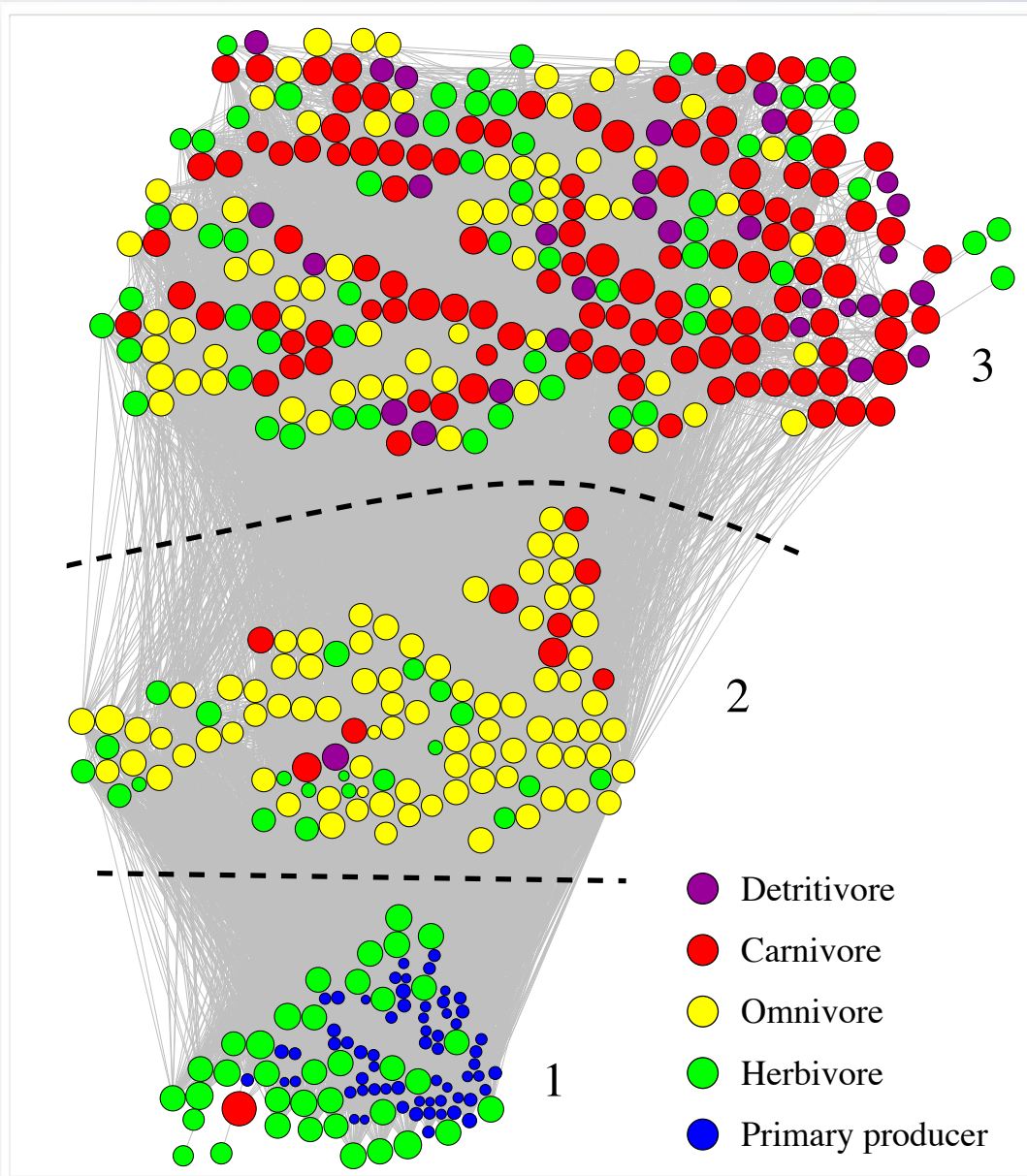
[1] Add Health network data, designed by Udry, Bearman & Harris

real-world networks

2. **marine food web**: predator-prey interactions among 488 species in Weddell Sea in Antarctica

   - $\mathbf{x} = \{$**species body mass**, feeding mode, oceanic zone$\}$
   - partition recovers known correlation between body mass, trophic level, and ecosystem role:



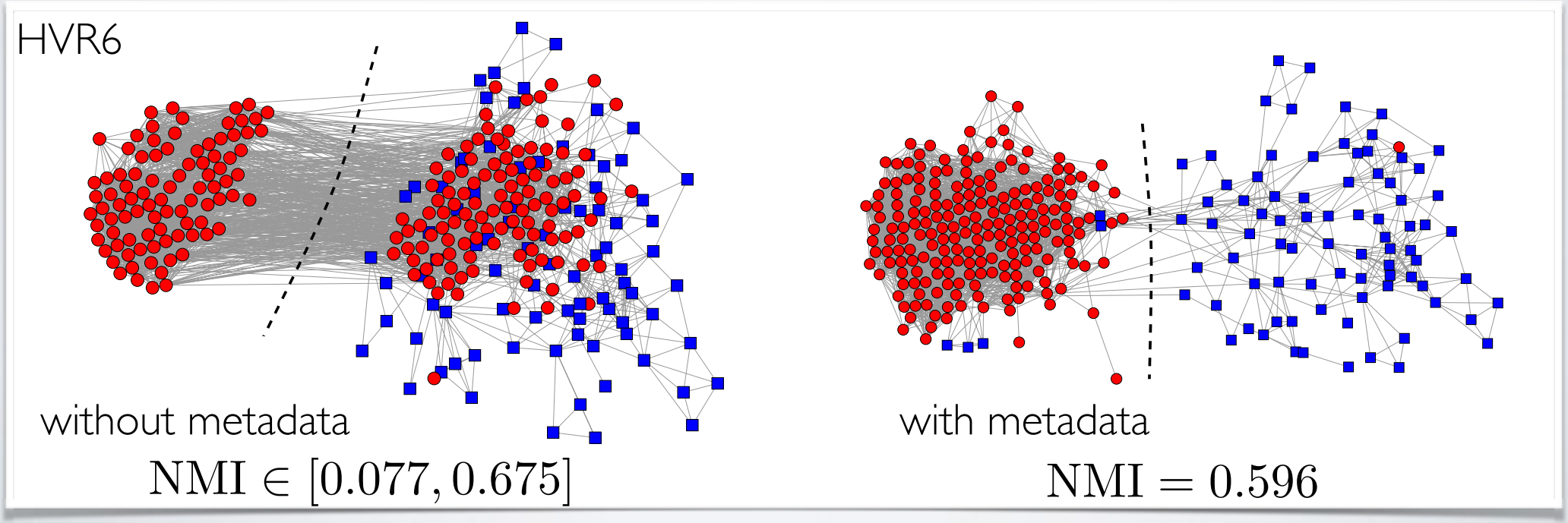


[1] here, we're using a continuous metadata model
[2] Brose et al. (2005)

real-world networks

3. **Malaria gene recombinations**: recombination events among 297 *var* genes

- $\mathbf{x} = \{$**Cys-PoLV labels for HVR6 region**$\}$
- with metadata, partition discovers correlation with Cys labels (which are associated with severe disease)

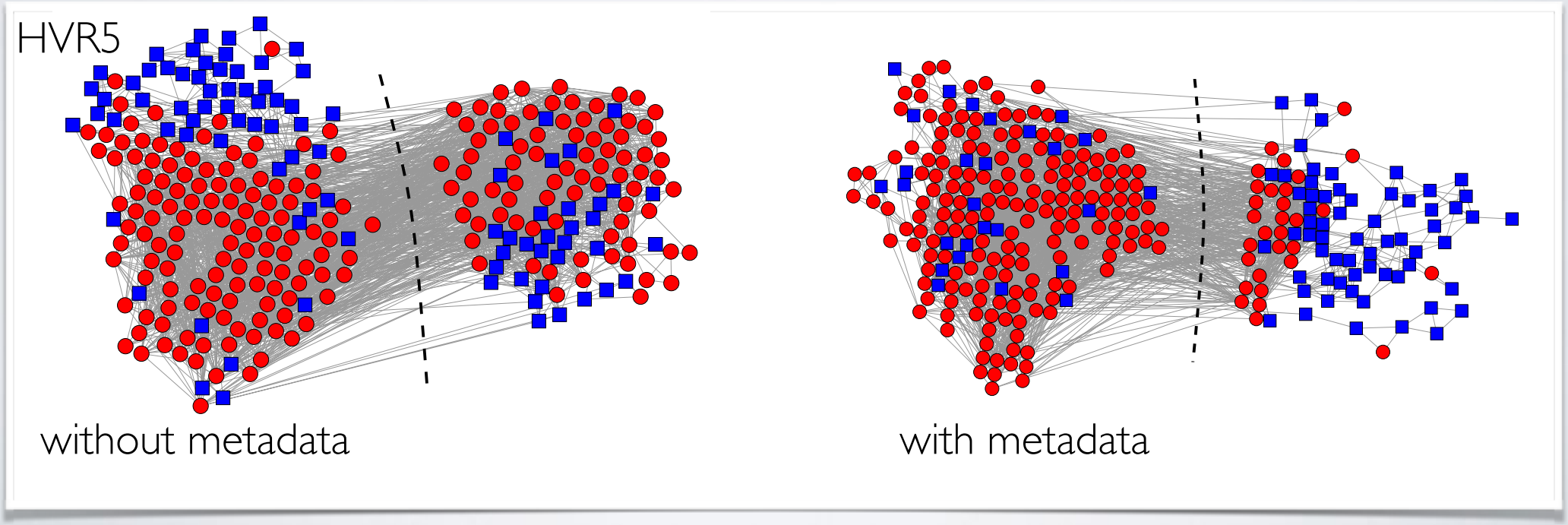


[1] Larremore, Clauset & Buckee (2013)

real-world networks

3. **Malaria gene recombinations**: recombination events among 297 *var* genes

- $\mathbf{x} = \{$**Cys-PoLV labels for HVR6 region**$\}$
- on adjacent region of gene, we find Cys-PoLV labels correlate with recombinant structure here, too



[1] Larremore, Clauset & Buckee (2013)

the ground truth about metadata

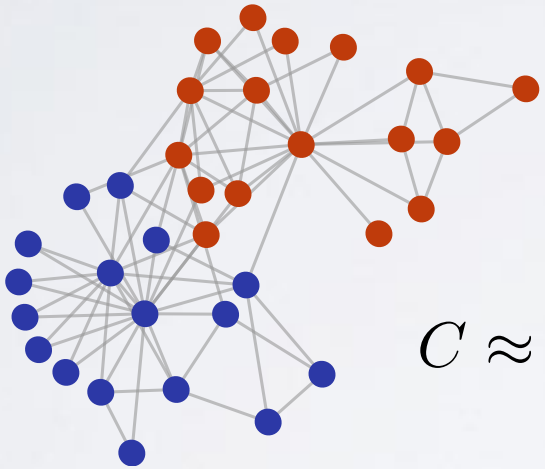
what is the goal of community detection?

network G + method $f \rightarrow$ communities $C = f(G)$ vs. M metadata

the ground truth about metadata

what is the goal of community detection?

network G + method $f \rightarrow$ communities $C = f(G)$ vs. M metadata



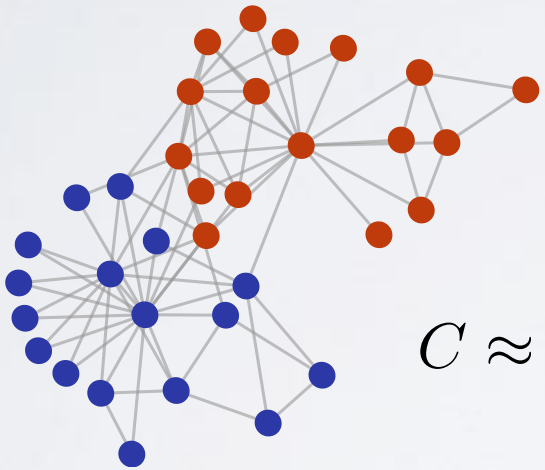
$C \approx M$

"this method works!"

the ground truth about metadata

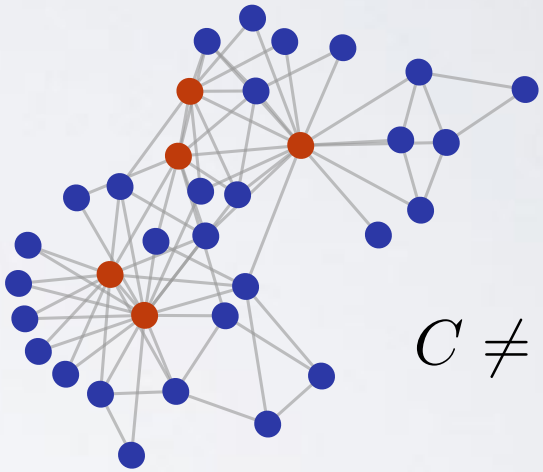
what is the goal of community detection?

network G + method $f \rightarrow$ communities $C = f(G)$ vs. M metadata

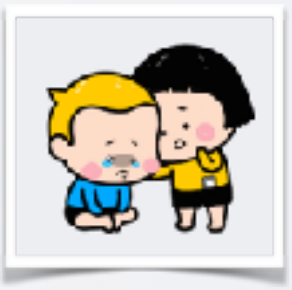


$C \approx M$

"this method works!"



$C \neq M$



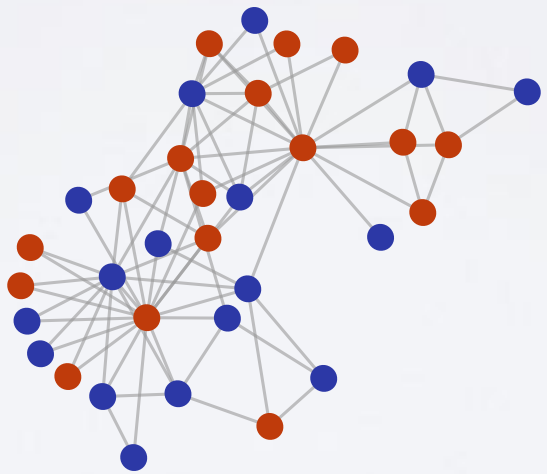
"this method stinks!"

the ground truth about metadata

what is the goal of community detection?

there are 4 indistinguishable reasons why we might find $f(G) = C \neq M$:

1. metadata M are unrelated to network structure G

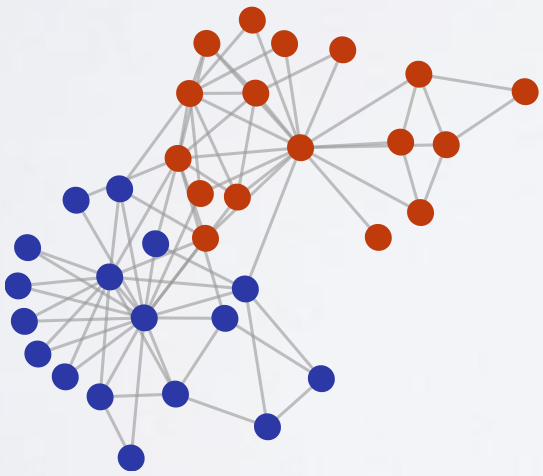


the ground truth about metadata

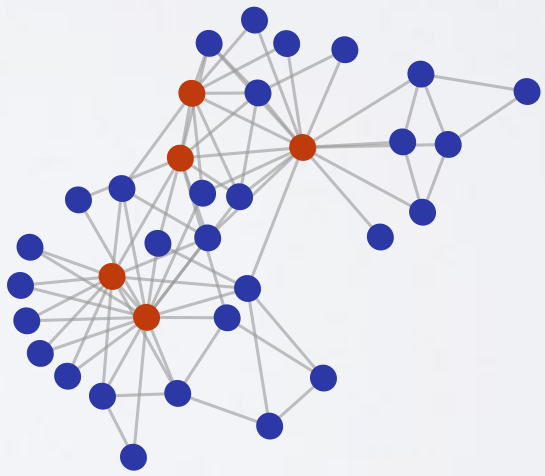
what is the goal of community detection?

there are 4 indistinguishable reasons why we might find $f(G) = C \neq M$:

1. metadata M are unrelated to network structure G
2. metadata M and communities C capture different aspects of structure



social groups



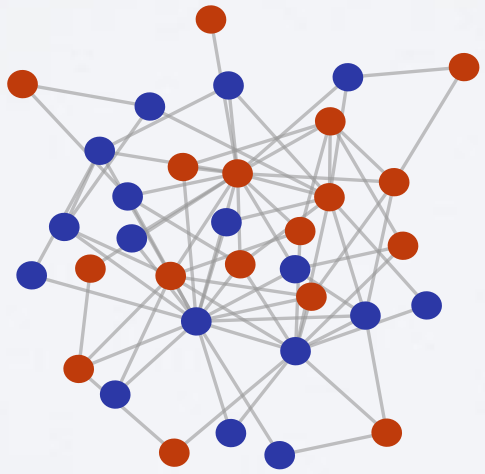
leaders and followers

the ground truth about metadata

what is the goal of community detection?

there are 4 indistinguishable reasons why we might find $f(G) = C \neq M$:

1. metadata M are unrelated to network structure G
2. metadata M and communities C capture different aspects of structure
3. network G has no community structure



the ground truth about metadata

what is the goal of community detection?

there are 4 indistinguishable reasons why we might find $f(G) = C \neq M$:

1. metadata M are unrelated to network structure G
2. metadata M and communities C capture different aspects of structure
3. network G has no community structure
4. algorithm f is bad

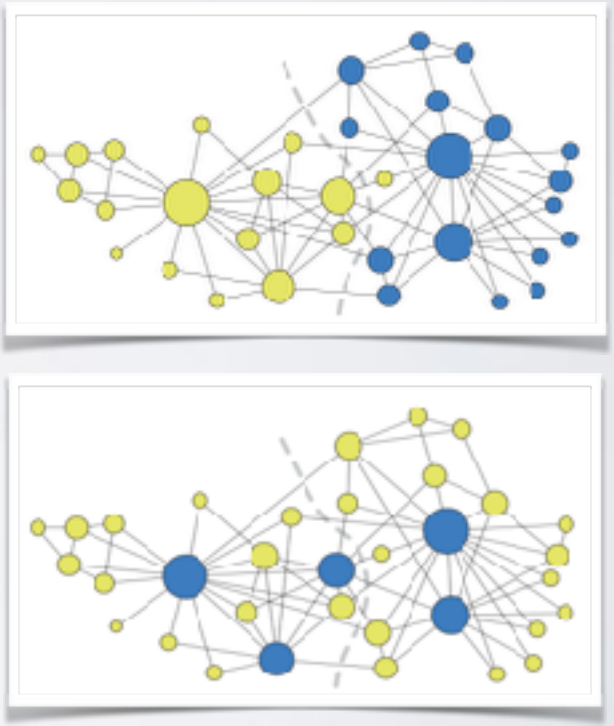
"this method stinks!"

theorems for community detection

theorems for community detection

1. Theorem: *no bijection between ground truth and communities*

$g(T) \rightarrow G \leftarrow g'(T')$ 2 different processes, on 2 different ground truths, can create the same observed network

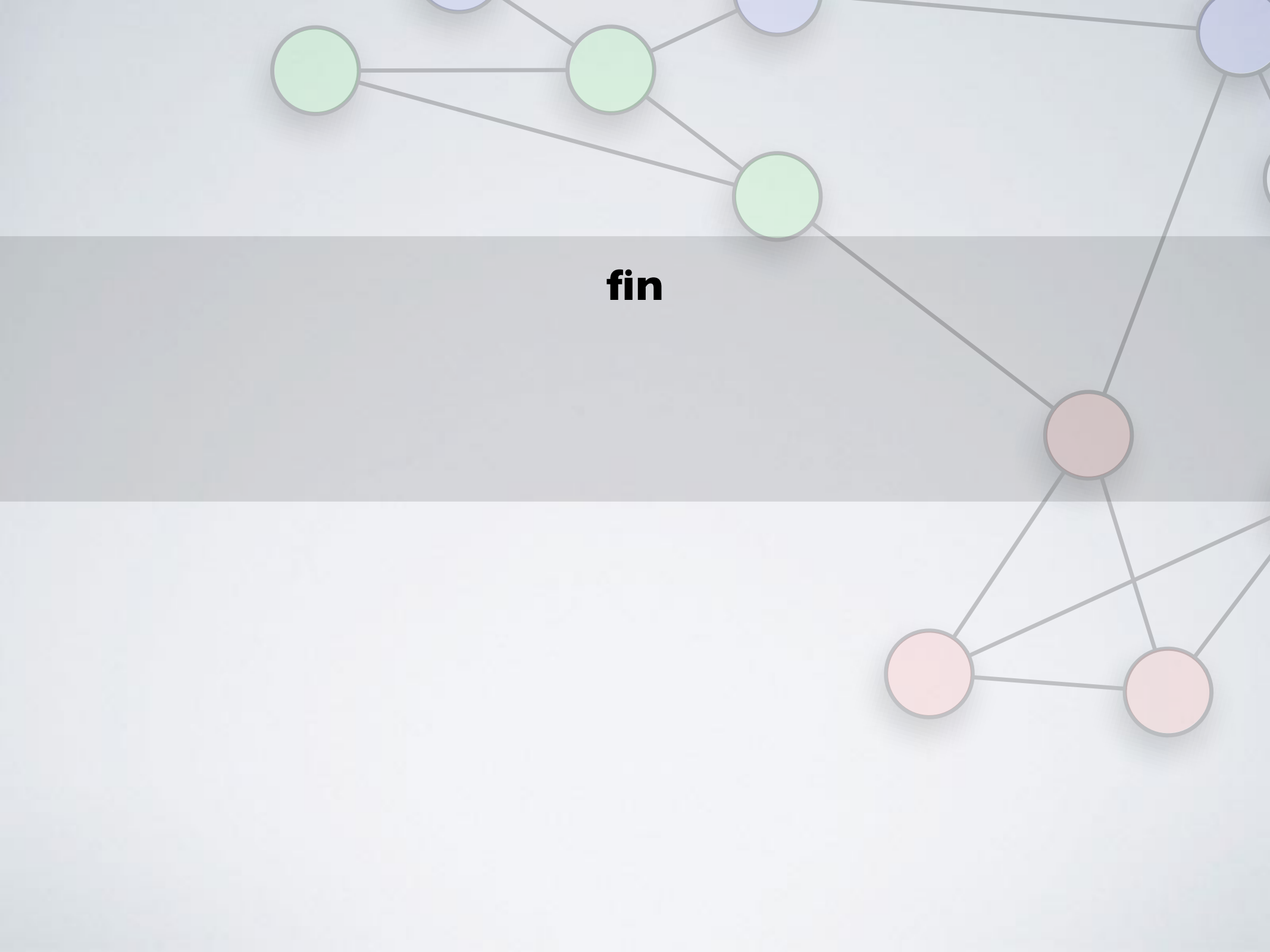


2. Theorem: *No Free Lunch in community detection*

no algorithm f has better performance than any other algorithm f', when averaged over all possible inputs $\{G\}$

$\rightarrow$ *good* performance comes from matching algorithm f to its preferred subclass of networks $\{G'\} \subset \{G\}$

[1] performance defined as adjusted mutual information (AMI), which is like the normalized mutual information, but adjusted for expected values
[2] original NFL theorem: Wolpert, *Neural Computation* (1996)
[3] proofs of these theorems is in Peel, Larremore, Clauset (2016)


fin

real-world networks

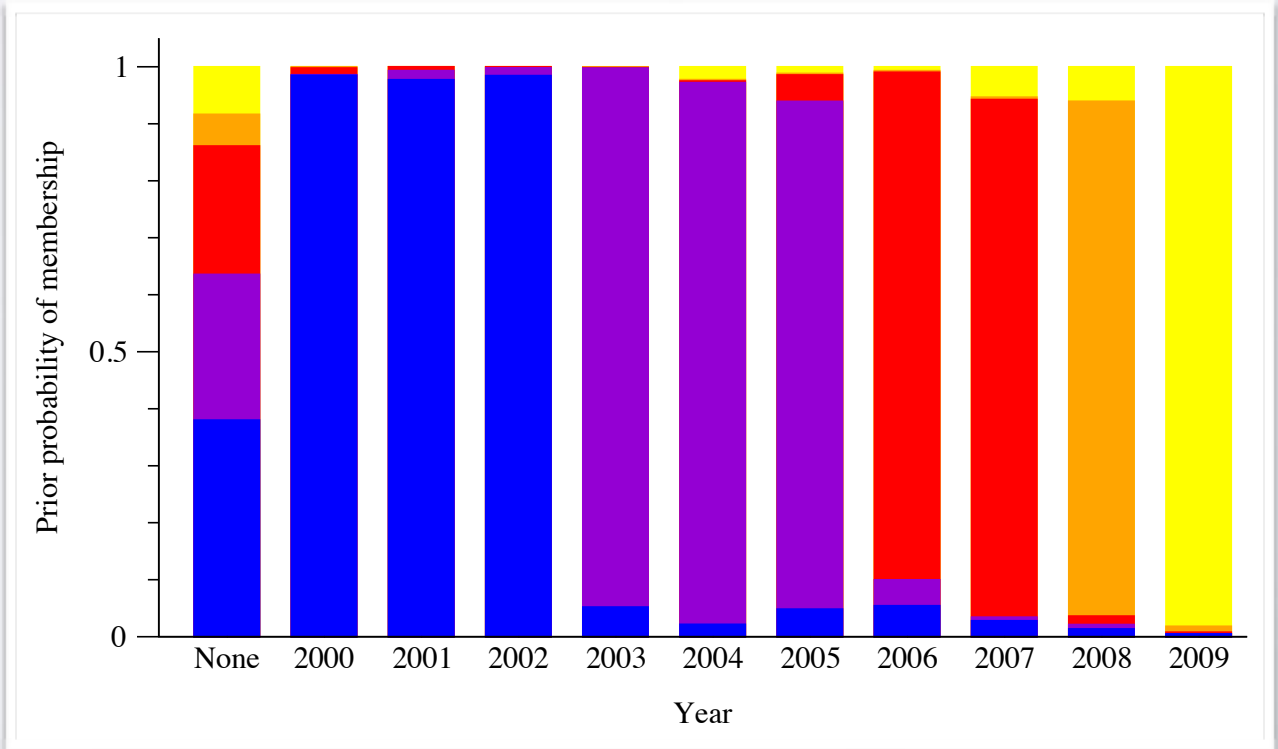
4. **Facebook friendships**: online friendships among 15,126 Harvard students and alumni (in Sept. 2005)

- $\mathbf{x} = \{$**graduation year**, dormitory$\}$
- method finds a good partition between alumni, recent graduates, upperclassmen, sophomores, and freshmen

$$\mathrm{NMI} = 0.668$$

- without metadata:

$$\mathrm{NMI} \in [0.573, 0.641]$$



[1] Traud, Mucha & Porter (20012)

real-world networks

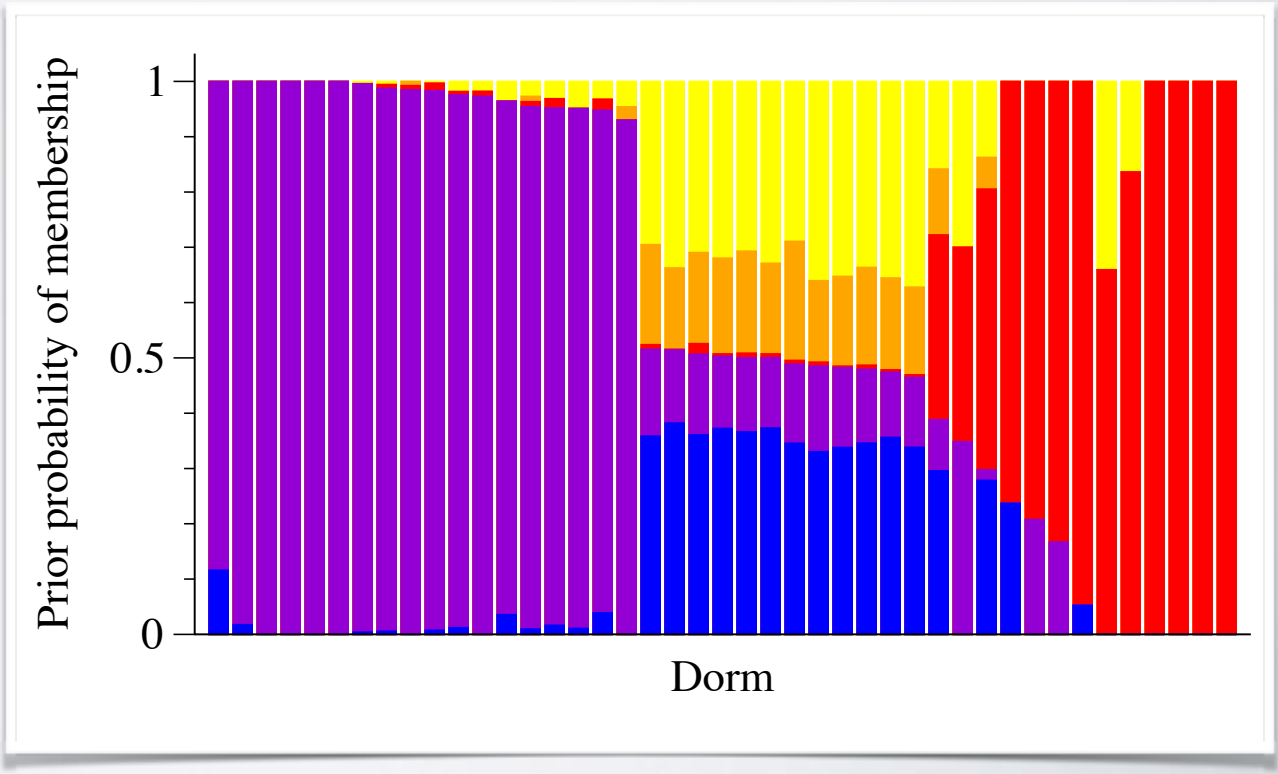
4. **Facebook friendships**: online friendships among 15,126 Harvard students and alumni (in Sept. 2005)

- $\mathbf{x} = \{\text{graduation year, } \textbf{dormitory}\}$
- method finds a good partition among the dorms

  $\text{NMI} = 0.255$

- without metadata:

  $\text{NMI} \in [0.074, 0.224]$



[1] Traud, Mucha & Porter (20012)