Lecture 9a:
Annotated Networks and … Trouble

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

what is community structure?

large-scale structure = community structure

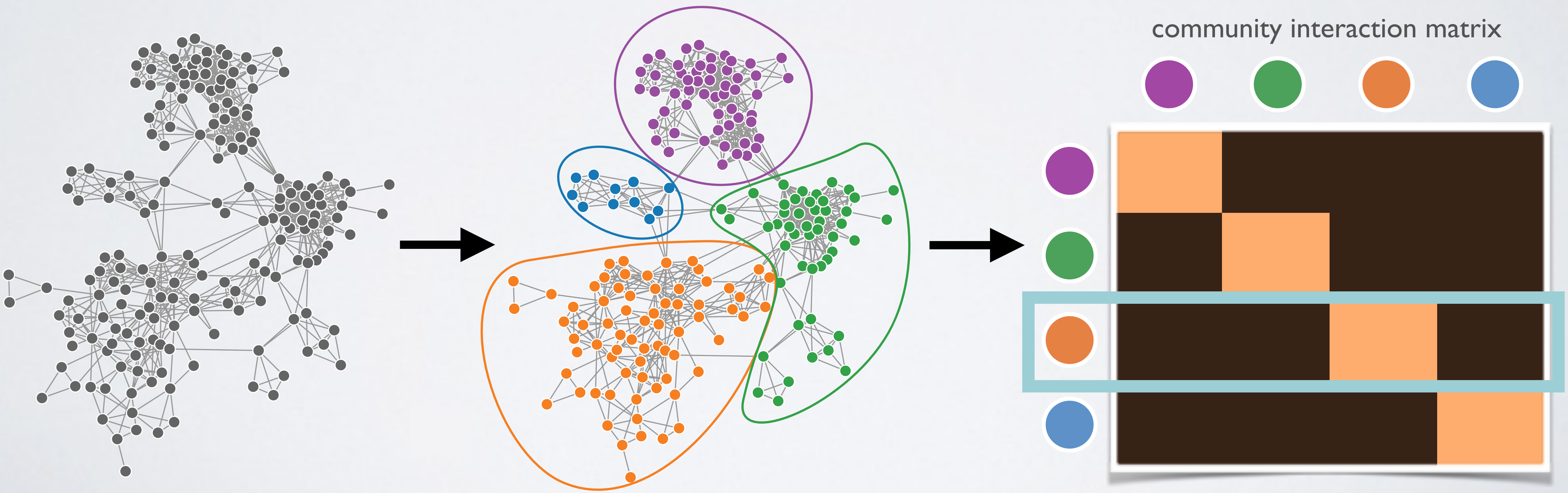
vertices with same pattern of inter-community connections

what is community structure?

large-scale structure = community structure

vertices with same pattern of inter-community connections

community detection:



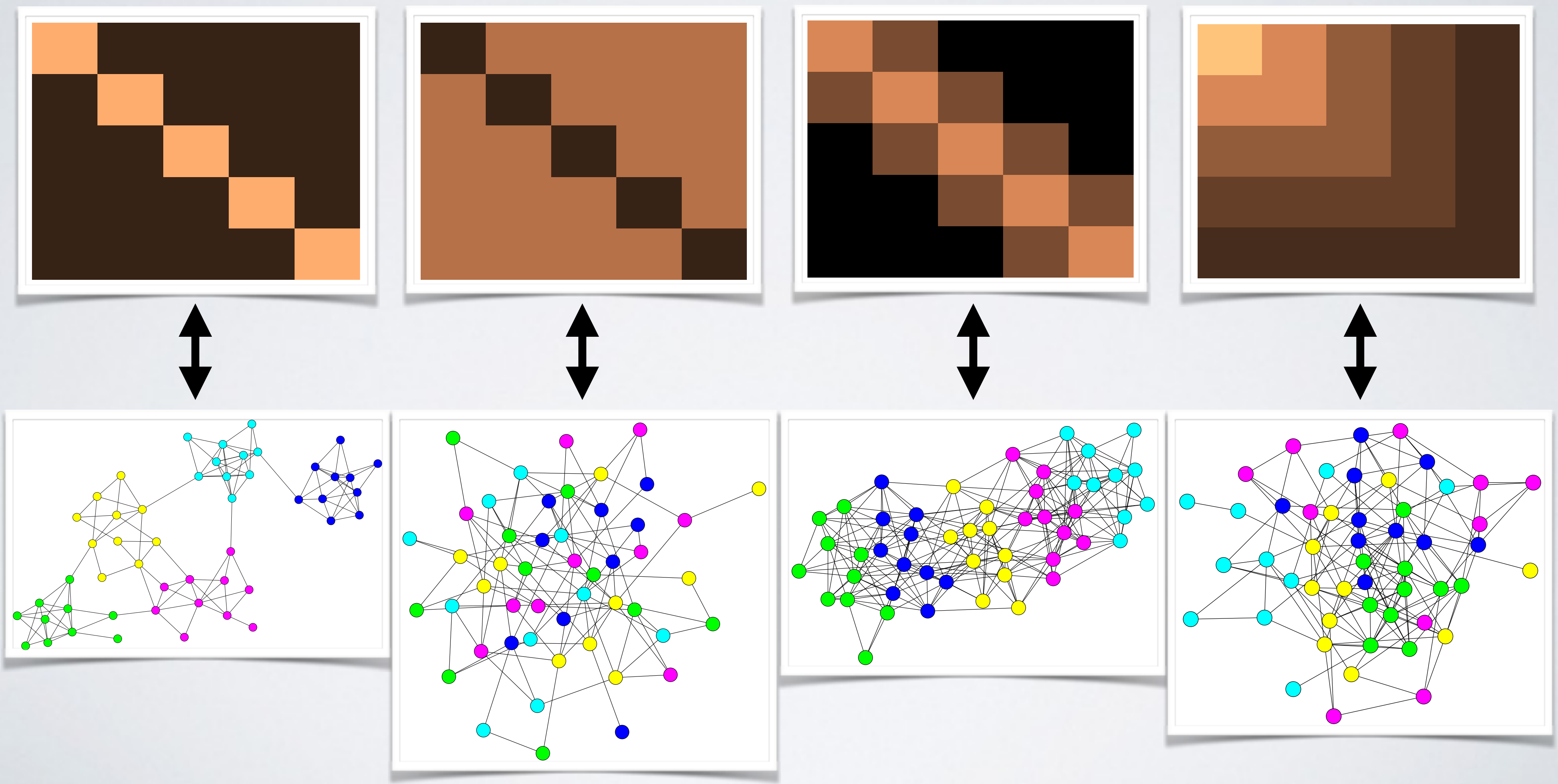
what is community structure?

assortative
edges within groups

disassortative
edges between groups

ordered
linear group hierarchy

core-periphery
dense core, sparse periphery



the trouble with community detection (part 1)

the trouble with community detection (part 1)

best understood for modularity maximization $Q = \sum_{i=1}^{k} \left[\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right]$

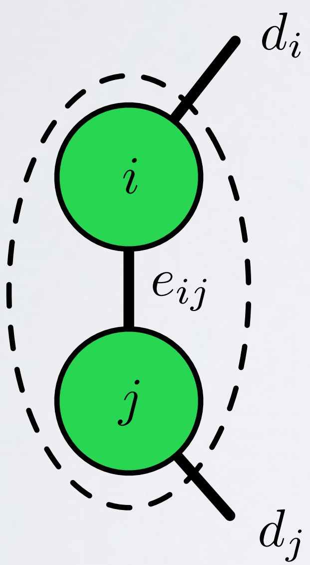
1. ***resolution limit***: in large, unweighted networks, a pair of "true" groups i, j will be incorrectly merged when maximizing Q

[1] Girvan & Newman (2002)
[2] Fortunato & Barthelemey (2007)

the trouble with community detection (part 1)

best understood for modularity maximization $Q = \sum_{i=1}^{k} \left[\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right]$

1. ***resolution limit***: in large, unweighted networks, a pair of "true" groups i, j will be incorrectly merged when maximizing Q



$$\Delta Q_{ij} = \frac{e_{ij}}{m} - \frac{d_i d_j}{2m^2}$$

merging is favored when

$$e_{ij} > (d_i d_j)/2m = \mathrm{E}[e_{ij}]$$

[1] Girvan & Newman (2002)
[2] Fortunato & Barthelemey (2007)

the trouble with community detection (part 1)

best understood for modularity maximization $Q = \sum_{i=1}^{k} \left[\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right]$

1. ***resolution limit***: in large, unweighted networks, a pair of "true" groups i, j will be incorrectly merged when maximizing Q

2. ***extreme degeneracies***: there are an exponential number of local optima with scores close to the maximum Q

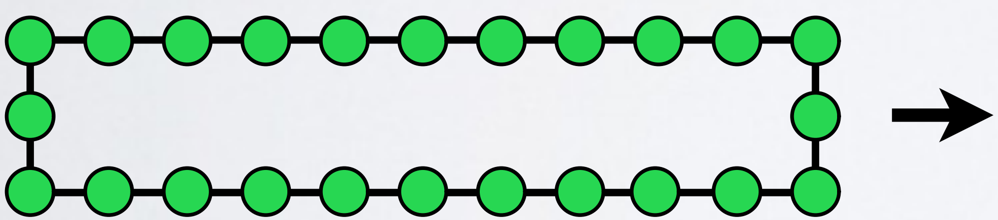
[1] Good, de Montjoye, & Clauset (2010)

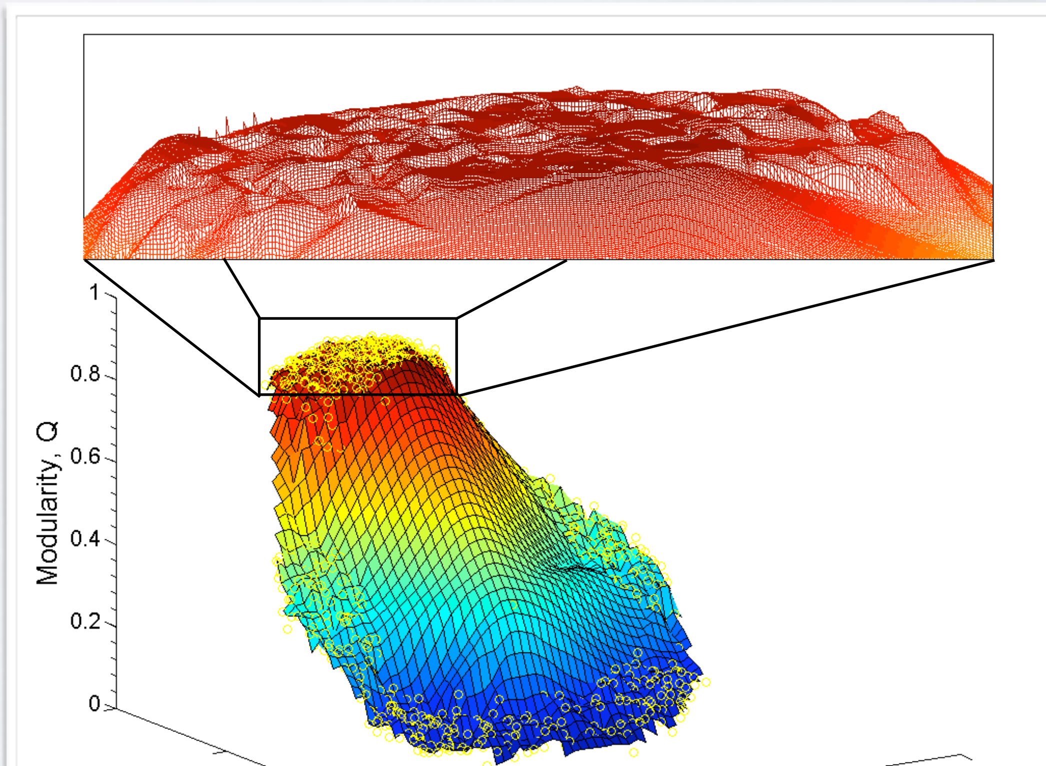
the trouble with community detection (part 1)

best understood for modularity maximization $Q = \sum_{i=1}^{k} \left[\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right]$

1. ***resolution limit***: in large, unweighted networks, a pair of "true" groups i, j will be incorrectly merged when maximizing Q

2. ***extreme degeneracies***: there are an exponential number of local optima with scores close to the maximum Q





[1] Good, de Montjoye, & Clauset (2010)

the trouble with community detection (part 1)

best understood for modularity maximization $Q = \sum_{i=1}^{k} \left[\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right]$

1. ***resolution limit***: in large, unweighted networks, a pair of "true" groups i, j will be incorrectly merged when maximizing Q
2. ***extreme degeneracies***: there are an exponential number of local optima with scores close to the maximum Q

the results of community detection should be interpreted with caution:

the "true" groups are likely merged in complicated ways

[1] Similar arguments can be constructed for other community structure objective functions, so these problems are likely fairly general.

the trouble with community detection (part 2)

the trouble with community detection (part 2)

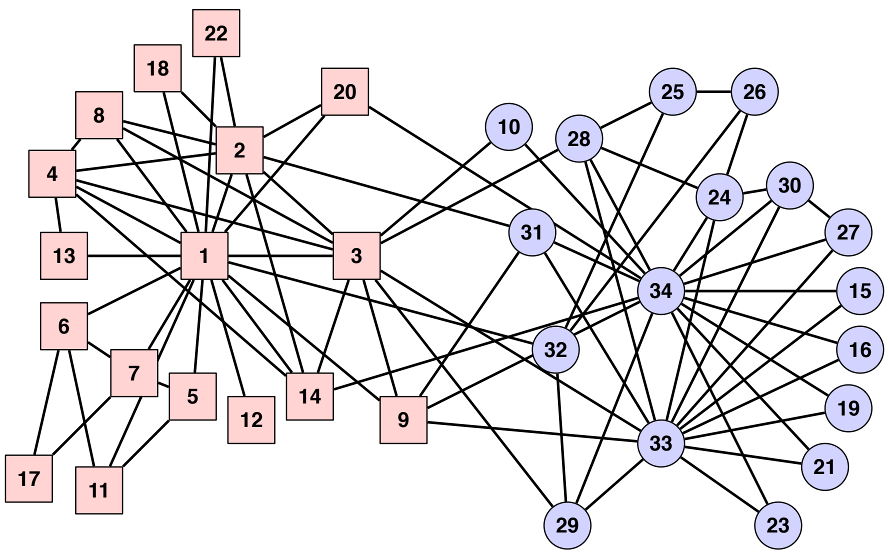
many networks include ***metadata*** on their nodes:

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.

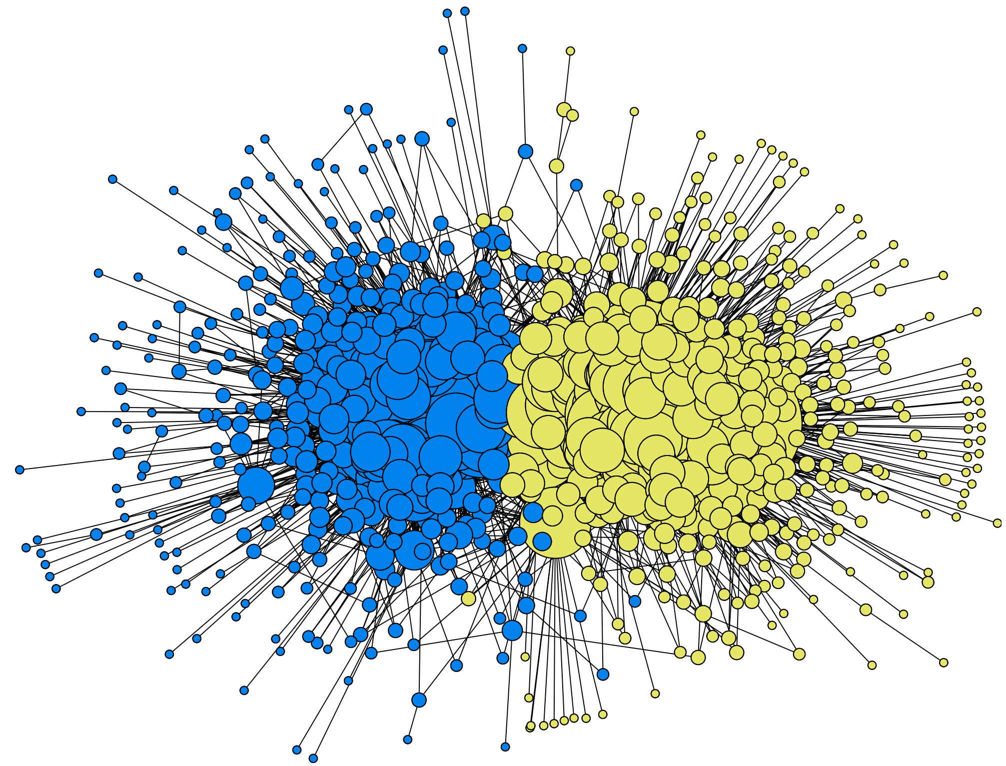
metadata $\mathbf{x}$ is often used to *evaluate the accuracy* of community detection algs.

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

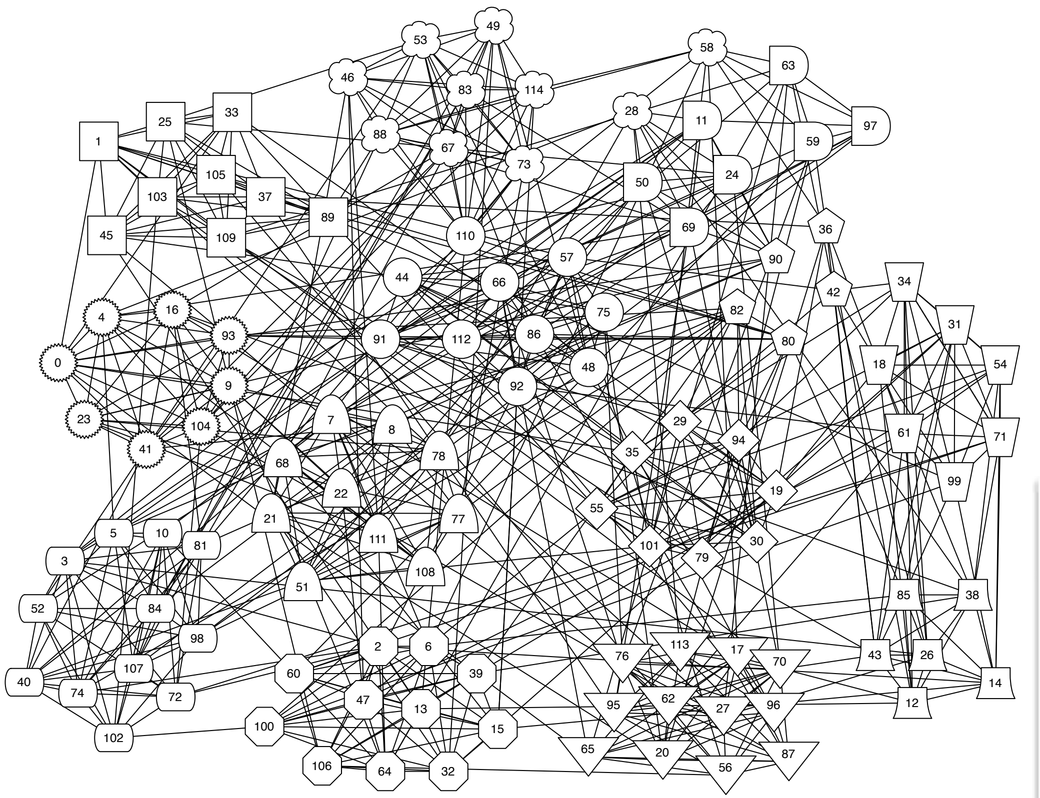


Zachary karate club



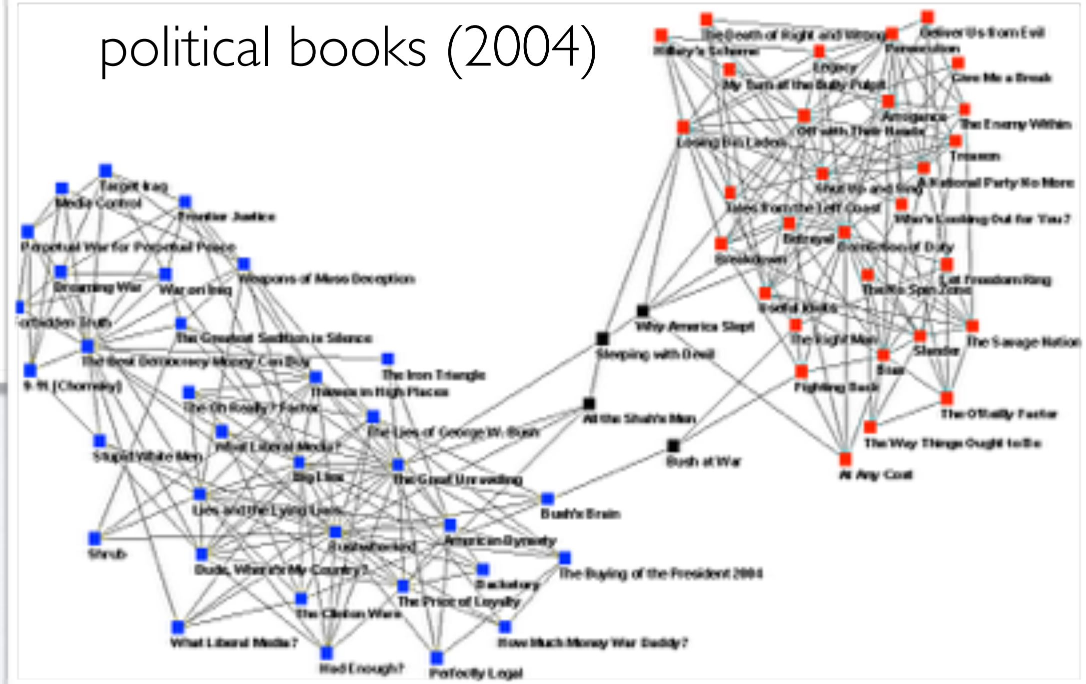
political blogs network

the trouble with community detection

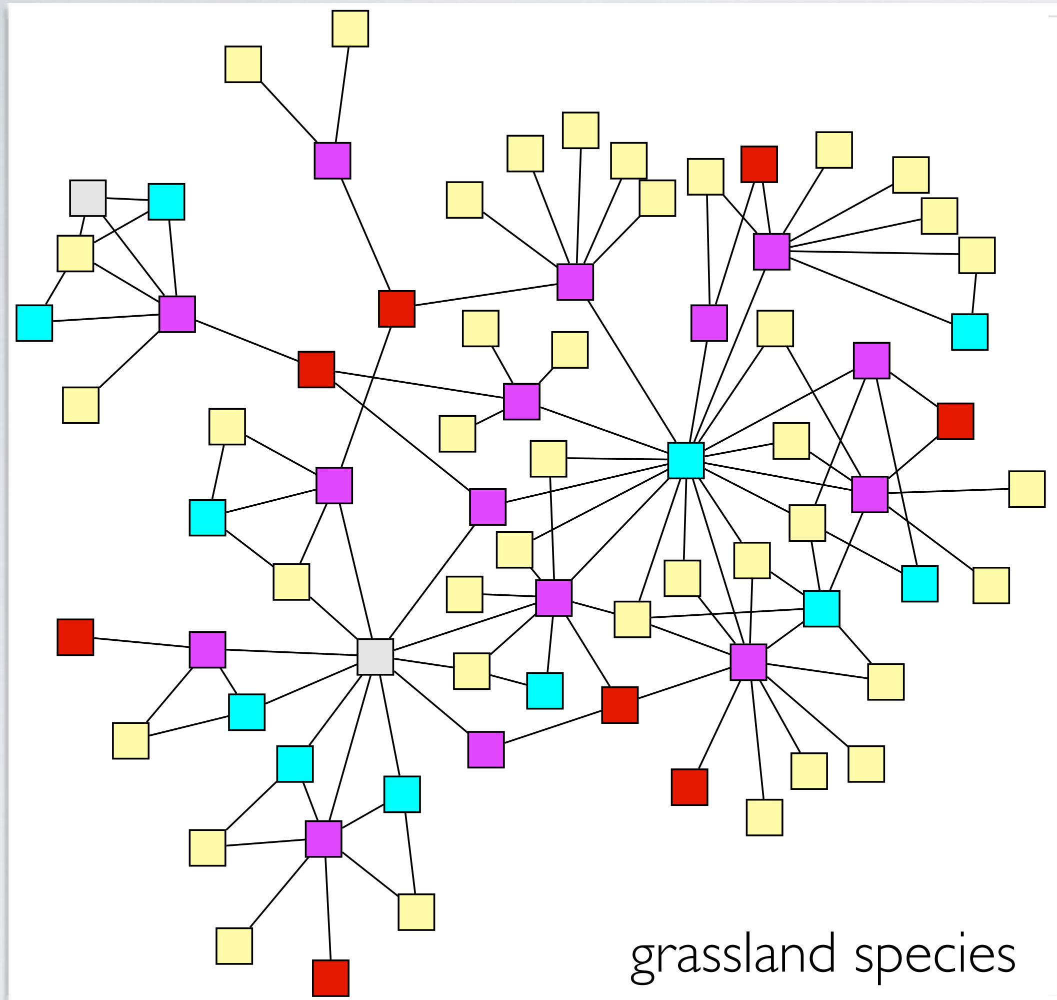


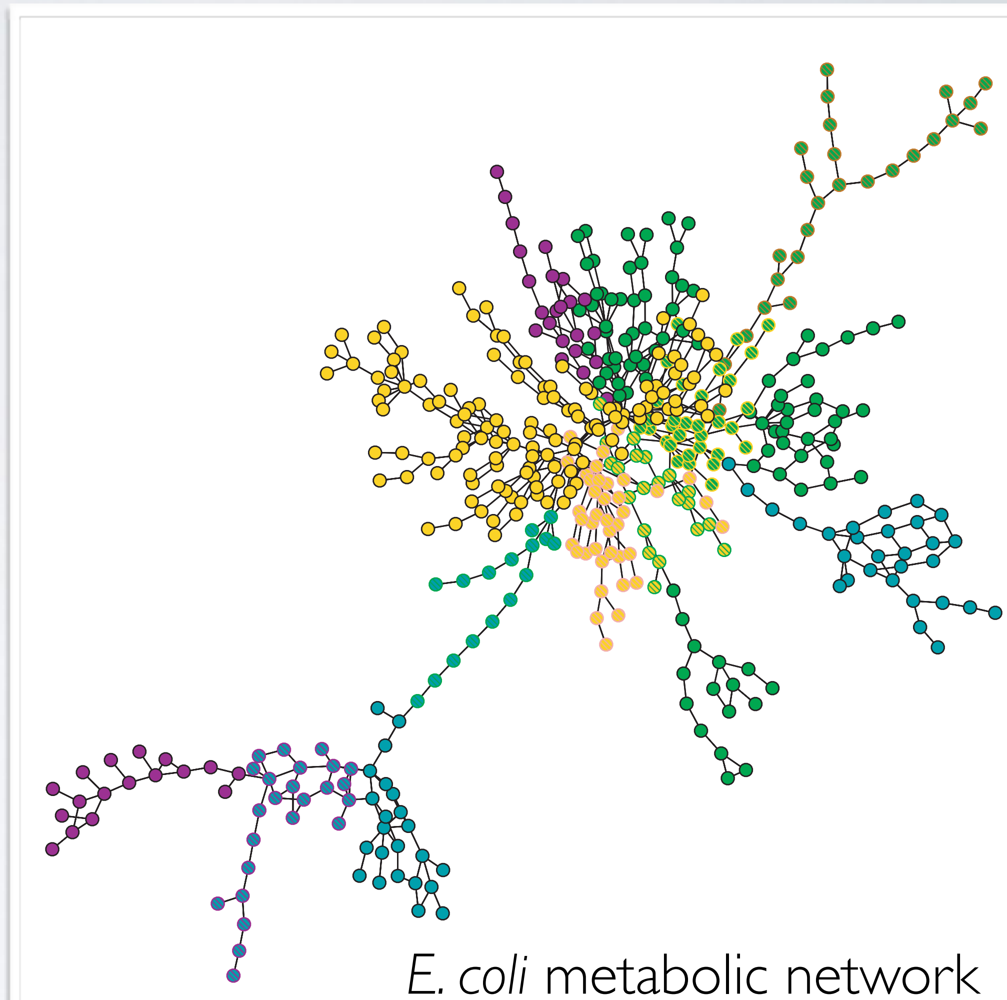
NCAA 2000 Schedule

political books (2004)



the trouble with community detection





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

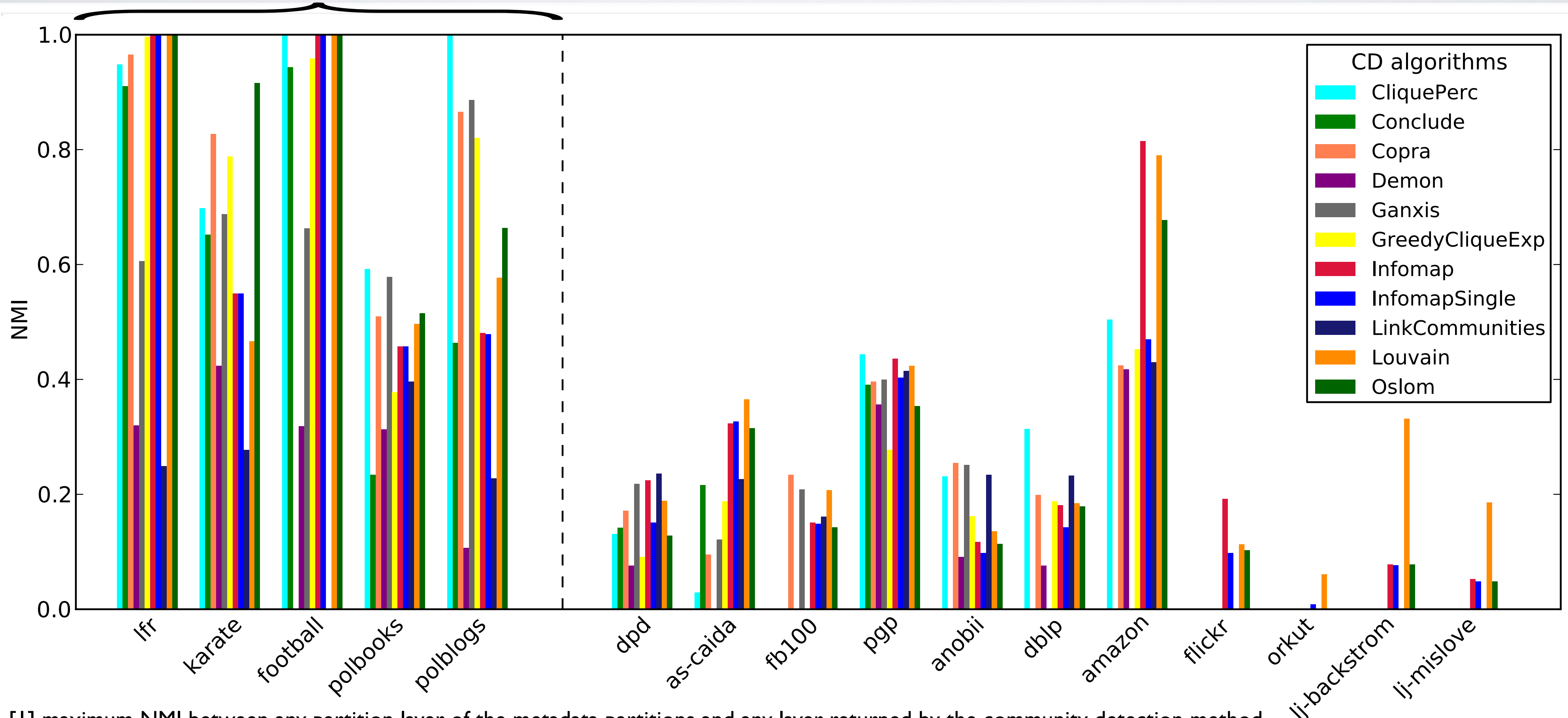
[1] fb100 is 100 networks

the trouble with community detection

Hric, Darst & Fortunato (2014)

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

"classic" data sets



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

the trouble with community detection

lies, damned lies, and community detection

"true" groups can be merged in complicated ways

best partition typically lost in exponential number of local optima

the groups we do find often don't correlate with what we think we are trying to recover

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

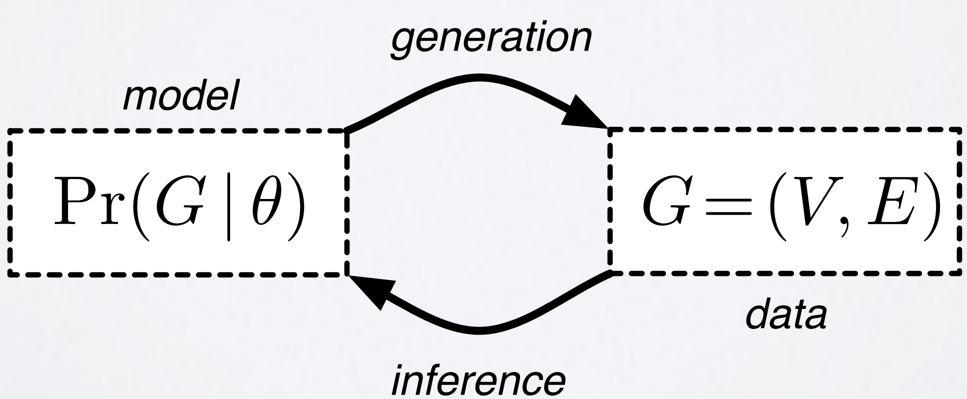
a solution

idea:

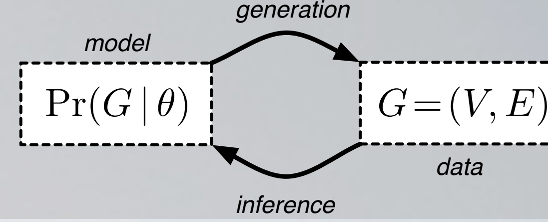
use metadata $\mathbf{x}$ to help select a partition $\mathcal{P}^* \in \{\mathcal{P}\}$ that correlates with $\mathbf{x}$, from among the exponential number of *plausible* partitions

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



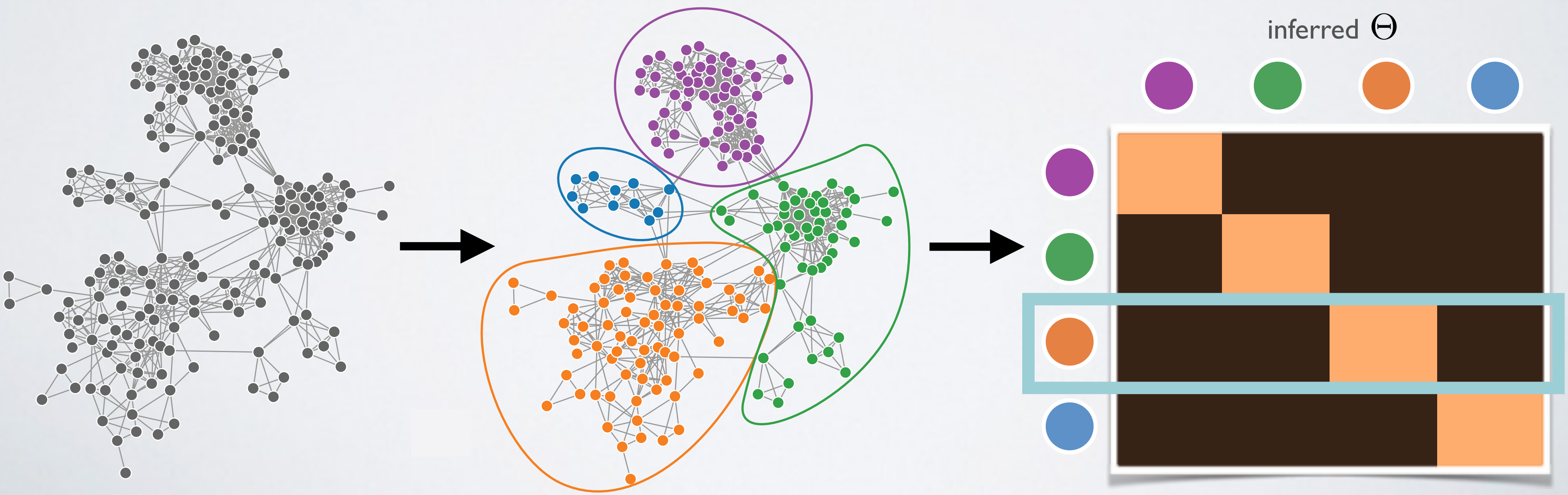
a solution



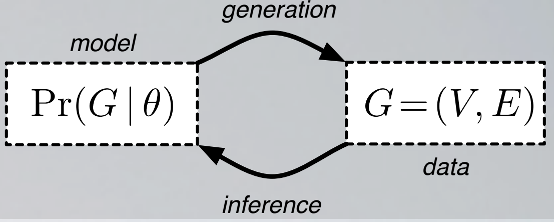
the stochastic block model

- each vertex u has type $s_u \in \{1, \ldots, k\}$
- stochastic block matrix Θ of group-level connection probabilities
- probability that $P(A_{uv} = 1) = \theta_{s_u, s_v}$

community = vertices with same pattern of inter-community connections



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

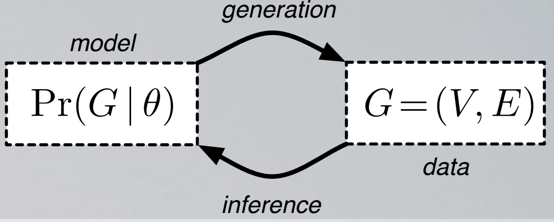
this is a degree-corrected stochastic block model (DC-SBM)

with a metadata-based prior on community labels

annotations are just more data, not "ground truth"

[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

$$P(\mathbf{A} \mid \Theta, \Gamma, \mathbf{x}) = \sum_{\mathbf{s}} \overset{\text{network}}{P(\mathbf{A} \mid \Theta, \mathbf{s})} \overset{\text{metadata}}{P(\mathbf{s} \mid \Gamma, \mathbf{x})}$$

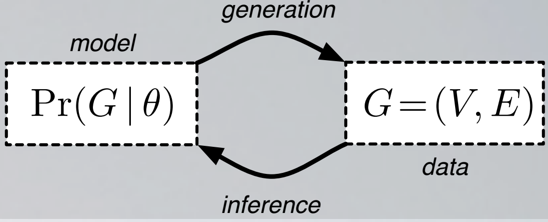
$$= \sum_{\mathbf{s}} \prod_{u<v} p_{uv}^{A_{uv}} (1 - p_{uv})^{1 - A_{uv}} \prod_{u} \gamma_{s_u, x_u}$$

- 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 $P(\mathbf{A} \mid \Theta, \Gamma, \mathbf{x})$ w.r.t. Θ and Γ

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

a metadata-aware stochastic block model



inference

with a little math, the optimal parameters are

$$\underbrace{\theta_{st} = \frac{\sum_{uv} A_{uv} q^{uv}_{st}}{\sum_u d_u q^u_s \sum_v d_v q^v_t}}_{\textit{community mixing matrix}} \qquad \underbrace{\gamma_{sx} = \frac{\sum_u \delta_{x,x_u} q^u_s}{\sum_u \delta_{x,x_u}}}_{\textit{metadata-community associations}}$$

where $\underbrace{q^u_s = \sum_{\mathbf{s}} q(\mathbf{s}) \delta_{s_i,s}}_{\textit{marginal posterior community membership}}$ and $\underbrace{q^{uv}_{st} = \sum_{\mathbf{s}} q(\mathbf{s}) \delta_{s_u,s} \delta_{s_v,t}}_{\textit{joint probability of community membership}}$

with $q(\mathbf{s}) = \dfrac{P(\mathbf{A} \,|\, \Theta, \mathbf{s}) P(\mathbf{s} \,|\, \Gamma, \mathbf{x})}{\sum_{\mathbf{s}} P(\mathbf{A} \,|\, \Theta, \mathbf{s}) P(\mathbf{s} \,|\, \Gamma, \mathbf{x})} = P(\mathbf{s} \,|\, \mathbf{A}, \Theta, \Gamma, \mathbf{x})$.

[1] where δ_{xy} is the Kronecker delta function
[2] the denominator of $q(\mathbf{s})$ is expensive to calculate. instead of MCMC, we use *belief propagation* to calculate it
[3] we may also define this model and derive estimators for when $\mathbf{x}$ is real-valued metadata

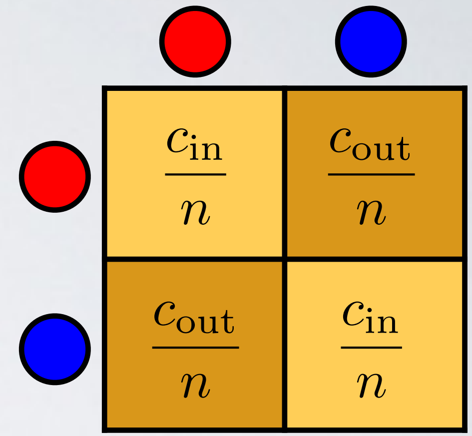
networks with planted structure

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)

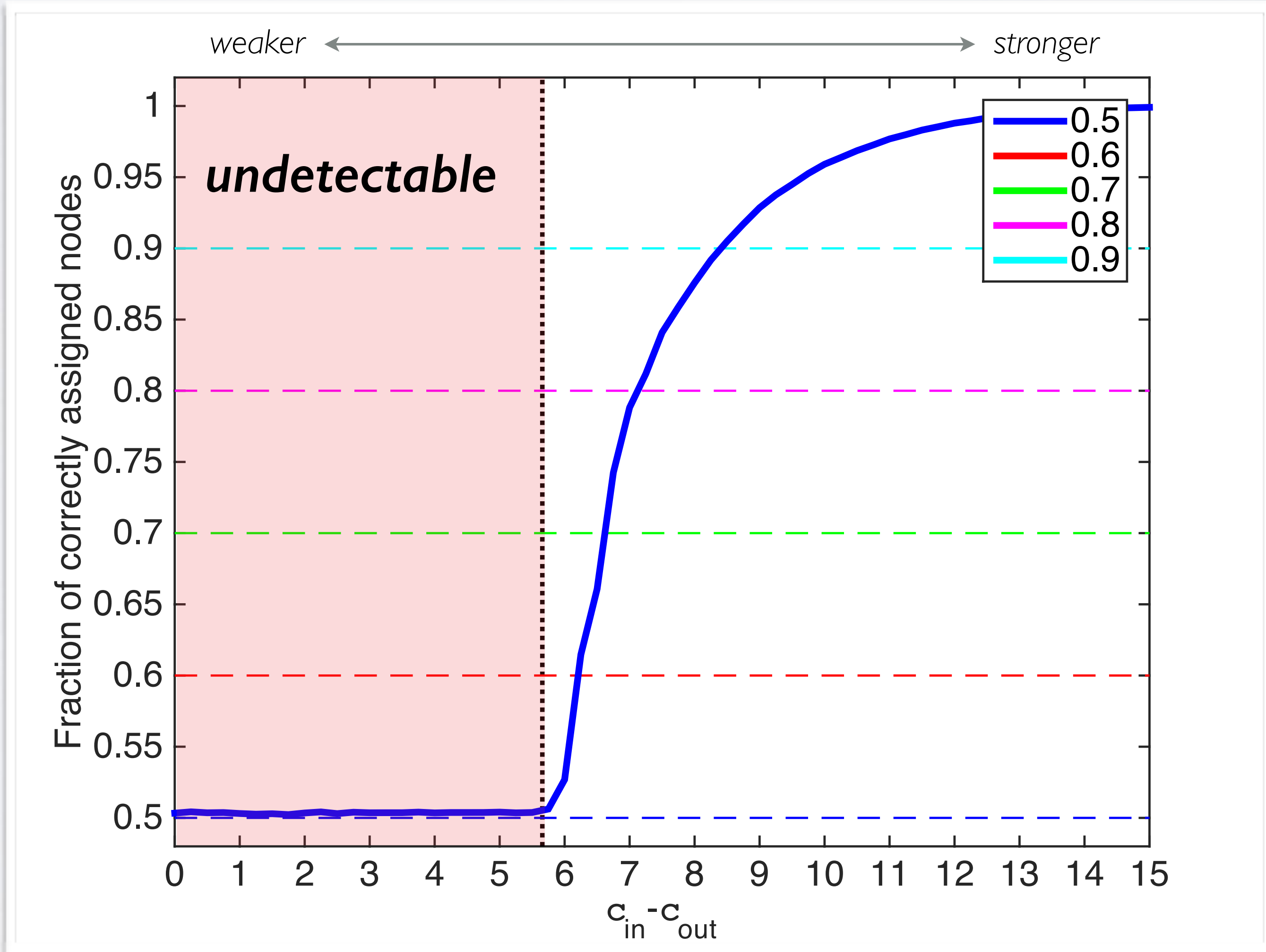


[1] Decelle, Krzakala, Moore & Zdeborova (2011)

networks with planted structure

let mean degree $c = 8$

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

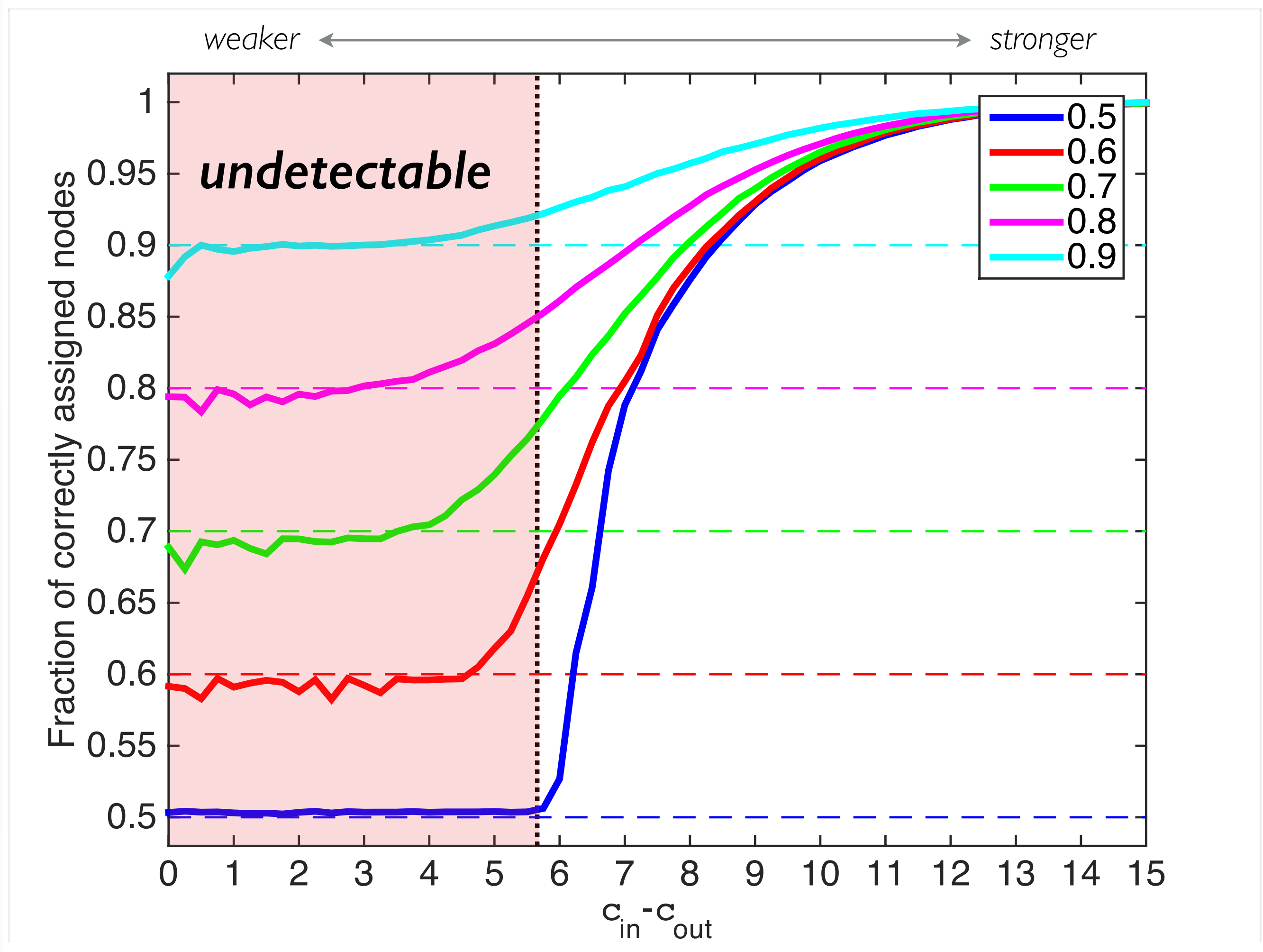


[!] $n = 10\,000$

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

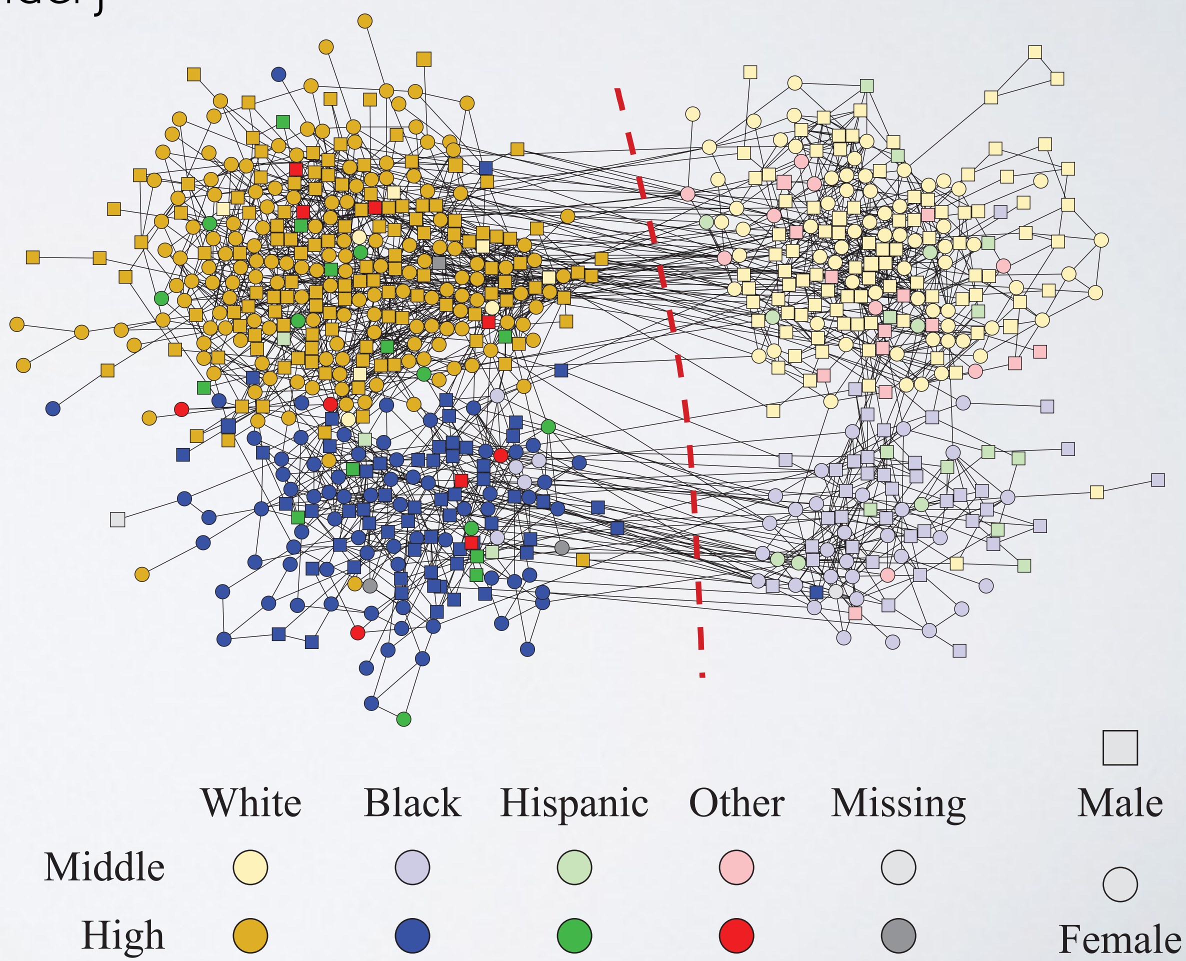
- $\mathbf{x} = \{$**grade 7-12**, ethnicity, gender$\}$

- 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

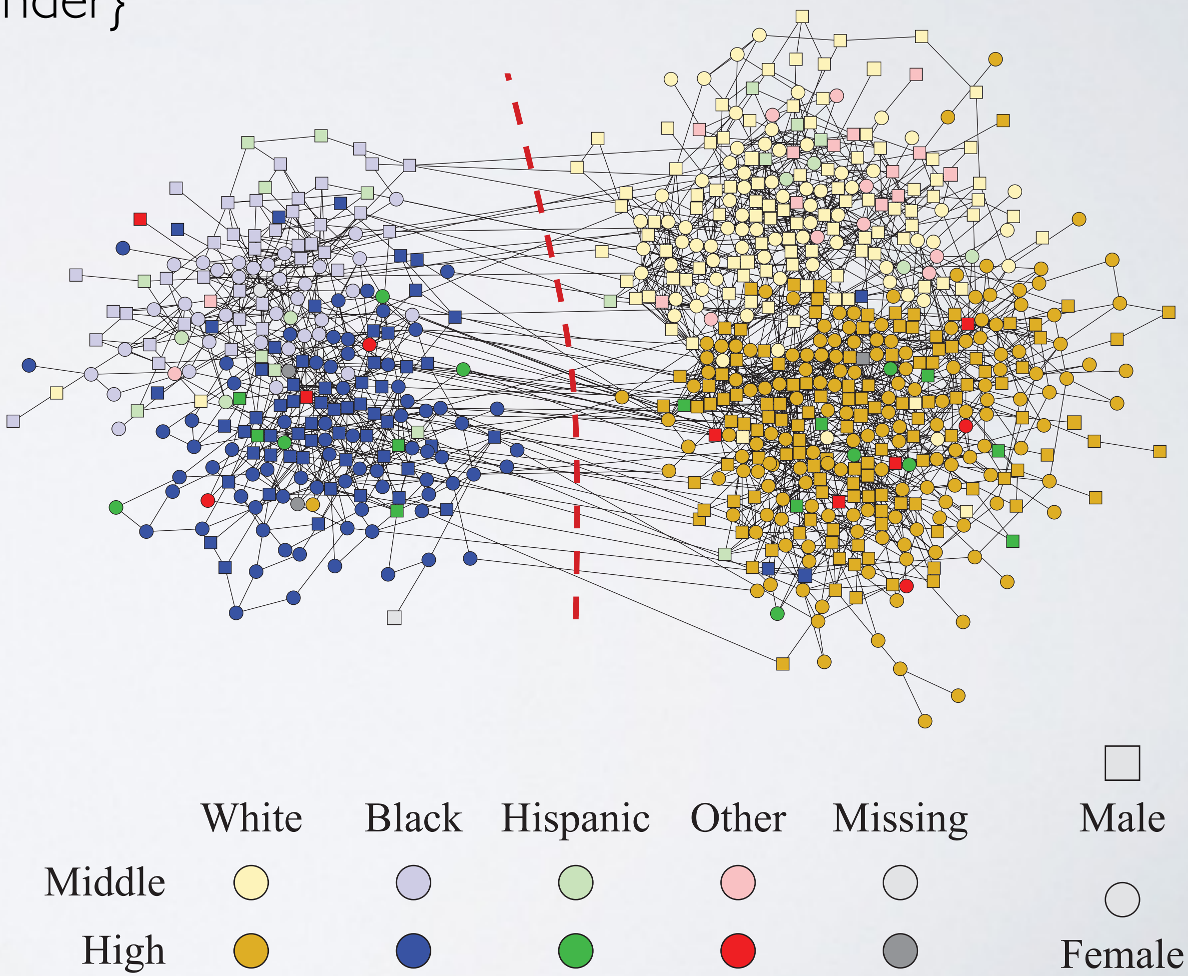
- $\mathbf{x} = \{$grade 7-12, **ethnicity**, gender$\}$

- 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

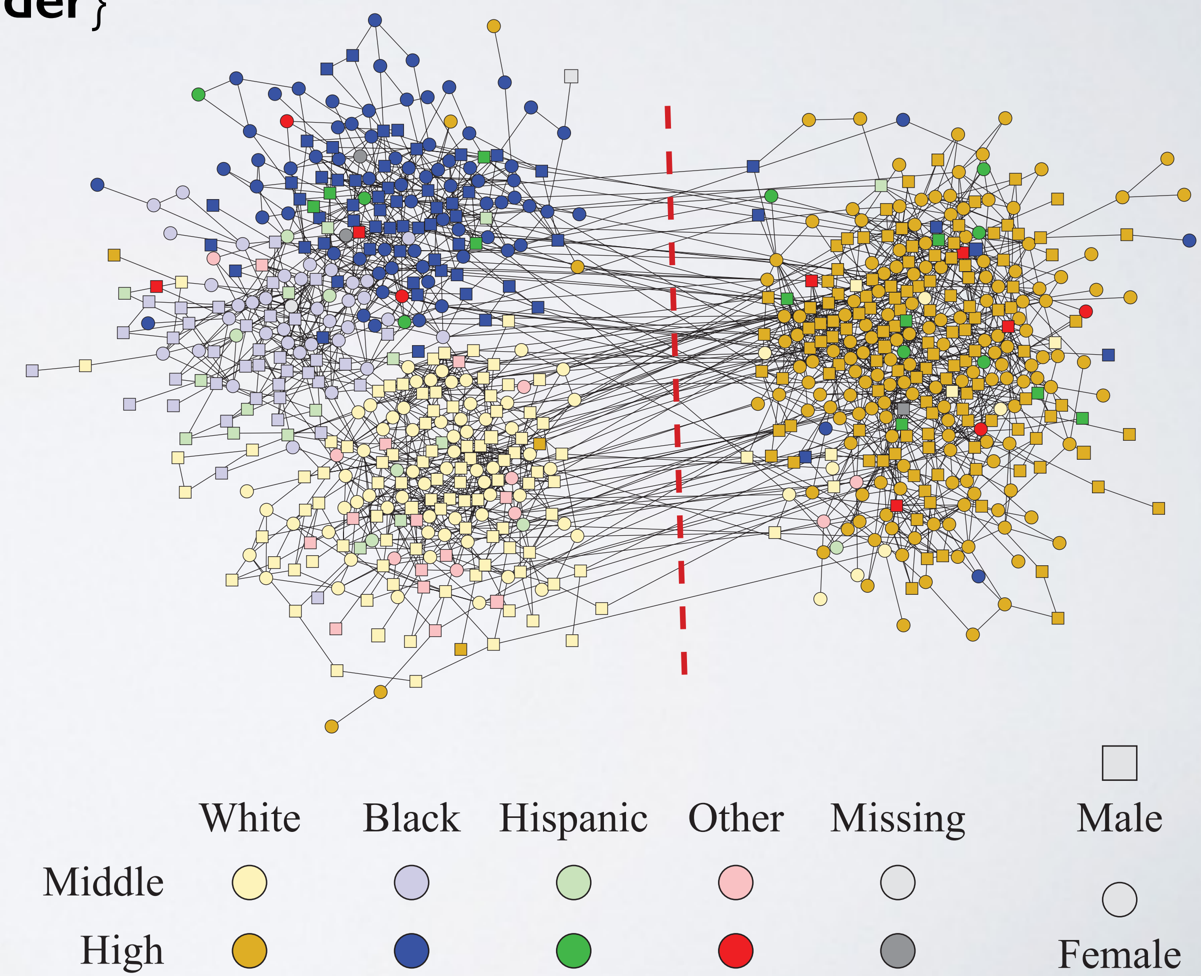
- $\mathbf{x} = \{$grade 7-12, ethnicity, **gender**$\}$

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

  $\text{NMI} = 0.003$

- without metadata:

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

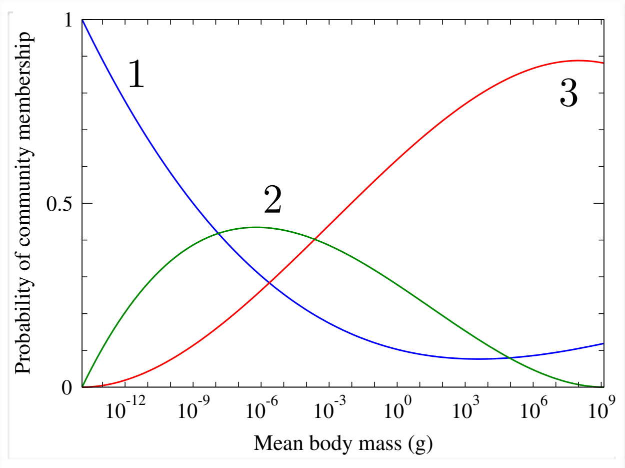


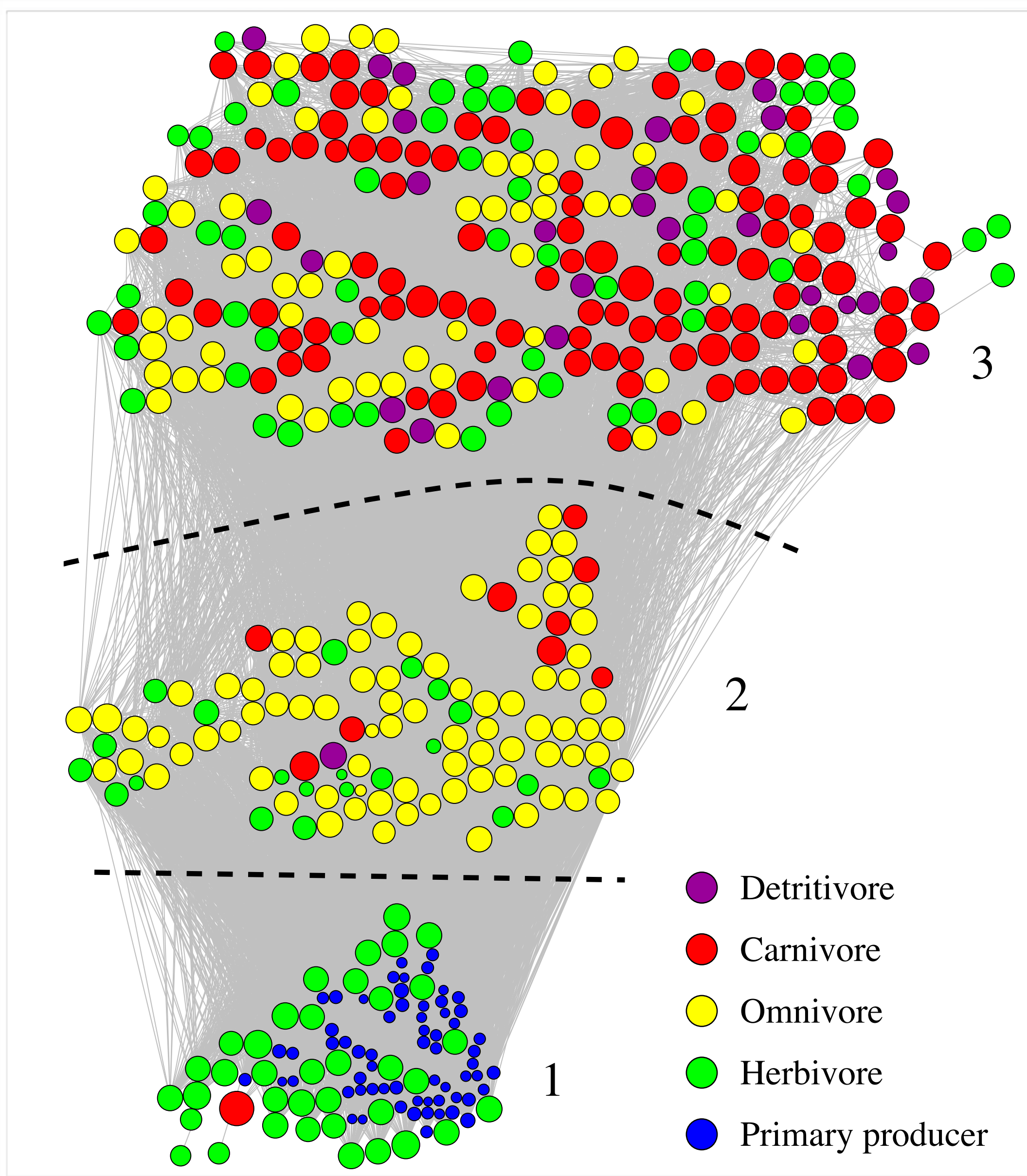
[1] Add Health network data, designed by Udry, Bearman & Harris

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:



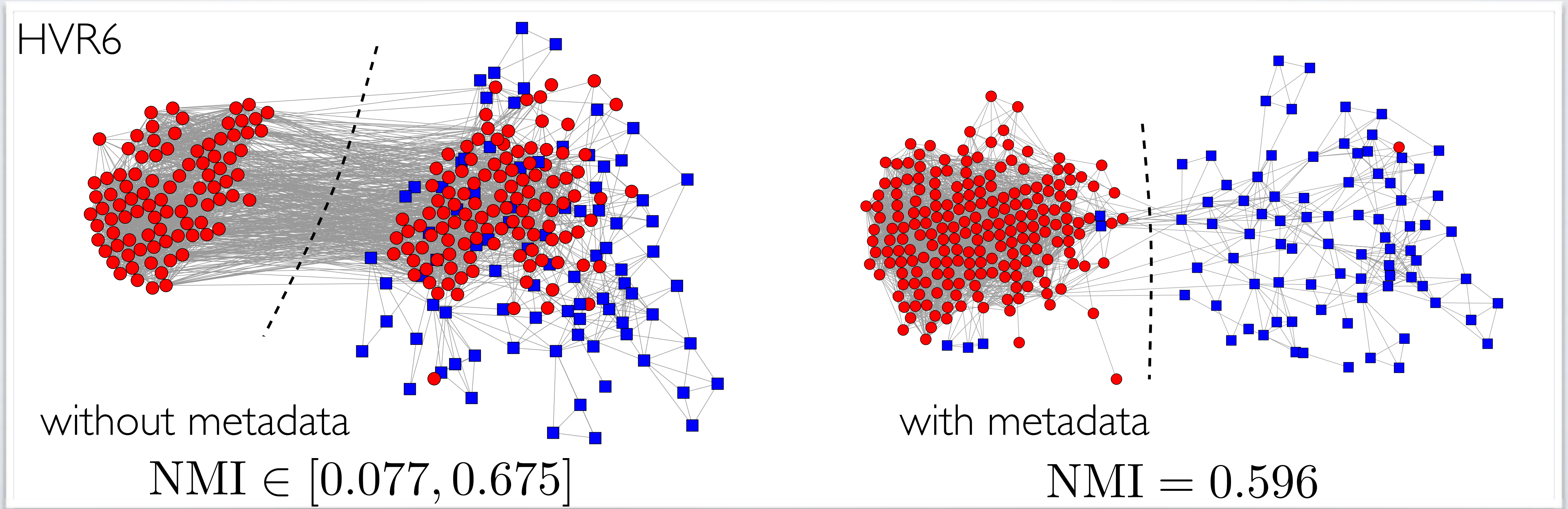


[1] here, we're using a continuous metadata model
[2] Brose et al. (2005)

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)

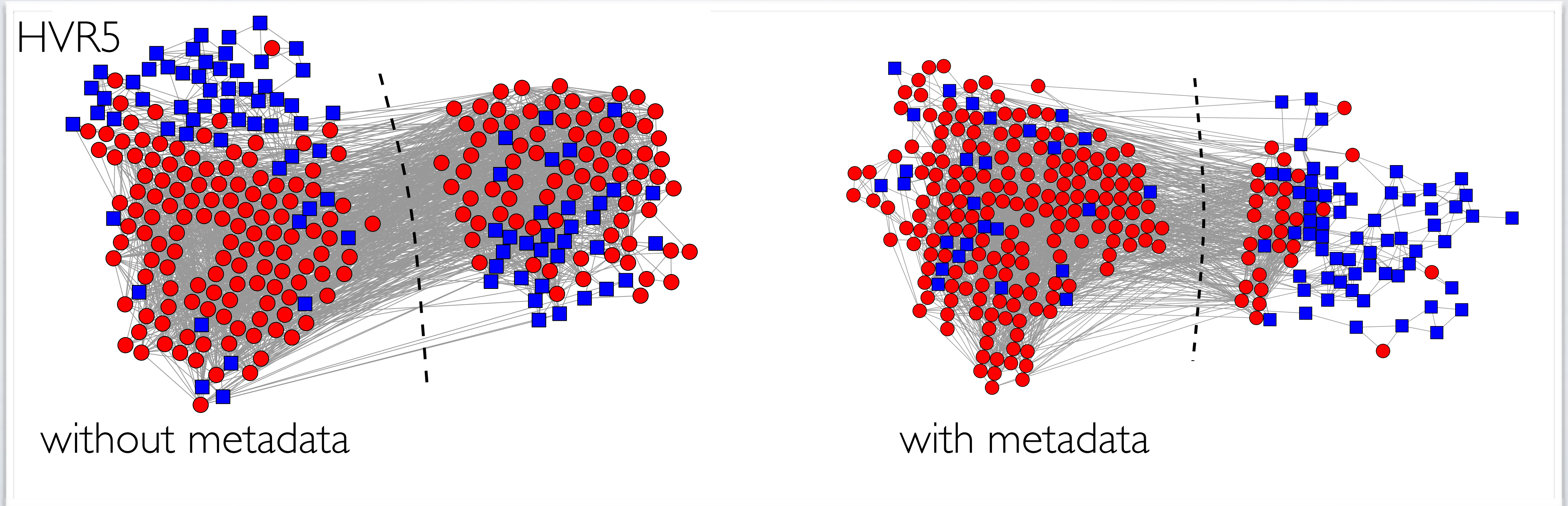


[1] Larremore, Clauset & Buckee (2013)

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)

real-world networks

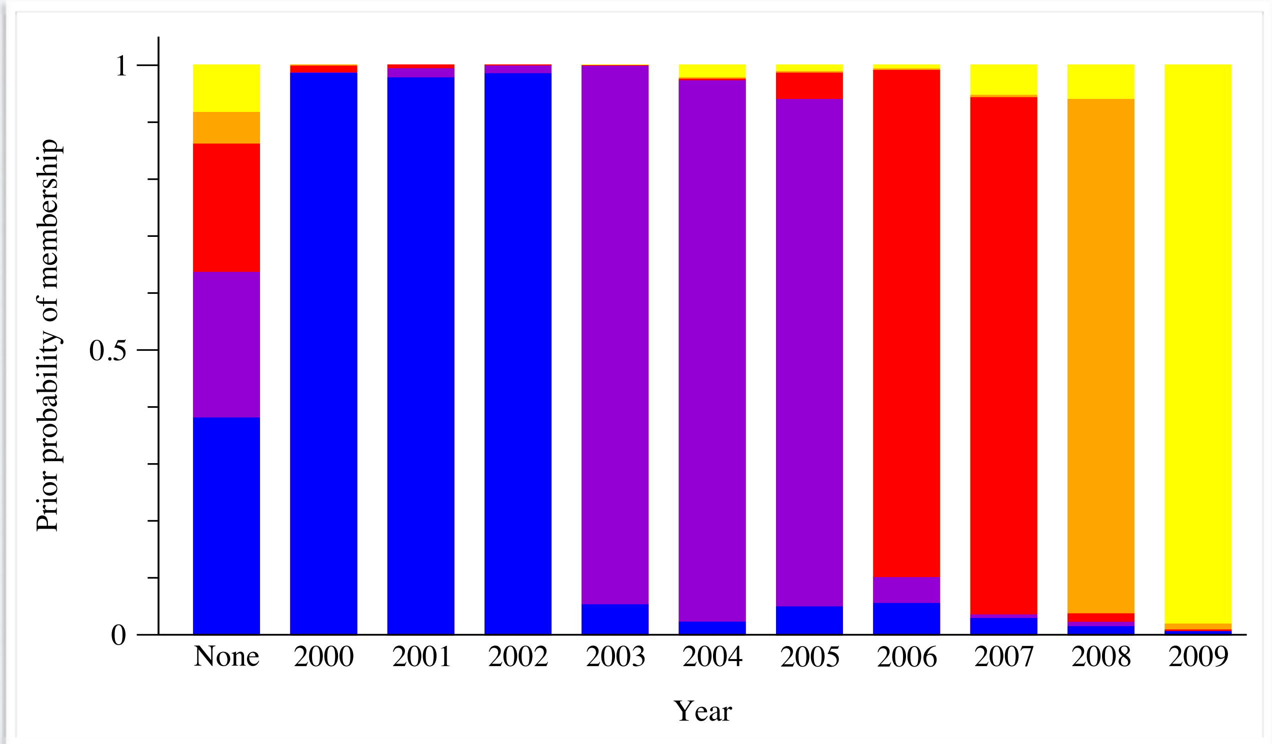
4. **Facebook friendships**: online friendships among 15,126 Harvard students and alumni (in Sept. 2005)

- $\mathbf{x} =$ {**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

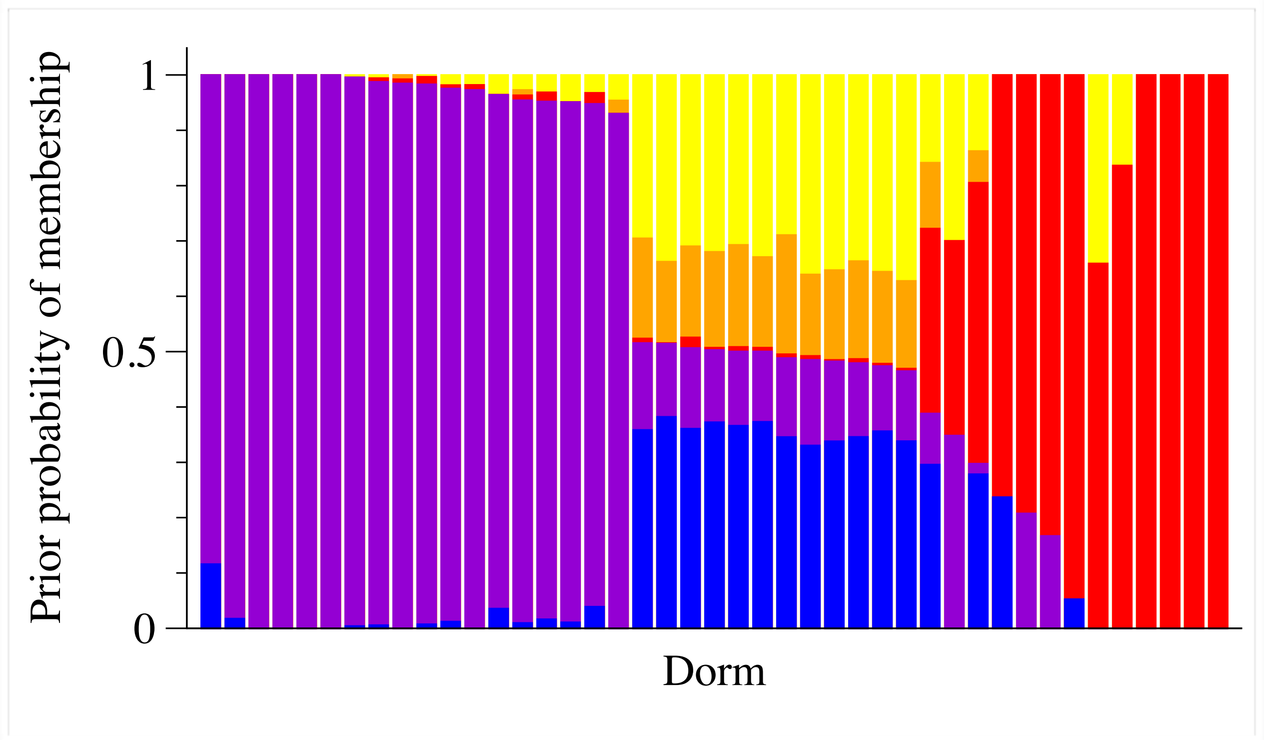
4. **Facebook friendships**: online friendships among 15,126 Harvard students and alumni (in Sept. 2005)

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

  $\mathrm{NMI} = 0.255$

- without metadata:

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



[1] Traud, Mucha & Porter (20012)

real-world networks

5. **Internet graph**: 262,953 peering relations among 46,676 Autonomous Systems

- $\mathbf{x} = \{$**country location of AS**$\}$
- method finds a good partition along the lines of the 173 countries

  $$\mathrm{NMI} = 0.870$$

- without metadata:

  $$\mathrm{NMI} \in [0.398, 0.626]$$

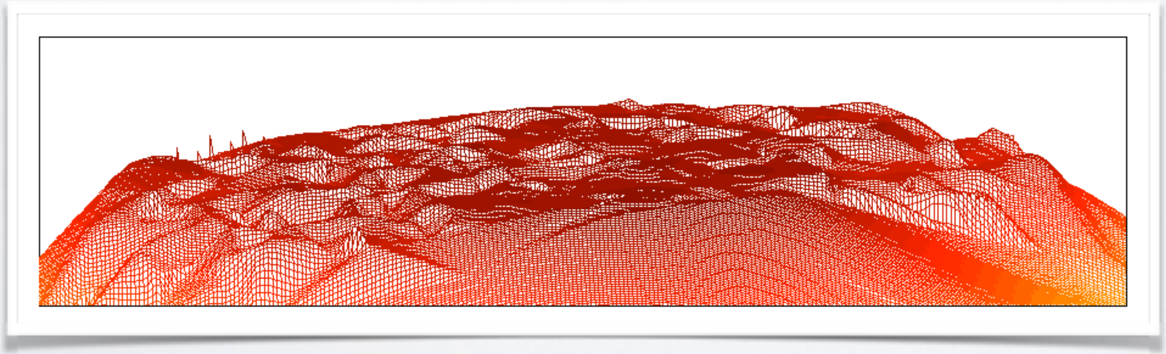
[1] here, we're using a continuous metadata model

conclusions

conclusions

the troubles with community detection…

1. exponential number of competitive local optima
2. inconsistent correlations with node metadata



[1] as it turns out, it is not known whether the SBM has resolution limits, but some evidence suggests it does

conclusions

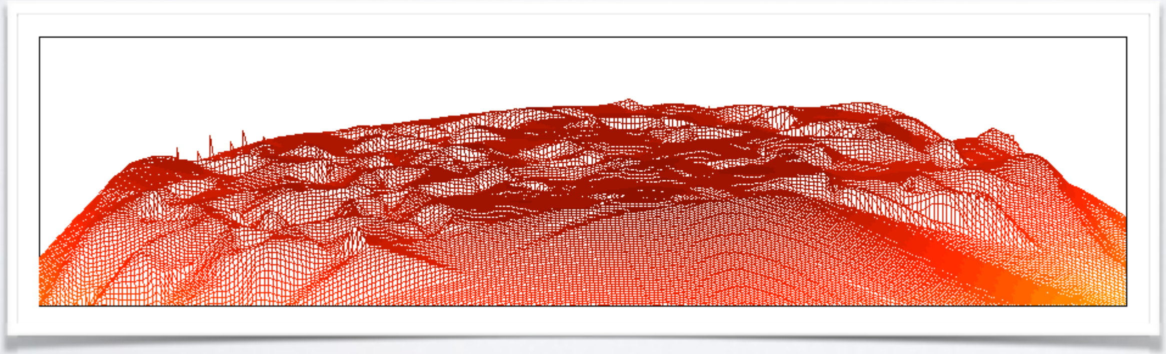
the troubles with community detection…

1. exponential number of competitive local optima
2. inconsistent correlations with node metadata

the virtue of community detection…

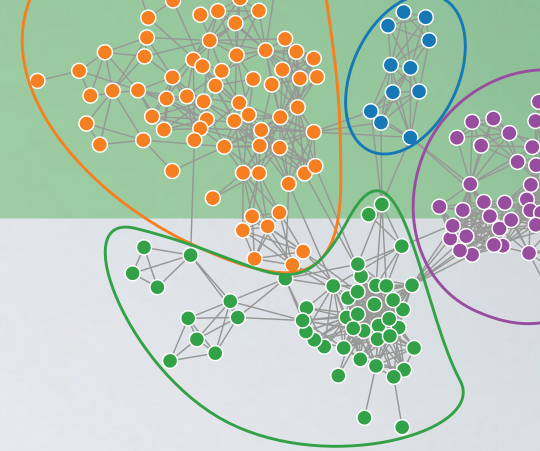
trouble 1 helps us solve trouble 2

there are a multiplicity of good partitions —
we simply needed a good way to choose from among them



[1] as it turns out, it is not known whether the SBM has resolution limits, but some evidence suggests it does

conclusions



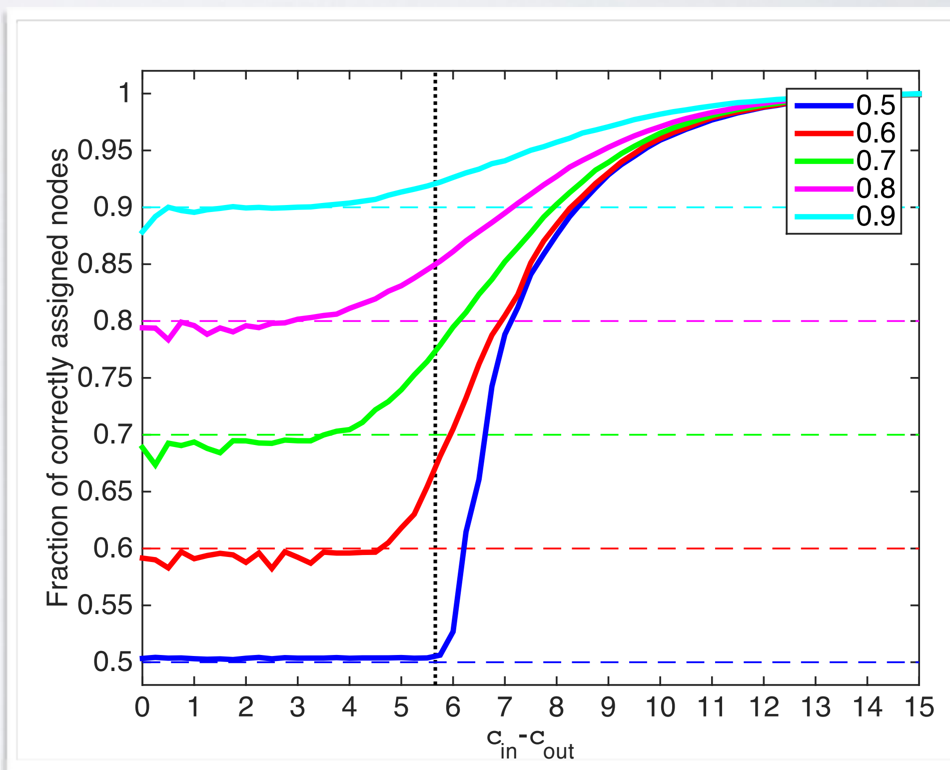
a metadata-aware stochastic block model

- probabilistic model of community structure and node metadata

$$P(\mathbf{A} \mid \Theta, \Gamma, \mathbf{x})$$

- ***yields***: posterior probabilities of community labels q, the mixing matrix Θ , and learned metadata-community association Γ

- metaSBM performs better than any either *structure-only* or *metadata-only algorithm*
- highly scalable, via EM + belief propagation
- works well in practice



future directions

- **a phase transition**

  does metadata eliminate or simply defer the detectability transition for sparse communities?

- **applications**

  how many networks with metadata have we overlooked because our algorithms didn't appear to work well on them?

  apply to all the data, learn all the science

- **extensions**

  to time-evolving networks, to multiplex networks, to other types of metadata (tags, vectors, etc.), etc.

- **more?**

[1] perhaps following Ghasemian et al. (2015) arXiv:1506.06179 or Peixoto (2015) arXiv:1504.02381

acknowledgements

Funding support: NSF

Mark Newman
(Michigan)

Structure and inference in annotated networks

M. E. J. Newman[1,2] and Aaron Clauset[2,3]

[1]Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109, USA
[2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501
[3]Department of Computer Science and BioFrontiers Institute, University of Colorado, Boulder, CO 80309, USA

arxiv:1507.0400

Benjamin H. Good
(Harvard)

Yva de Montjoye
(MIT)

PHYSICAL REVIEW E **81**, 046106 (2010)

Performance of modularity maximization in practical contexts

Benjamin H. Good,[1,2,*] Yves-Alexandre de Montjoye,[3,2,†] and Aaron Clauset[2,‡]

[1]Department of Physics, Swarthmore College, Swarthmore, Pennsylvania 19081, USA
[2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA
[3]Department of Applied Mathematics, Université Catholique de Louvain, 4 Avenue Georges Lemaitre, B-1348 Louvain-la-Neuve, Belgium

(Received 1 October 2009; published 15 April 2010)

[1] a C implementation of the metadata SBM is available by request

fin

[1] a C implementation of the metadata SBM is available by request