

Large-scale structures in networks: Hidden communities and blocks

Daniel Larremore

Assistant Professor

Dept. of Computer Science
& BioFrontiers Institute

daniel.larremore@colorado.edu
@danlarremore



University of Colorado **Boulder**

PDF of slides available

http://danlarremore.com/CommunityDetection_and_Ranking_Larremore_2018.pdf

Goals for this talk:

1. **Why** do we look for large-scale structure? 🤔
2. **How** do we find communities and hierarchies? 🙄
3. **Where** can we read more details? 📚

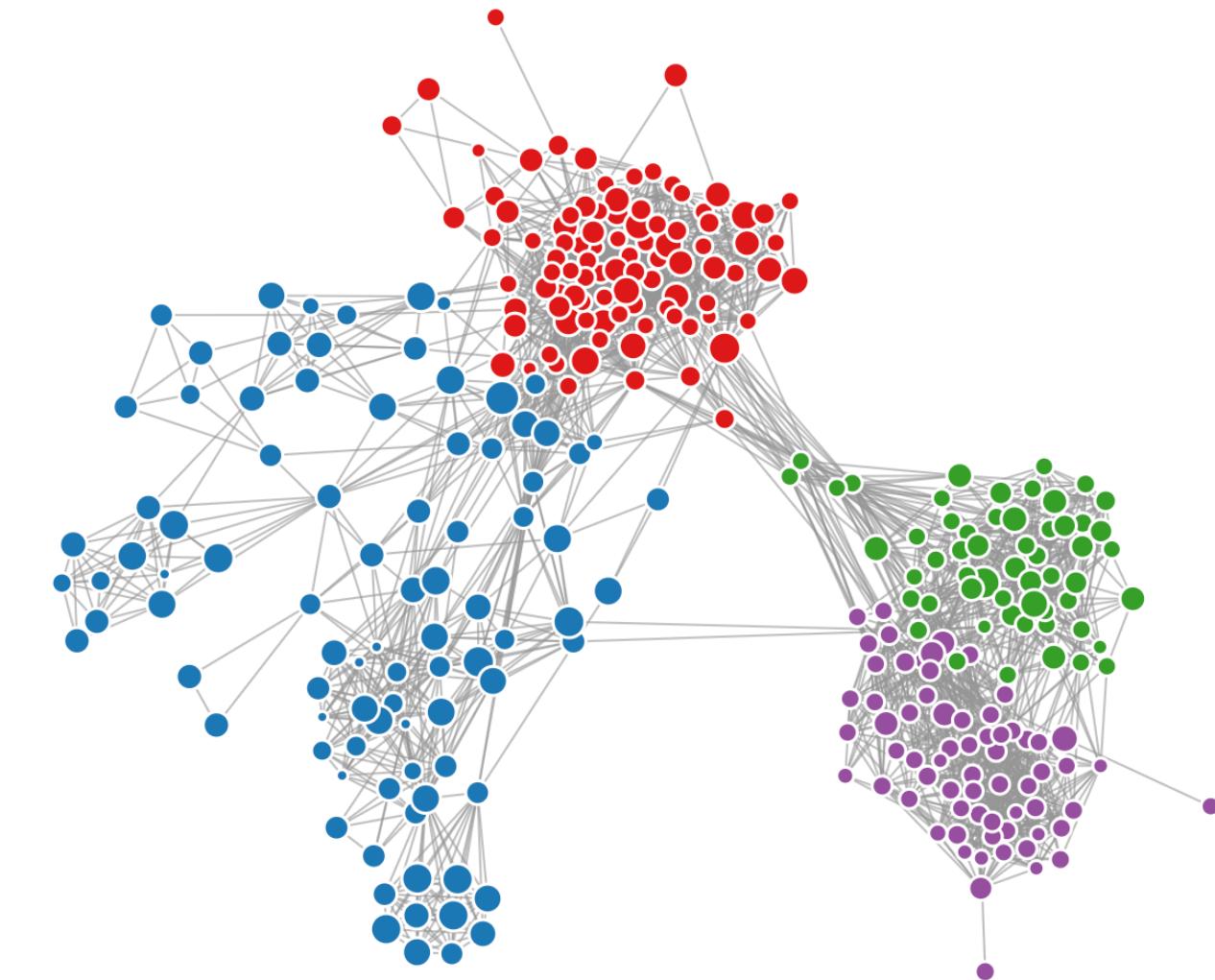
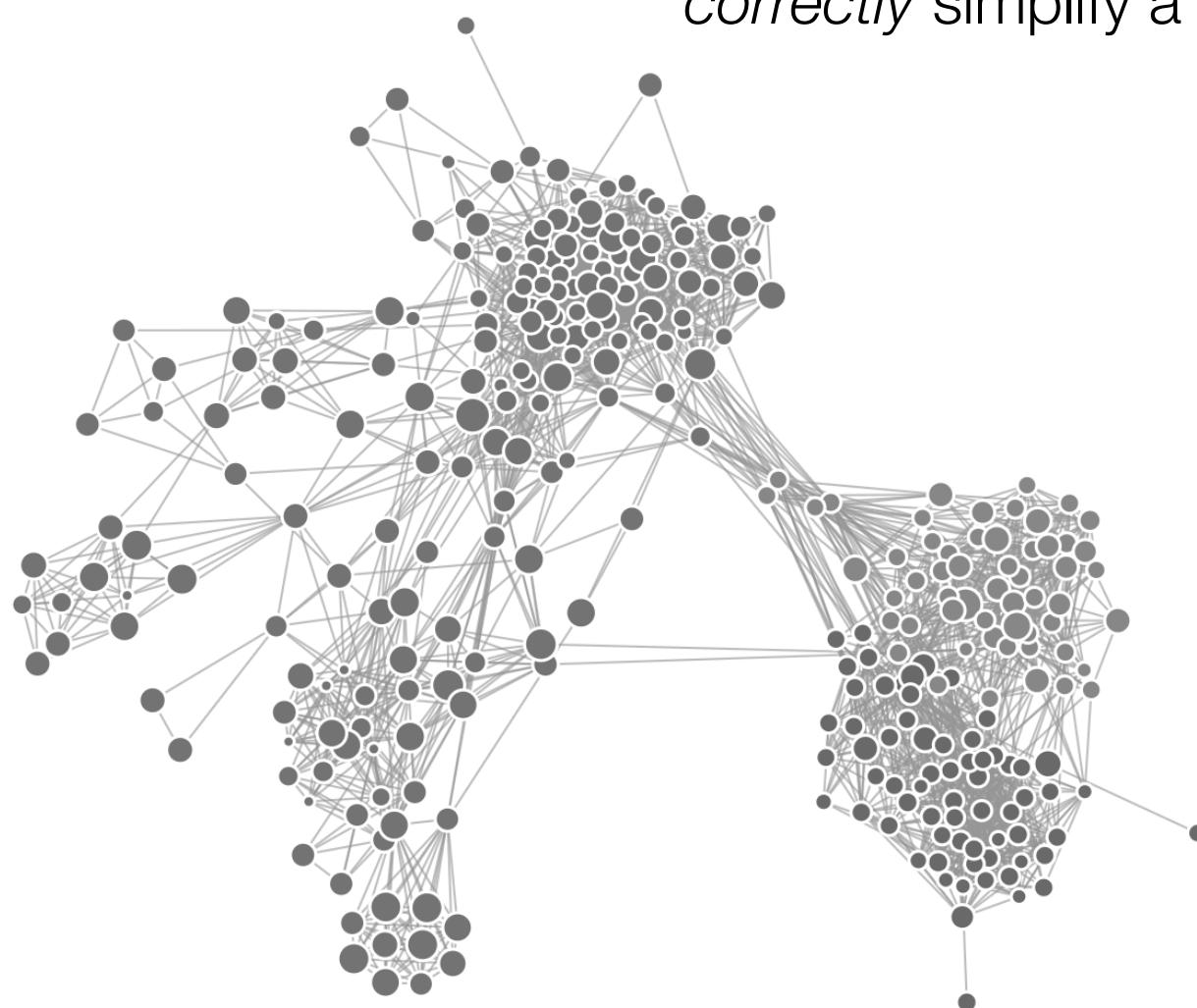
Simplicity is a great virtue but it requires hard work to achieve it and education to appreciate it. And to make matters worse: complexity sells better.

E. W. Dijkstra

We can interpret this in two ways:

The Cynic: Pictures of networks can be *really cool* but our goal is to do good science, not make pretty pictures.

The Scientist: The most beautiful science is when we *correctly* simplify a complex system.



What do we mean by “large-scale structure” ?

Structure is what makes data different from noise.
It's what makes a network different from a random graph.

Networks are often too large and complex to be adequately summarized by a few scalars, like the number of nodes, the number of edges, or the mean degree.

However, they are also often too large and complex to be analyzed *without* some kind of simplification!

Therefore, understanding what the network means requires that we identify key structures.

Searching for large-scale structures in a network reflects a belief that in all the complexity there are patterns that make the network less complicated.

We define these large-scale structures—models, really—to compress complex networks.

Goal: understanding, not a list of parts and dimensions



Finding large-scale structures
is the same as anything else:

We want a simplified model of
something very complicated.

We want to know what the
important pieces are,
and how they fit together.

Many uses for models of large-scale structure

Treat the network like a system:

Extrapolation. Make predictions for as-yet unseen nodes (in “space” or time).

Interpolation. Identify missing links.

Generalization. Nodes of this type are like others of the same type.

Treat the network like an artifact:

Mechanisms. How did this network arise? What rules governed its assembly?

Explanations. Coarse-graining or compression.

Treat the network like a means to an end; an intermediate data structure:

Useful division. Need groups so that we can assign treatments in an A/B test.

Simplification. Downstream regression model needs ranks or groups.

intuition: compare this list with the list you would write for regression

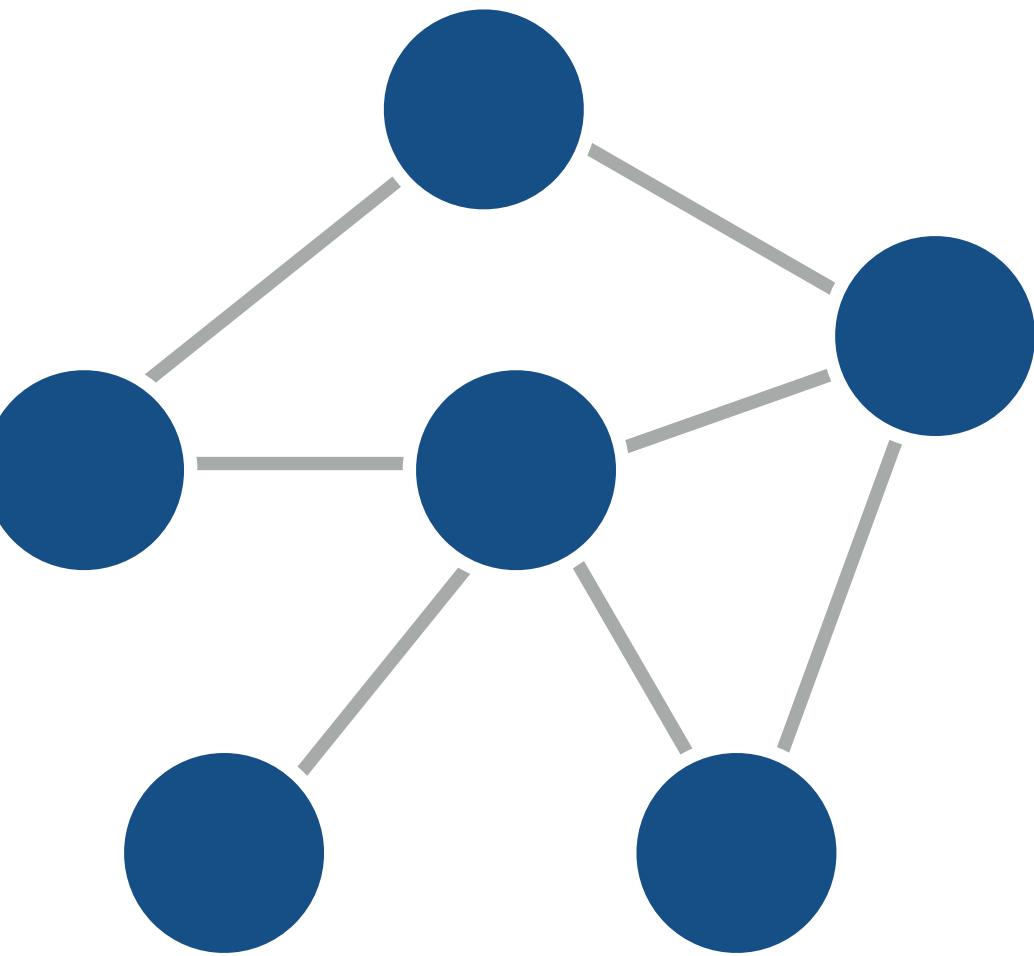
Community structure



Homophily & assortative mixing

like links with *like*

Assortativity coefficient r measures extent of homophily.

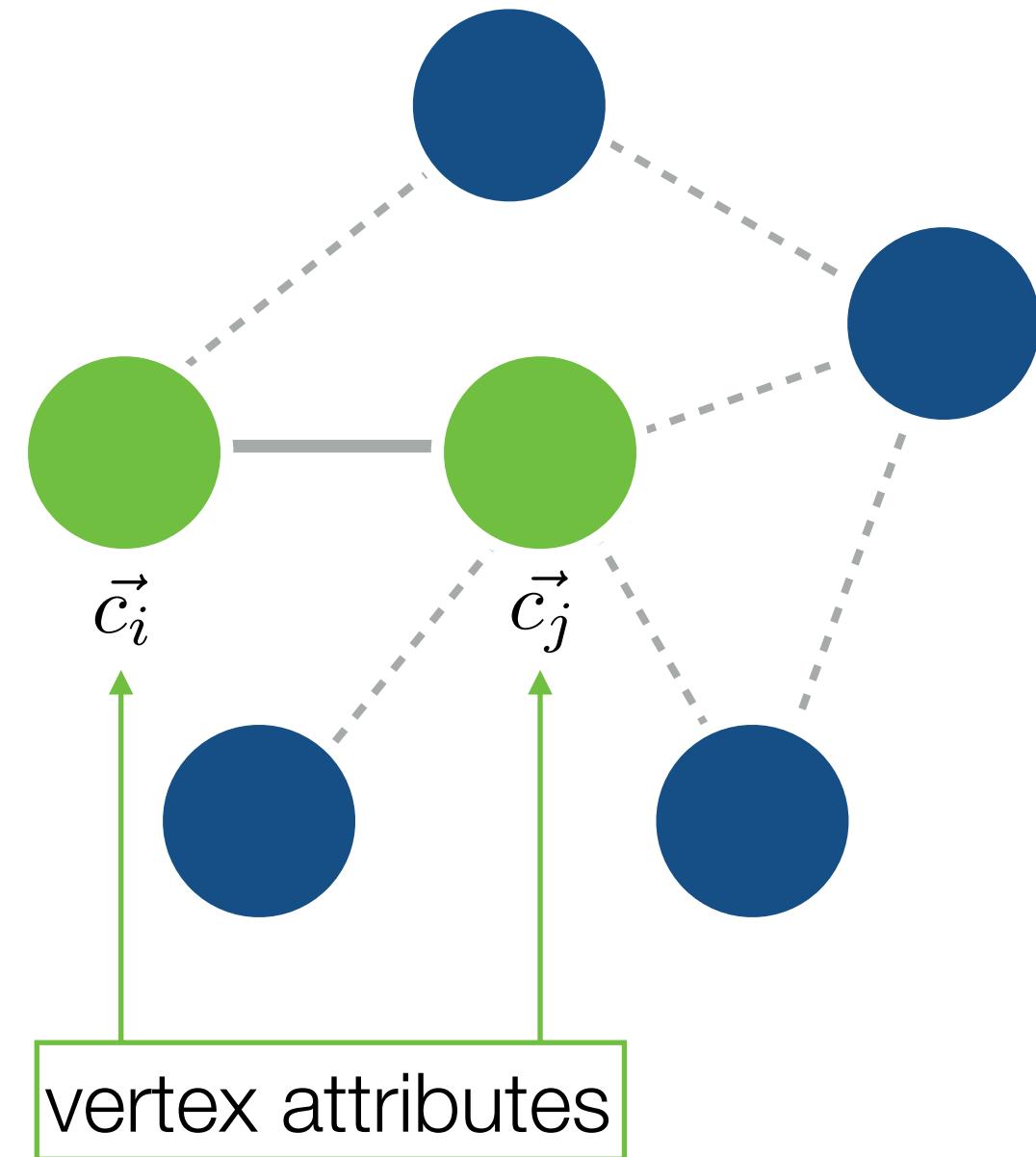


Homophily & assortative mixing

like links with *like*

Assortativity coefficient r measures extent of homophily.

Three types:
scalar attributes
vertex degrees
categorical variables



Homophily & assortative mixing

like links with *like*

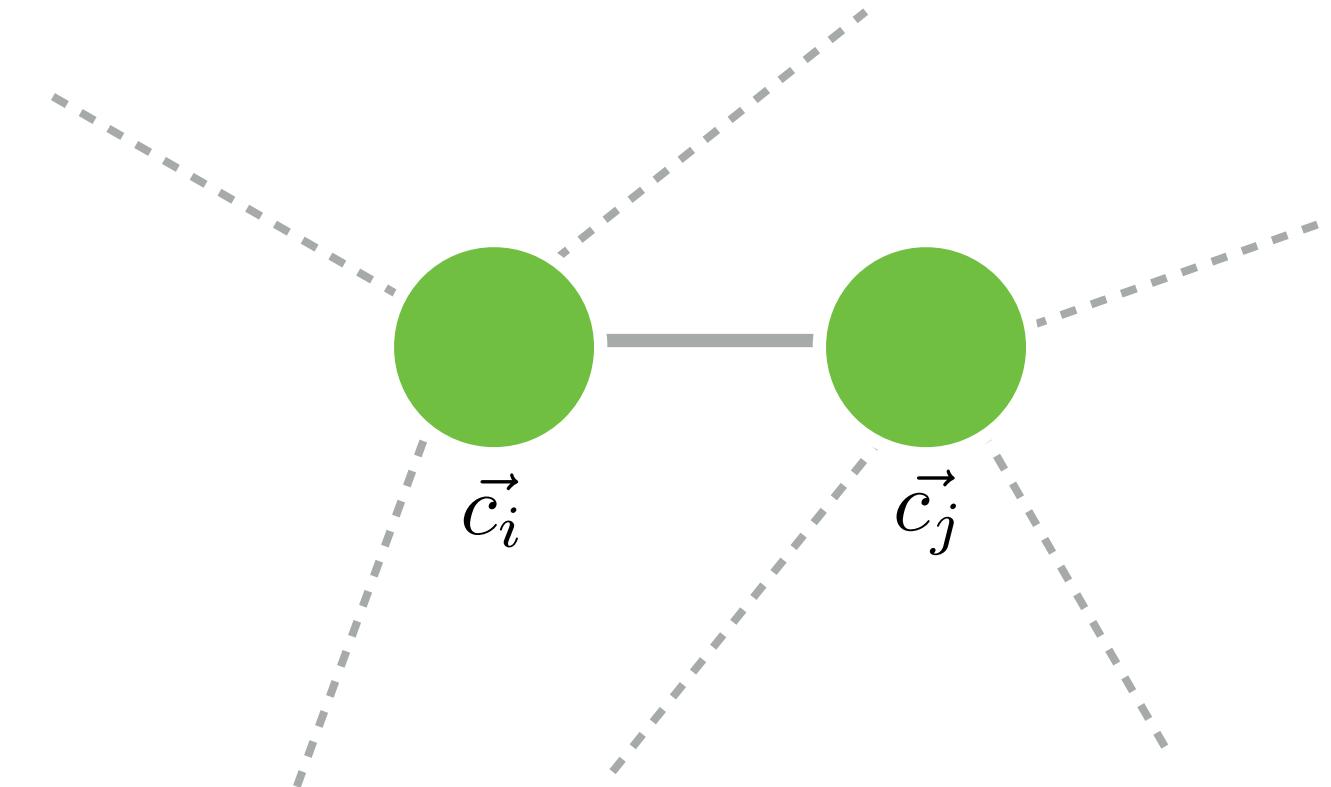
Assortativity coefficient r measures extent of homophily.

scalar attributes:

compute a Pearson correlation over edges.

start with the mean of c across ties:

$$\mu = \frac{1}{2m} \sum_i \sum_j A_{ij} c_i = \frac{1}{2m} \sum_i k_i c_i$$



Homophily & assortative mixing

like links with *like*

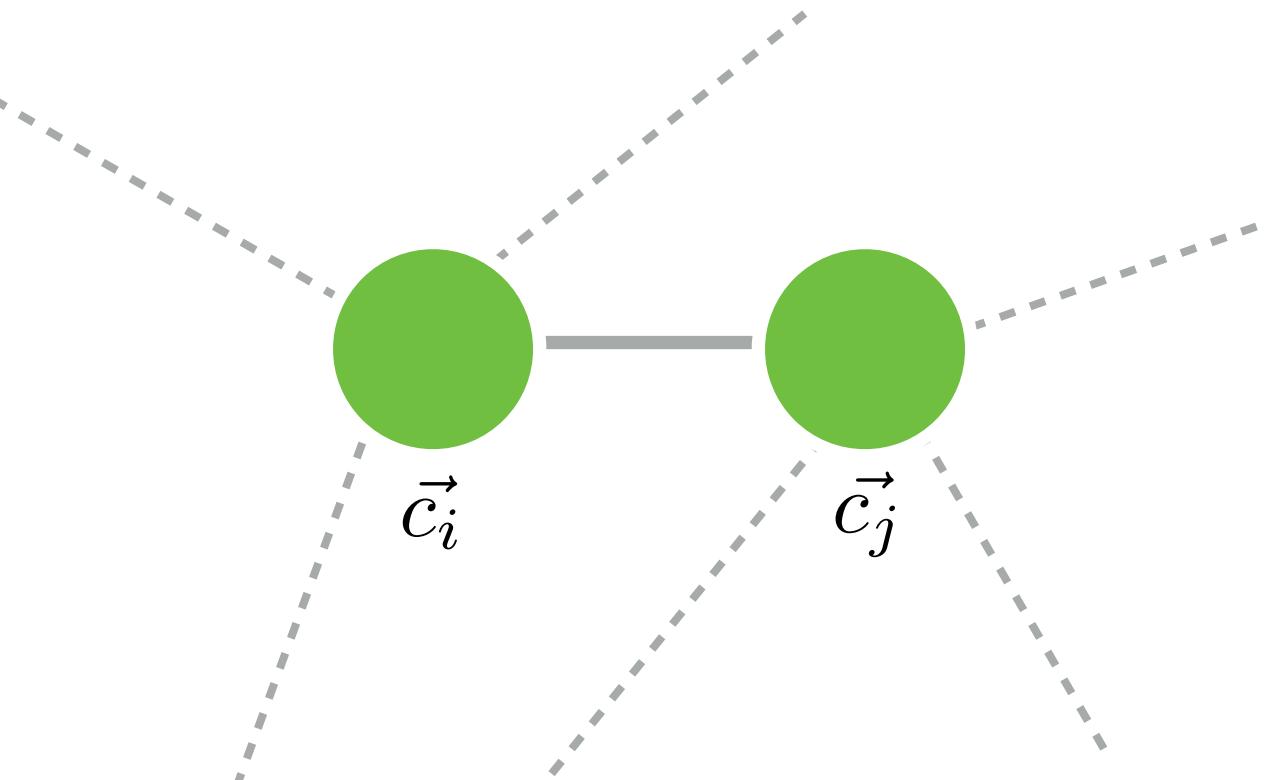
Assortativity coefficient r measures extent of homophily.

compute covariance:

$$\text{cov}(c_i, c_j) = \frac{\sum_{ij} A_{ij}(c_i - \mu)(c_j - \mu)}{\sum_{ij} A_{ij}}$$

$$= \frac{1}{2m} \sum_{ij} A_{ij} c_i c_j - \mu^2$$

$$= \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) c_i c_j$$



substitute $\mu = \frac{1}{2m} \sum_i k_i c_i$

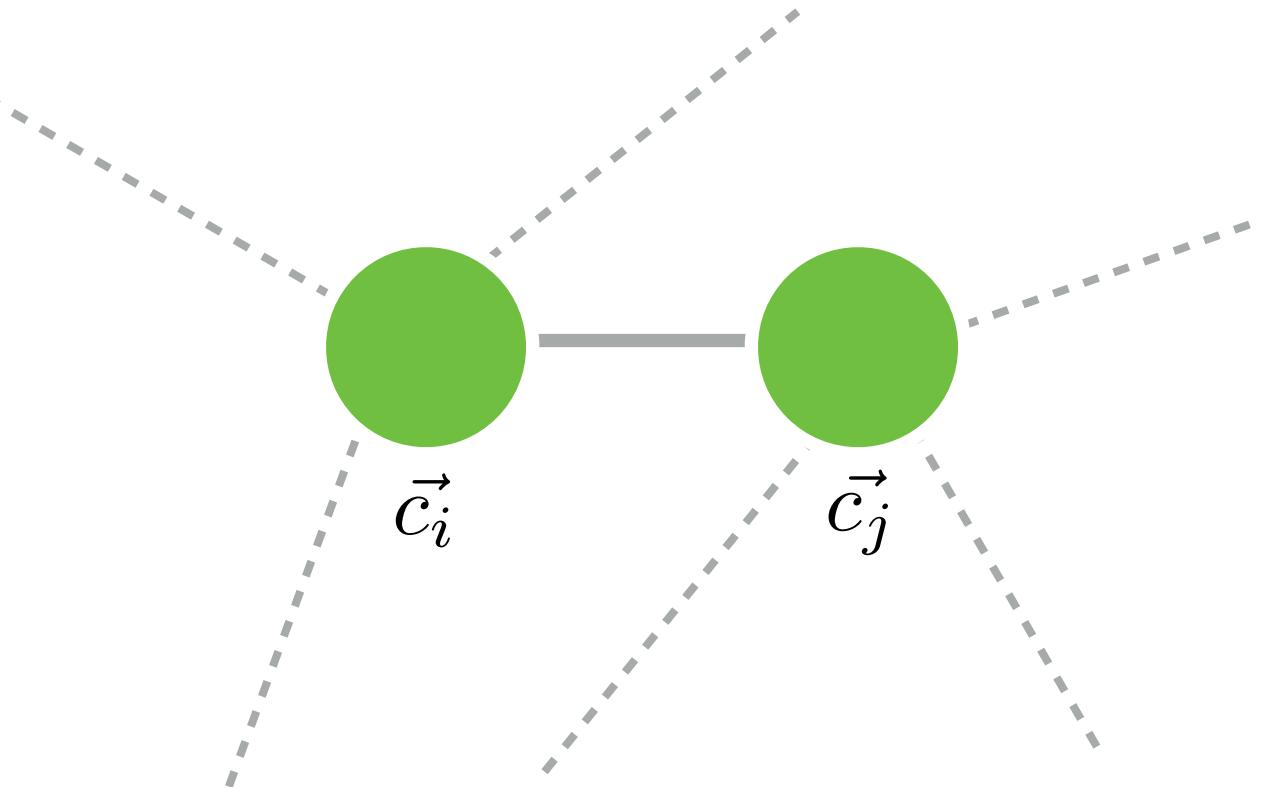
Homophily & assortative mixing

like links with *like*

Assortativity coefficient (scalar).

$$r = \frac{\text{cov}(c_i, c_j)}{\text{var}(c_i, c_j)}$$

$$r = \frac{\sum_{ij} (A_{ij} - k_i k_j / 2m) c_i c_j}{\sum_{ij} k_i \delta_{ij} - k_i k_j / 2m}$$



“it’s that easy!” 😎

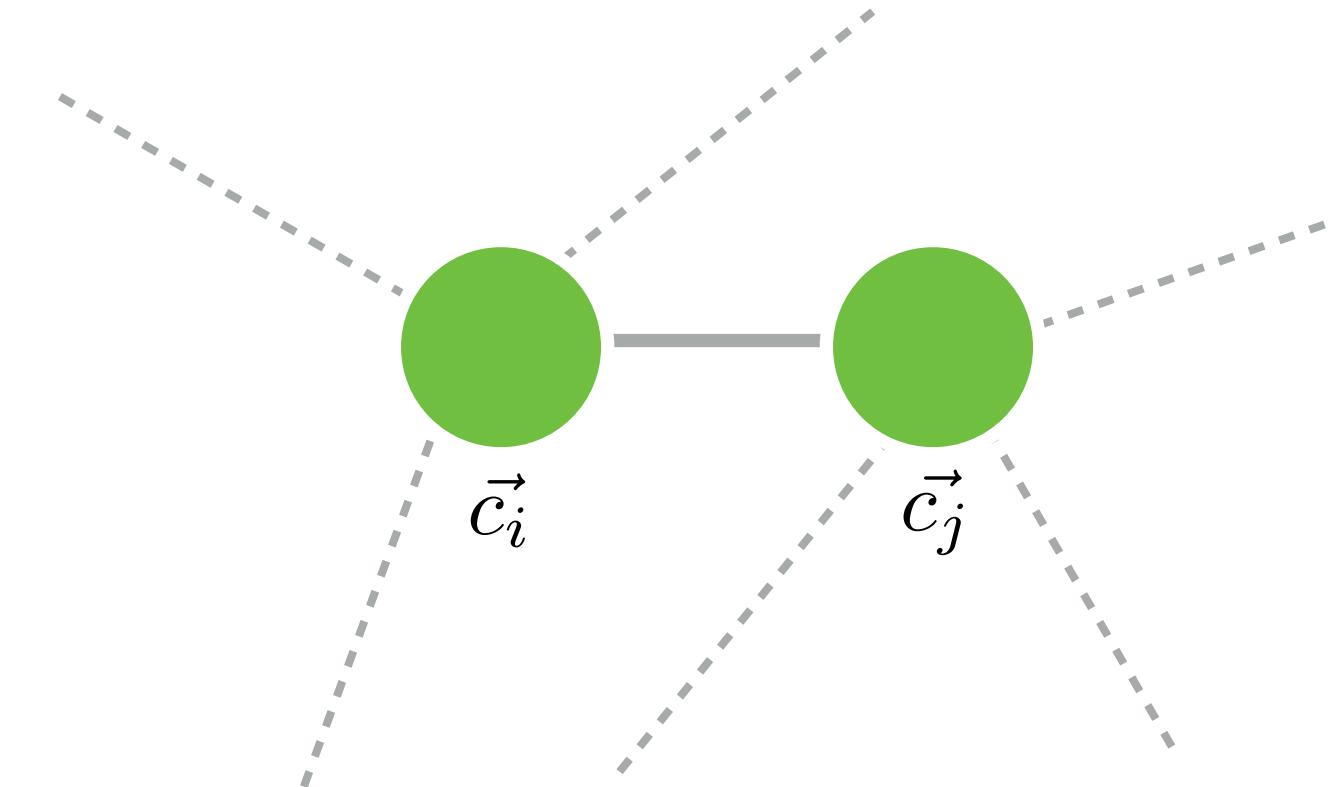
$$-1 \leq r \leq 1$$

Homophily & assortative mixing

like links with *like*

degree: just another scalar.

(very well studied!)



Homophily & assortative mixing

like links with *like*

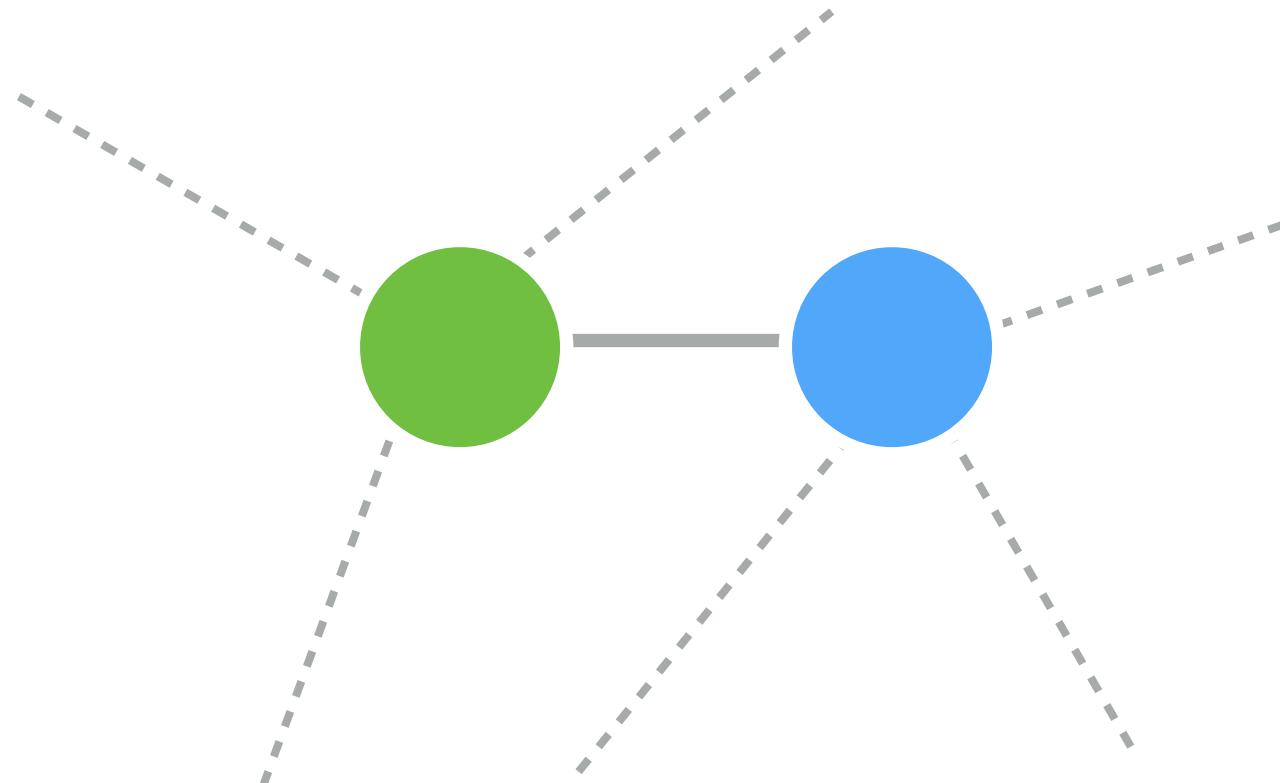
categories: not just another scalar.

Imagine that c are labels, categories.

let e_{rs} be the fraction of edges
between nodes of type r and s .

$$\sum_{rs} e_{rs} = 1 \quad \sum_r e_{rs} = a_r \quad \sum_s e_{rs} = b_s$$

$$\text{then, } r = \frac{\sum_r e_{rr} - \sum_r a_r b_r}{1 - \sum_r a_r b_r}$$



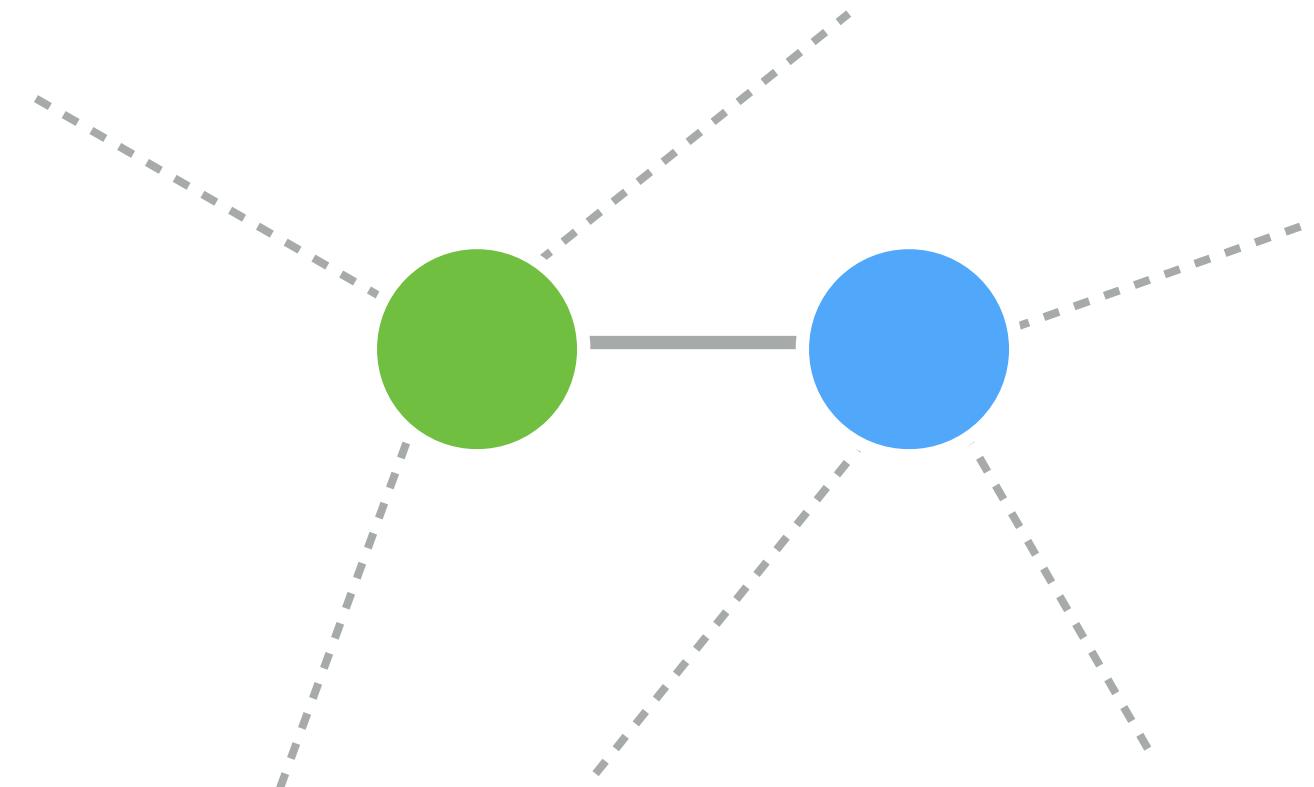
consider: if there were *only* edges
between nodes of the same type...

Homophily & assortative mixing

like links with *like*

It is often convenient to write this in a different way, and call it Q .

Same principle: what fraction of edges fall between nodes of the same community?*



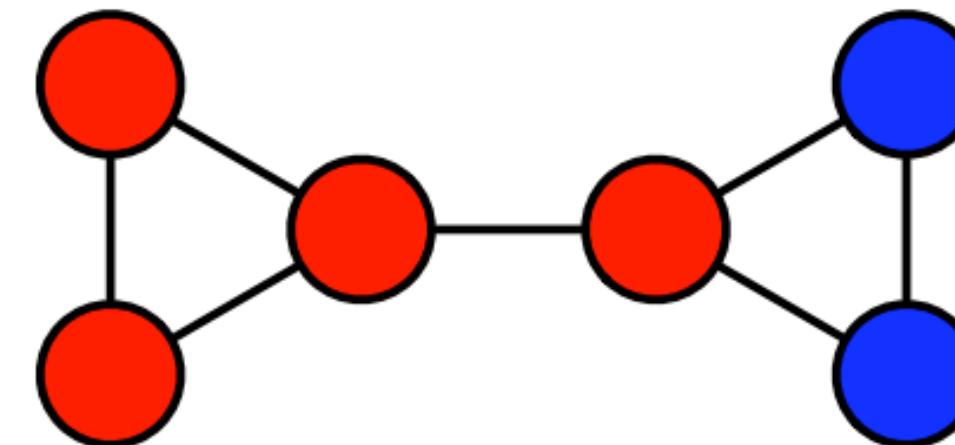
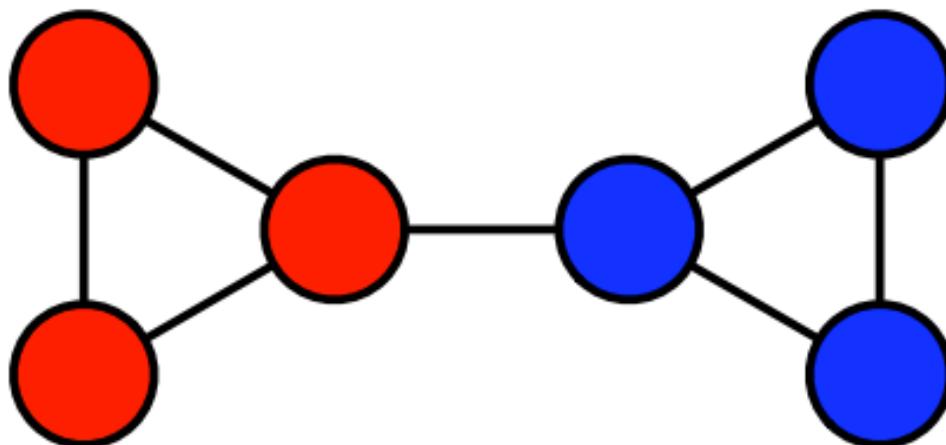
$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{ij}$$

*compared to what we'd expect if the network were random!

Practice makes the master

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{b_i, b_j}$$

$$Q = \sum_r e_{rr} - a_r^2$$



		red	blue
red	3/7	1/14	
blue	1/14	3/7	

$$Q_1 = 5/14 = 0.357$$

		red	blue
red	4/7	2/14	
blue	2/14	1/7	

$$Q = 6/49 = 0.122$$

Modularity

Modularity is easily *the* most popular method for community detection. But why?

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{b_i, b_j}$$

Why is this more powerful than simply a measure of correlation over node labels?

Community structure in social and biological networks

[M Girvan](#), [MEJ Newman](#) - Proceedings of the national ..., 2002 - National Acad Sciences

Abstract A number of recent studies have focused on the statistical properties of networked systems such as social networks and the Worldwide Web. Researchers have concentrated particularly on a few properties that seem to be common to many networks: the small-world

☆ 99 Cited by 10319  Related articles All 54 versions

Finding and evaluating community structure in networks

[MEJ Newman](#), [M Girvan](#) - Physical review E, 2004 - APS

Abstract We propose and study a set of algorithms for discovering community structure in networks—natural divisions of network nodes into densely connected subgroups. Our algorithms all share two definitive features: first, they involve iterative removal of edges from

☆ 99 Cited by 9247  Related articles All 44 versions

Key: let's reverse our thinking of what Q does

Don't use Q to compute correlation of some given labels.

Instead, **experiment with the labels** and see how you can **maximize Q!**

Now, we have a computer science problem:
how do you search the space of partitions?

(This space is really big!)

How would you do it? 🤔

People like modularity. Why?

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{b_i, b_j}$$

- Intuitive
- Works for weighted and unweighted networks.
- Corresponds to our social network ideas of what (cohesive) communities are.
 - Automatically choose k , the number of groups.
 - Rapid approximate solutions.
 - Follows the usual methods trajectory: idea, demonstration, optimization.
- Fun customizations:
 - Resolution parameter to “zoom in” and “zoom out.”
 - Find the clusters. Then cluster the clusters. Then cluster those clusters...
 - Directed. Bipartite.

$$Q = \frac{1}{m} \sum_{ij} \left(A_{ij} - \frac{k_i^{\text{out}} k_j^{\text{in}}}{m} \right) \delta_{b_i, b_j}$$

modularity for directed networks

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta_{b_i, b_j}$$

modularity with a resolution parameter

Why aren't we done here?

Physicists like to minimize things because rocks fall.

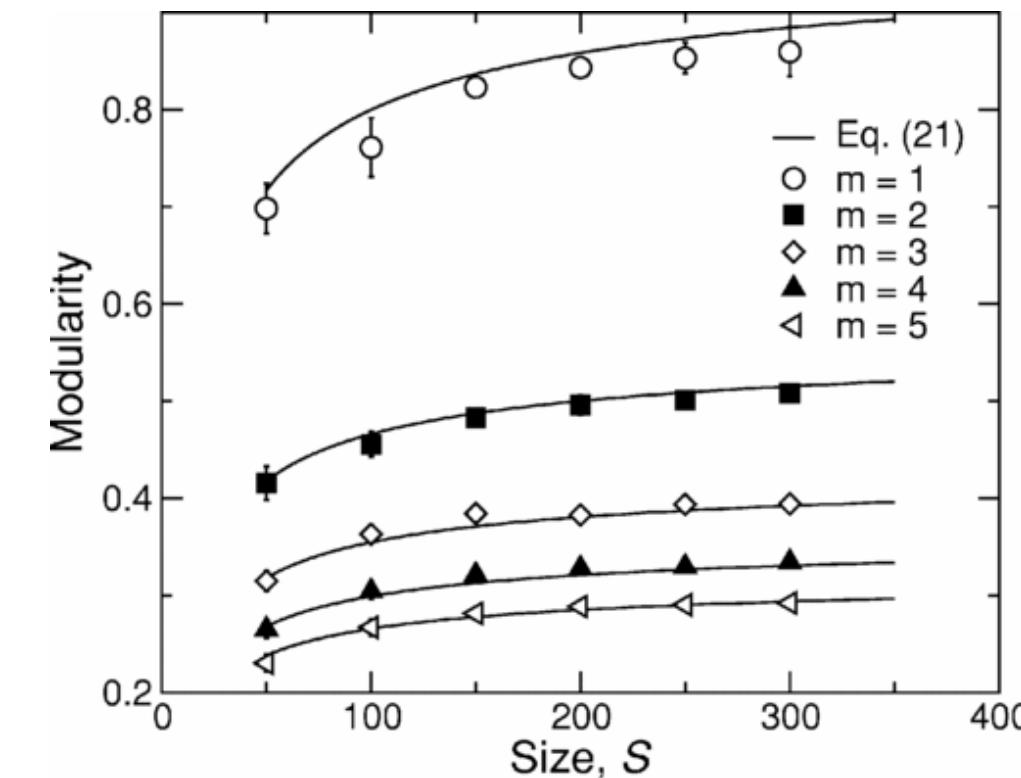
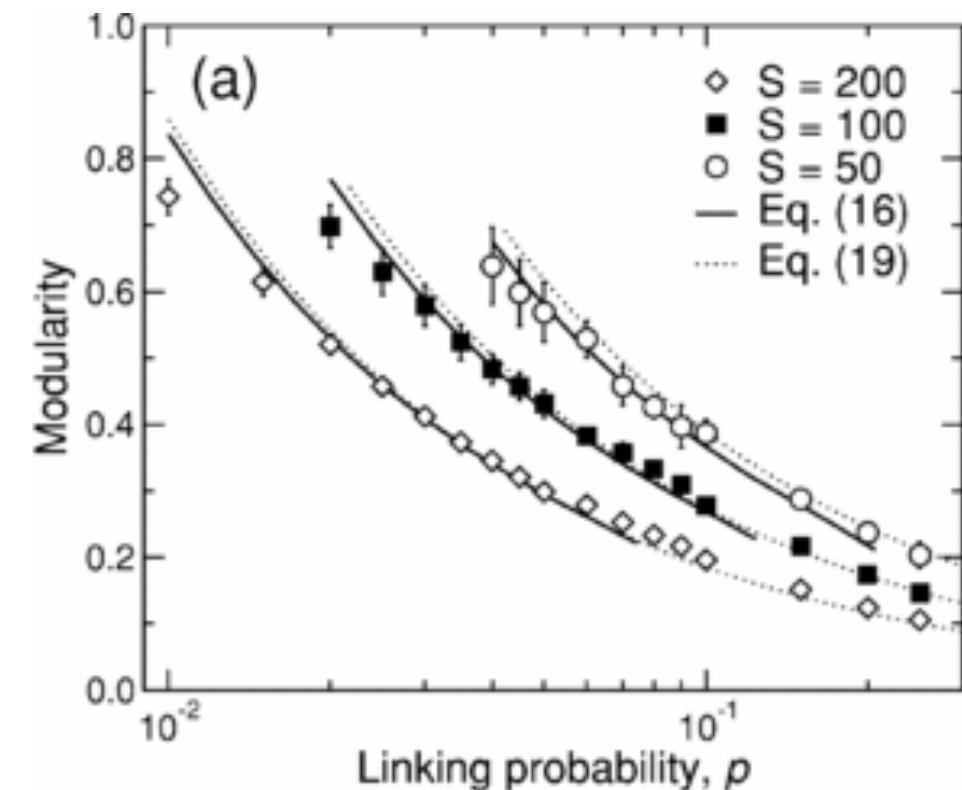
— Cris Moore

We can always maximize Q to find a partition, but is it meaningful?

Fooled by “structure” in totally random networks

As it turns out, you can find high-modularity partitions in random networks.

*Structure is what makes data different from noise.
It's what makes a network different from a random graph.*

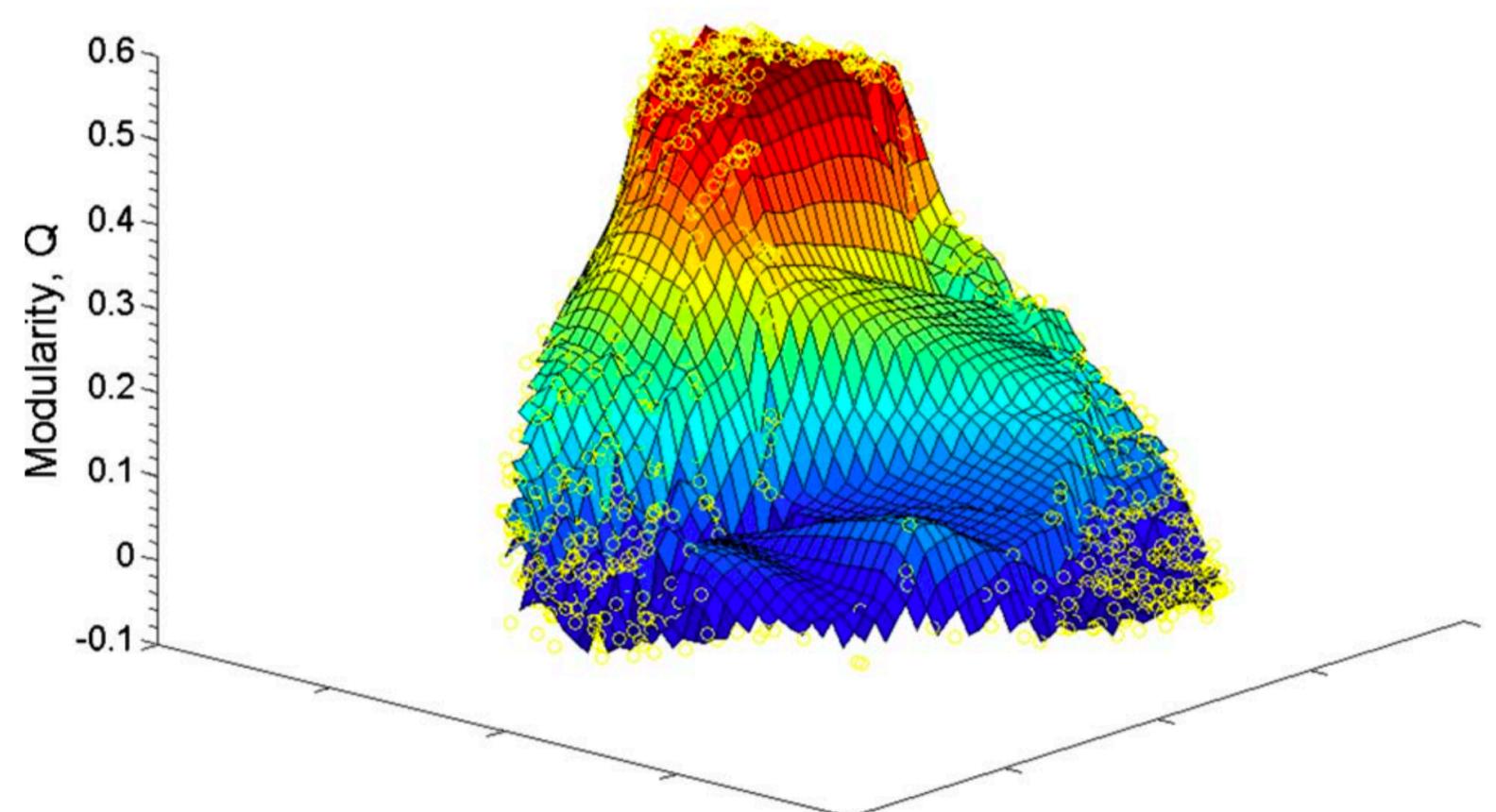
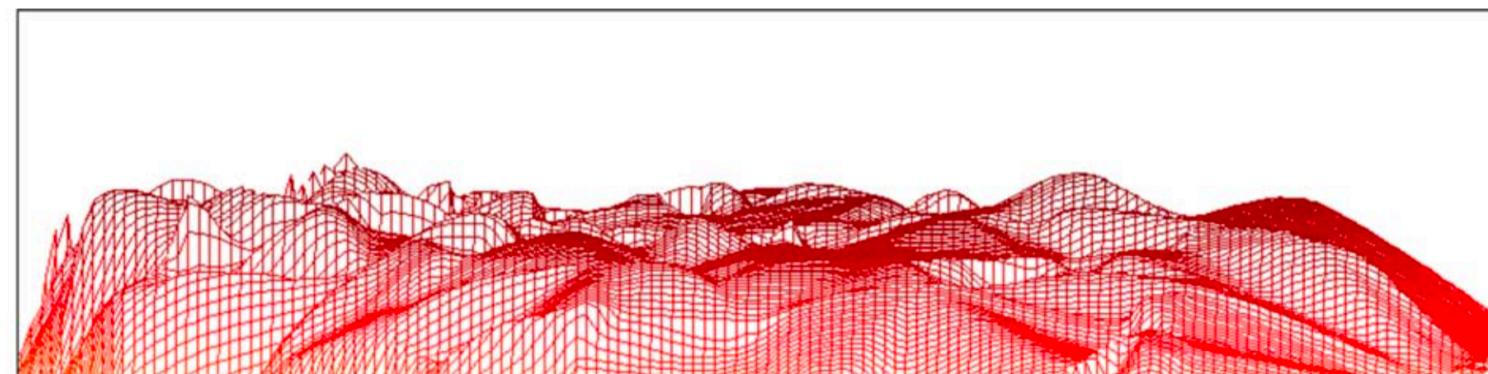


We prefer that our methods fail gracefully, and tell us when they fail. (like R^2)
[alternative perspective: maybe you want to find clusters in randomness?]

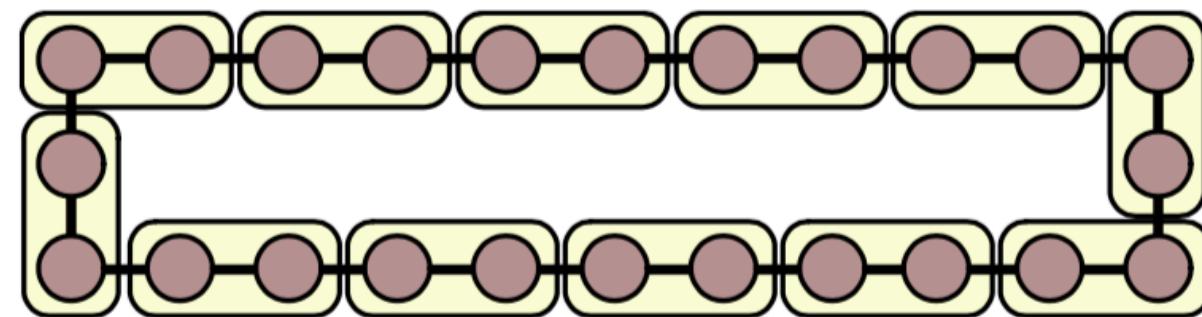
Modularity: degeneracy and strange behavior

Lots of different but nearly-as-good partitions.

The optimization landscape is *degenerate*.



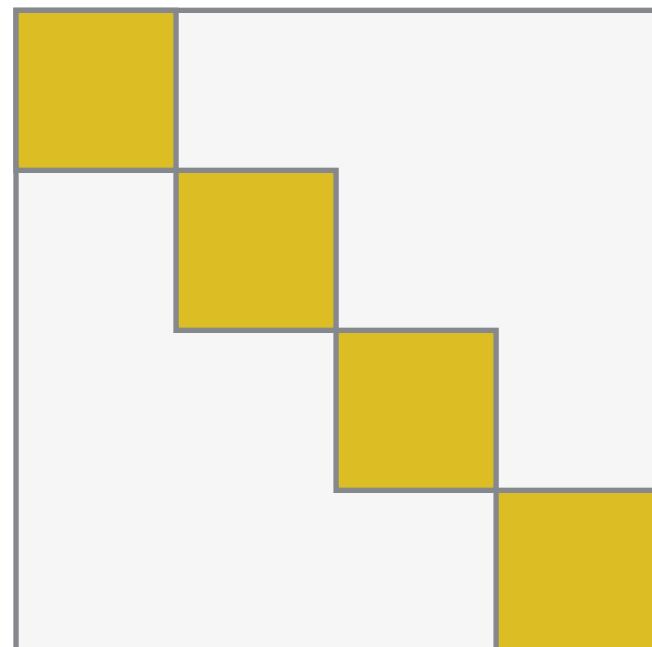
Unintuitive behavior: find the communities in a chain of cliques and you get pairs of cliques...not the cliques themselves!



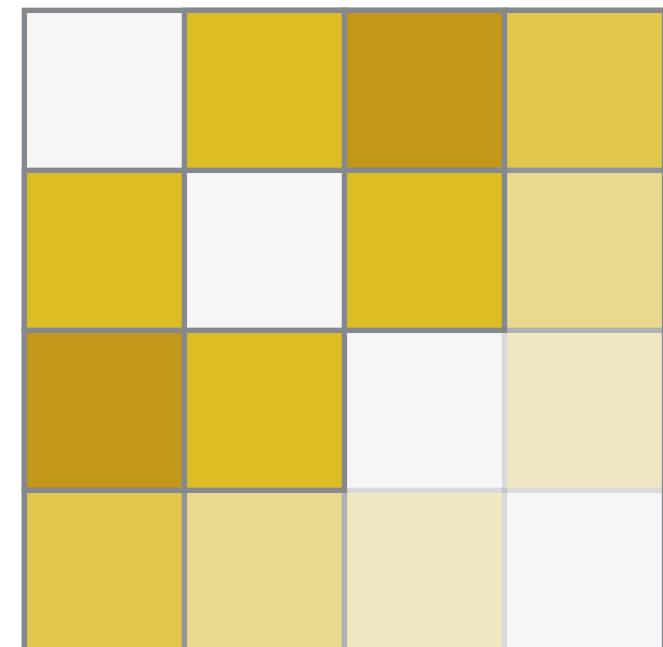
Q is restricted to *assortative* community structure

The zoo of possible structures is diverse and interesting!

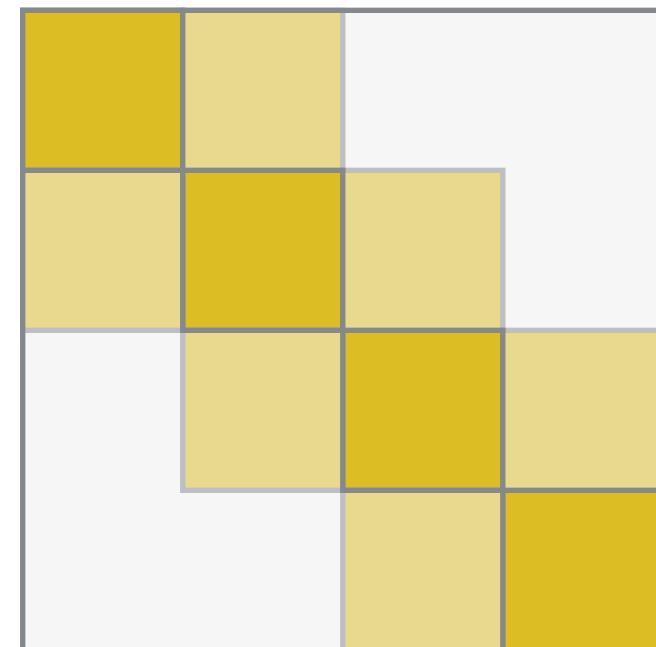
Build intuition: what do these networks look like?



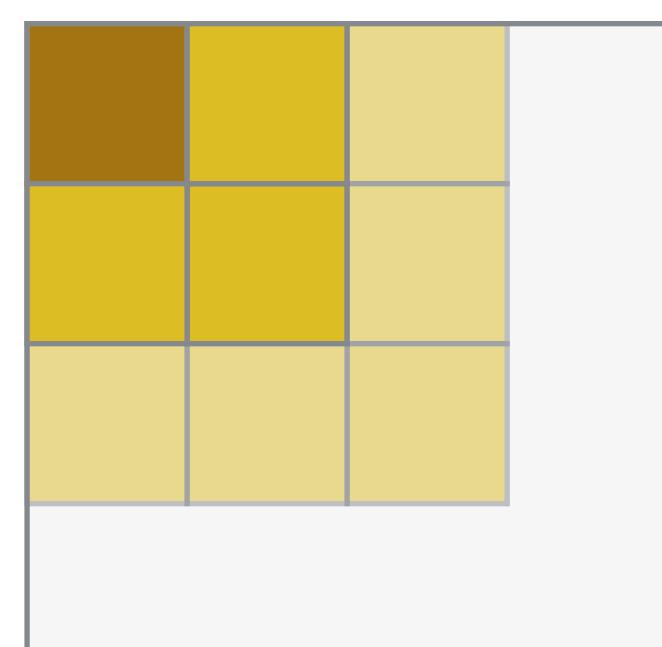
Assortative



Disassortative



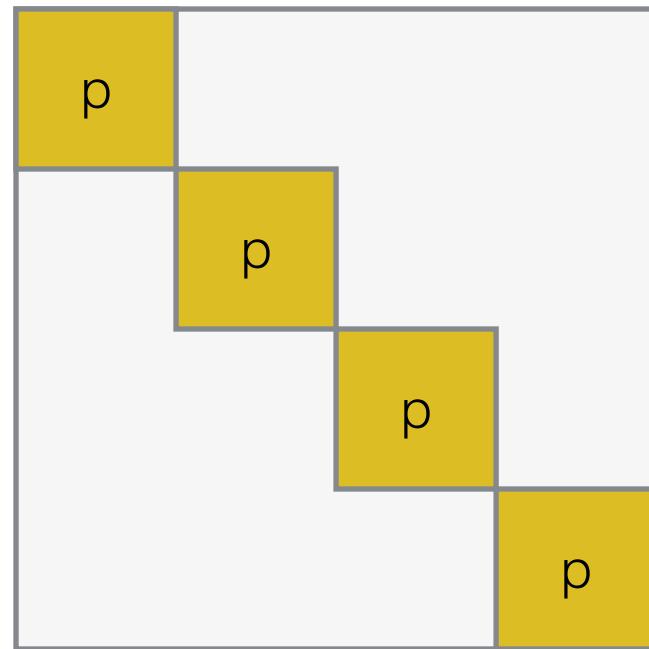
Ordered



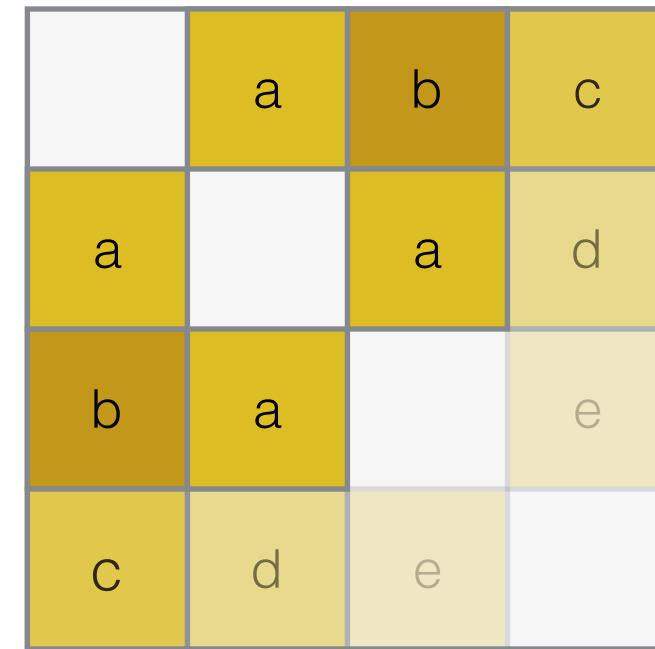
Core-periphery

Beyond assortativity: block models

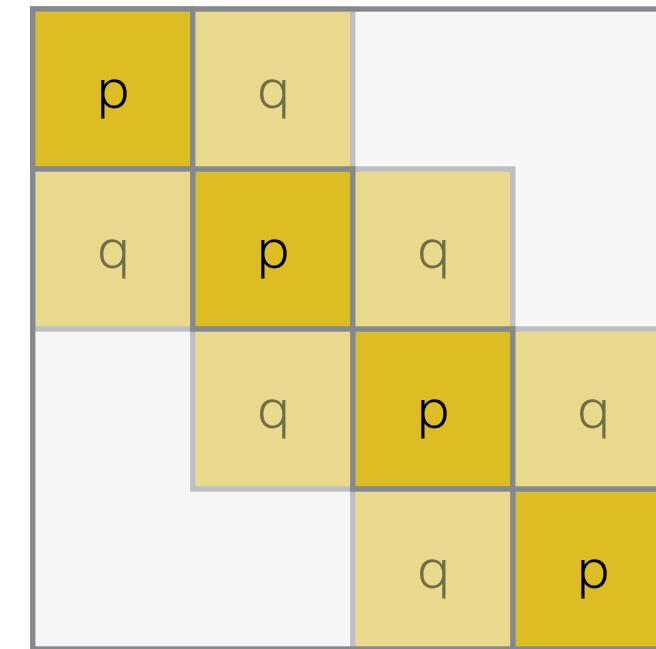
What do these have in common?



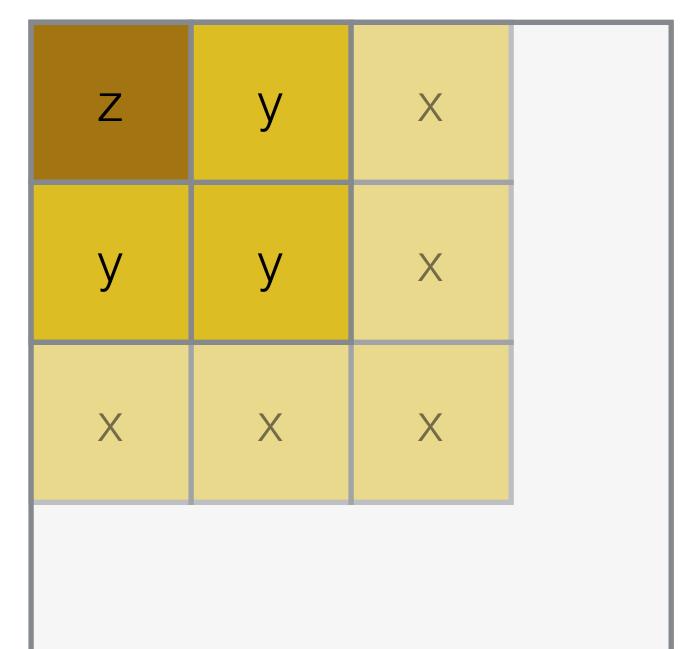
Assortative



Disassortative



Ordered



Core-periphery

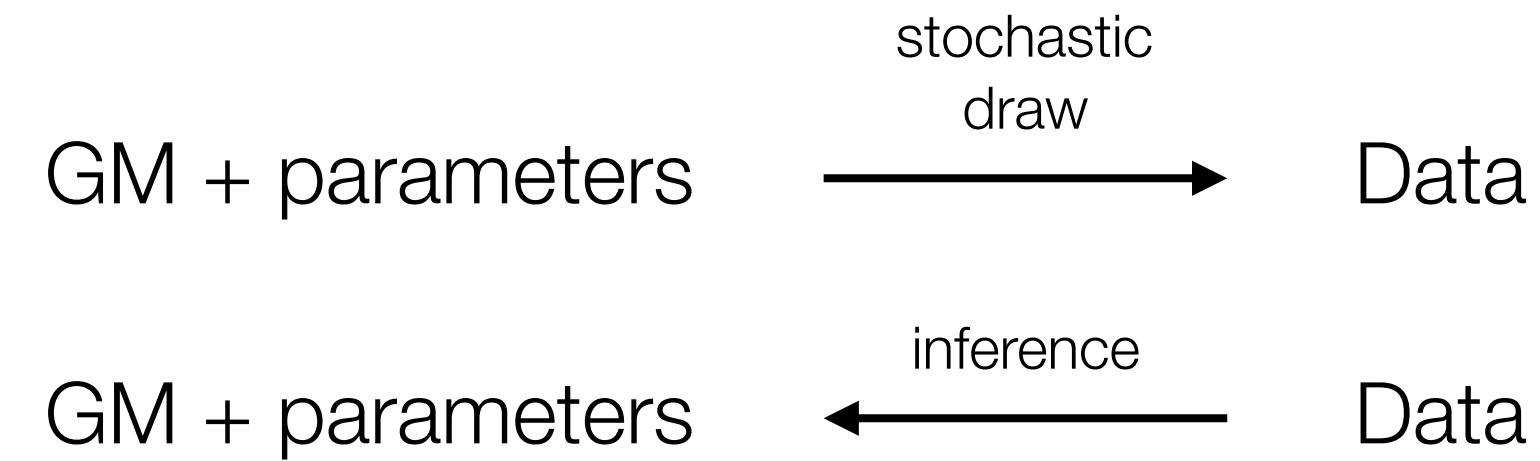
Nodes are in groups with other nodes that connect to other groups *in similar ways*.

Key idea: all nodes in a group are *stochastically equivalent*.

Generative model approach

Generate the structure you wish to infer.

We like generative models because they open the door to inference:



In other words: let's write down a recipe for generating block structure.



The stochastic block model

GM + parameters



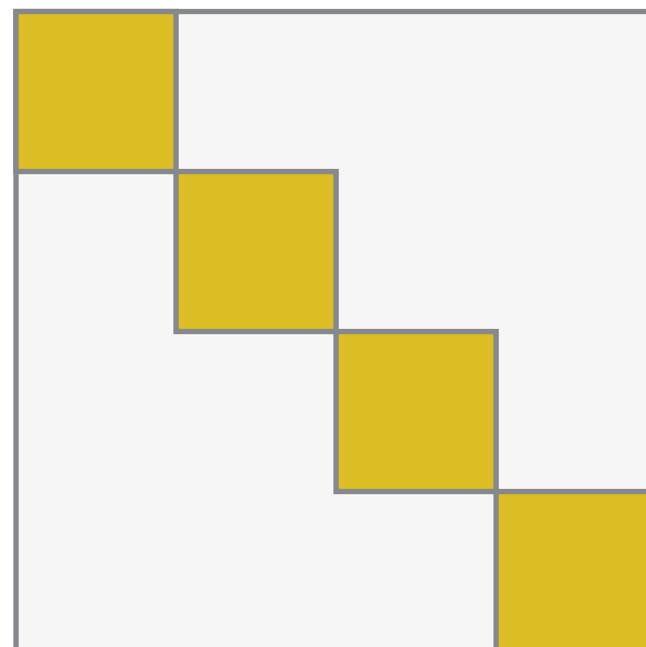
Assign each node to one of B blocks.

$$b_i$$

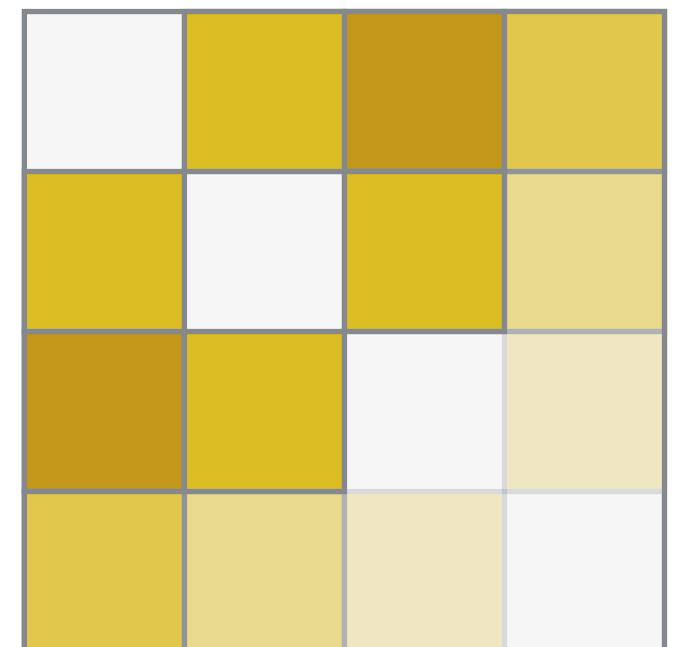
Let the probability that two nodes connect depend *only* on their blocks:

$$\Pr(A_{ij}|b_i, b_j) = \omega_{b_i, b_j}$$

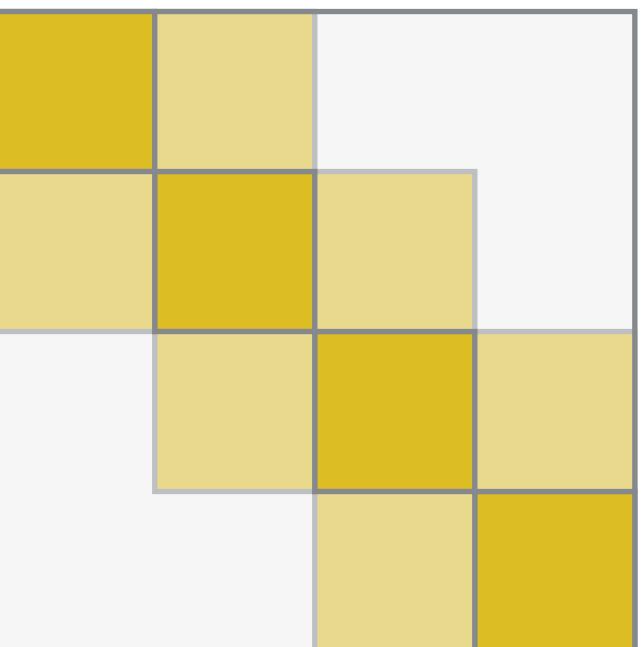
Then we can choose the matrix ω to have whatever structure we want!



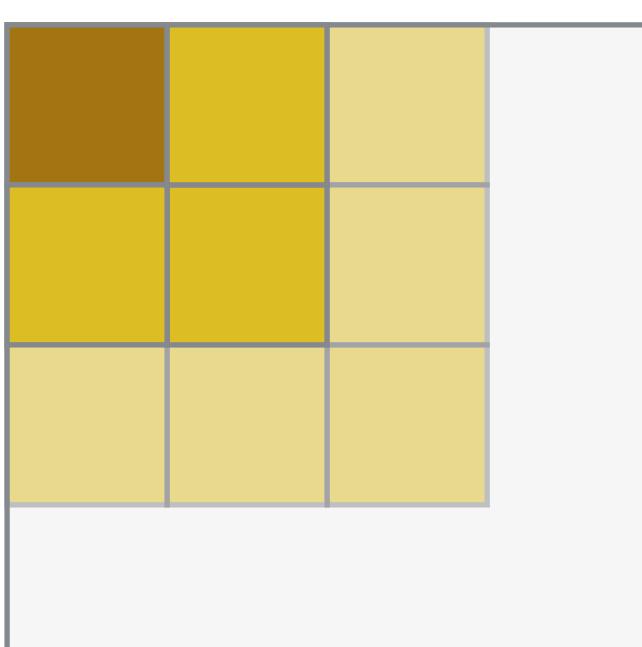
Assortative



Disassortative



Ordered



Core-periphery

SBM inference

GM + parameters



Data

no more math on slides 😭

(boringness prevention intervention)

Karrer, Newman. Stochastic blockmodels and community structure in networks.
Phys. Rev. E 83, 016107 (2011).

example matrix of parameters, $B=4$

Summary:

1. Write down the SBM *likelihood function* for a fixed number of blocks B .
2. Maximize the likelihood with respect to matrix parameters.
3. Search over divisions into B blocks to find the best blocks.

The problem with parameterized models...

You have to choose their parameters!

How shall we choose B , the number of blocks?

Hint: we can't simply maximize the likelihood over all choices of B :

Why? If we place each node in its own community, we can get Likelihood=1.
[Actually, this wouldn't model the data at all: it would *memorize* it.]

We need a way to penalize the complexity of the model. Any ideas?

Description length & Occam's razor

The Description Length of a message is:

bits required to send the compressed message + # bits in encoding scheme.

Occam's razor: among all possible explanation for a phenomenon, choose the simplest one. Therefore, choose the model with **Minimum Description Length** (MDL).

The stochastic block model also has a Description Length:

$$\Sigma = \boxed{\mathcal{S}} + \boxed{\mathcal{L}}$$

description length = entropy of data, given the model (fit SBM) + entropy of model

Consider the original problem:

what happens to this equation when I increase the number of blocks B?

MDL criterion suggests an algorithm:

Fit the SBM with 1 block and record the Description Length.

Fit the SBM with 2 blocks and record the Description Length.

...

when the Description Length starts to increase, go back one step and stop.*

Bonus: what happens if I *try to trick you* and give you a *random network with no blocks*?

MDL approach will tell you: your network is a random network with *one* block.

*Actually, use something clever, like Golden Ratio / Fibonacci search

Press et al. Numerical Recipes: The Art of Scientific Computing, (Cambridge University Press, Cambridge, England, 2007), 3rd ed.

So how does the search part work?

Markov-chain Monte Carlo:

Wander from one partition to another partition by proposing to take a node from one group and move it to a new group.

If this move increases the likelihood score, then keep the move.

If this move decreases the likelihood score, then maybe keep it, depending on how bad it is.

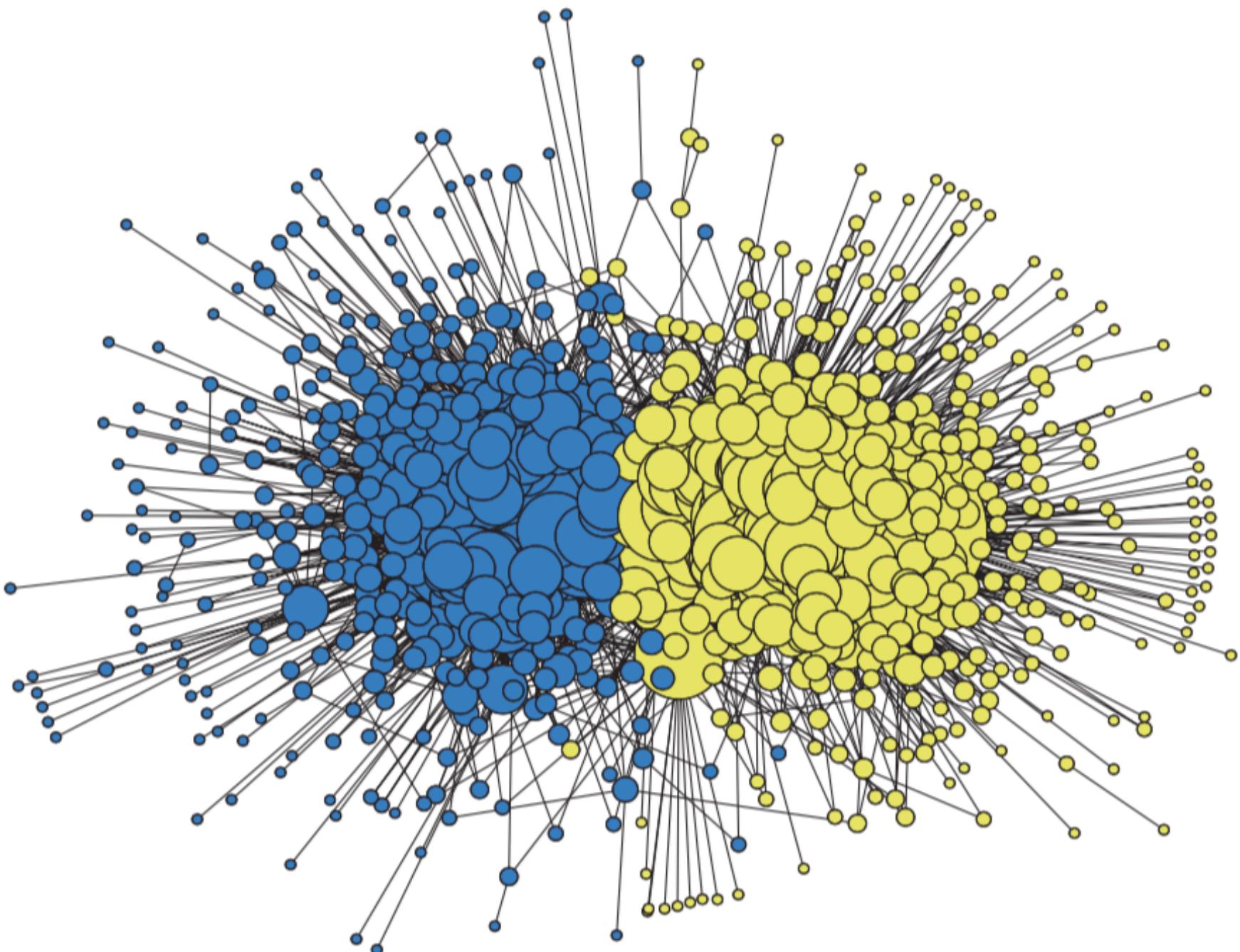
Thorough details in the documentation for graph-tool. <https://graph-tool.skewed.de>

Does it work?

Adamic & Glance mapped the link structure of USA political blogs in 2004.

Karrer & Newman used this network as a testbed for community detection using the SBM.

What does this say about the process that may be generating (or pruning?) the links in this network?

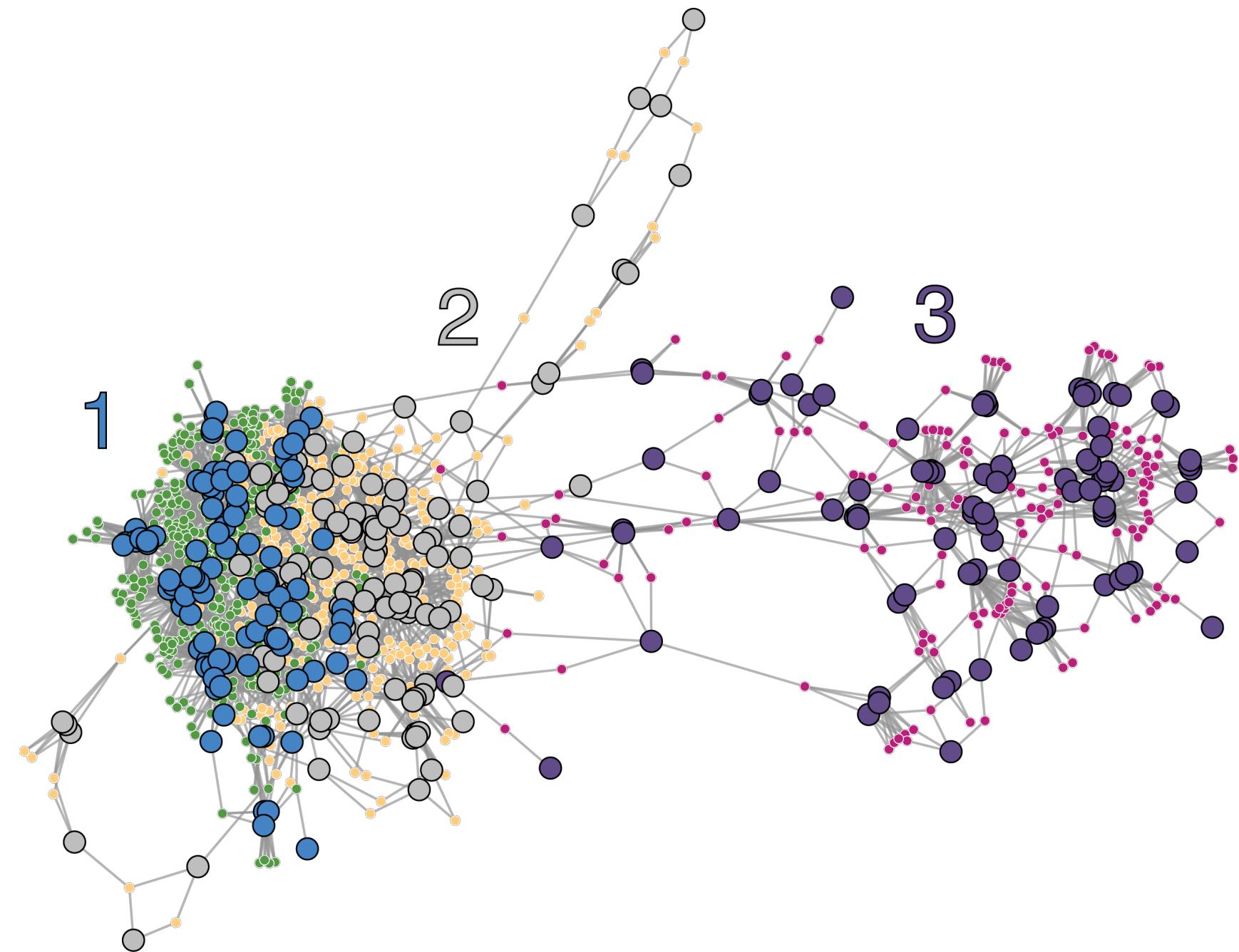


Does it work?

In bipartite networks, we know the major split in the data already.

Methodologically, we found that exploiting this split improved speed and quality of the partitions we found.

Scientifically, this opened new directions to analyze (and understand) evolutionary constraints on malaria parasites.



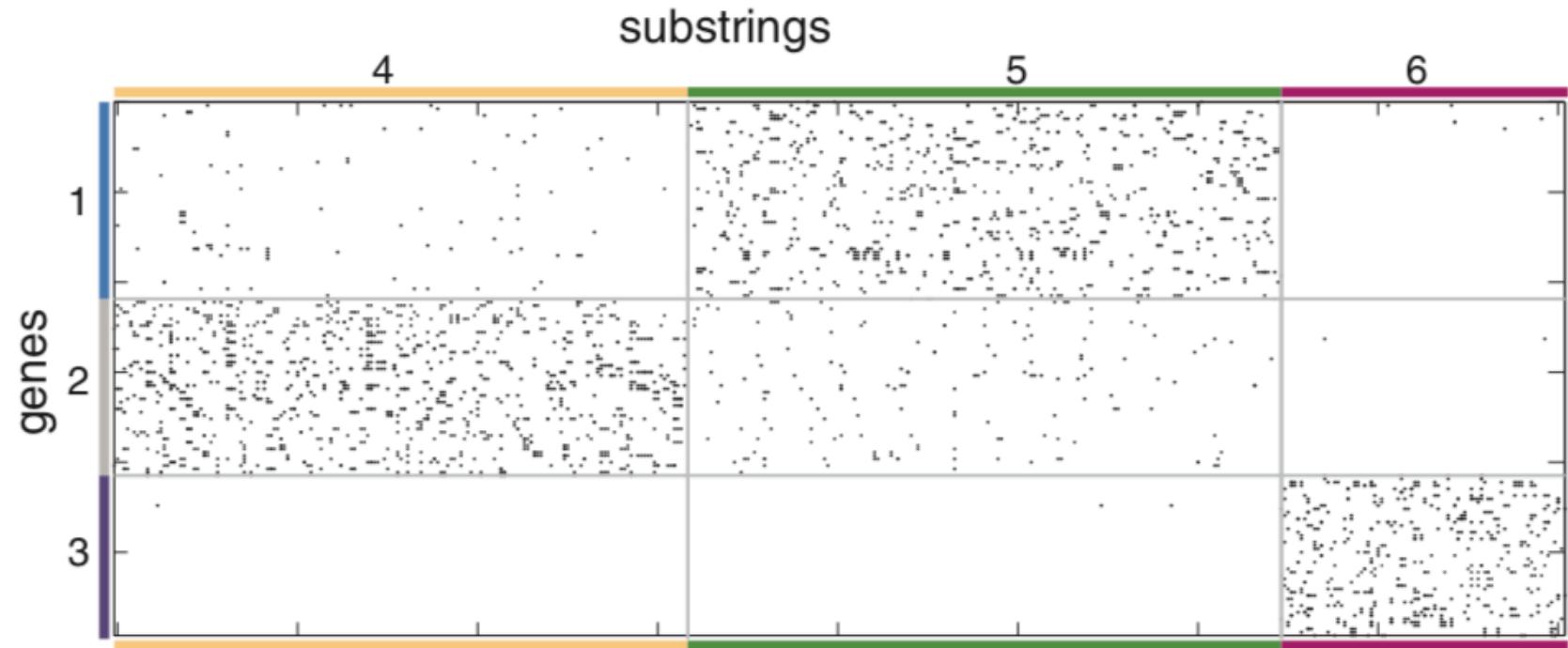
Genes & substrings,
malaria immune evasion

Does it work?

In bipartite networks, we *know* the major split in the data already.

Methodologically, we found that exploiting this split improved speed and quality of the partitions we found.

Scientifically, this opened new directions to analyze (and understand) evolutionary constraints on malaria parasites.



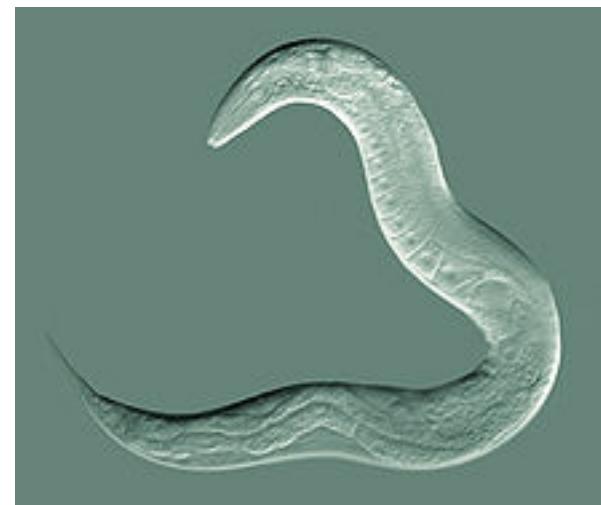
Genes & substrings,
malaria immune evasion

Larremore, Clauset, Buckee, *PLoS Comp Biol*, 2013.

Larremore, Clauset, Jacobs, *Physical Review E*, 2014.

Does it work?

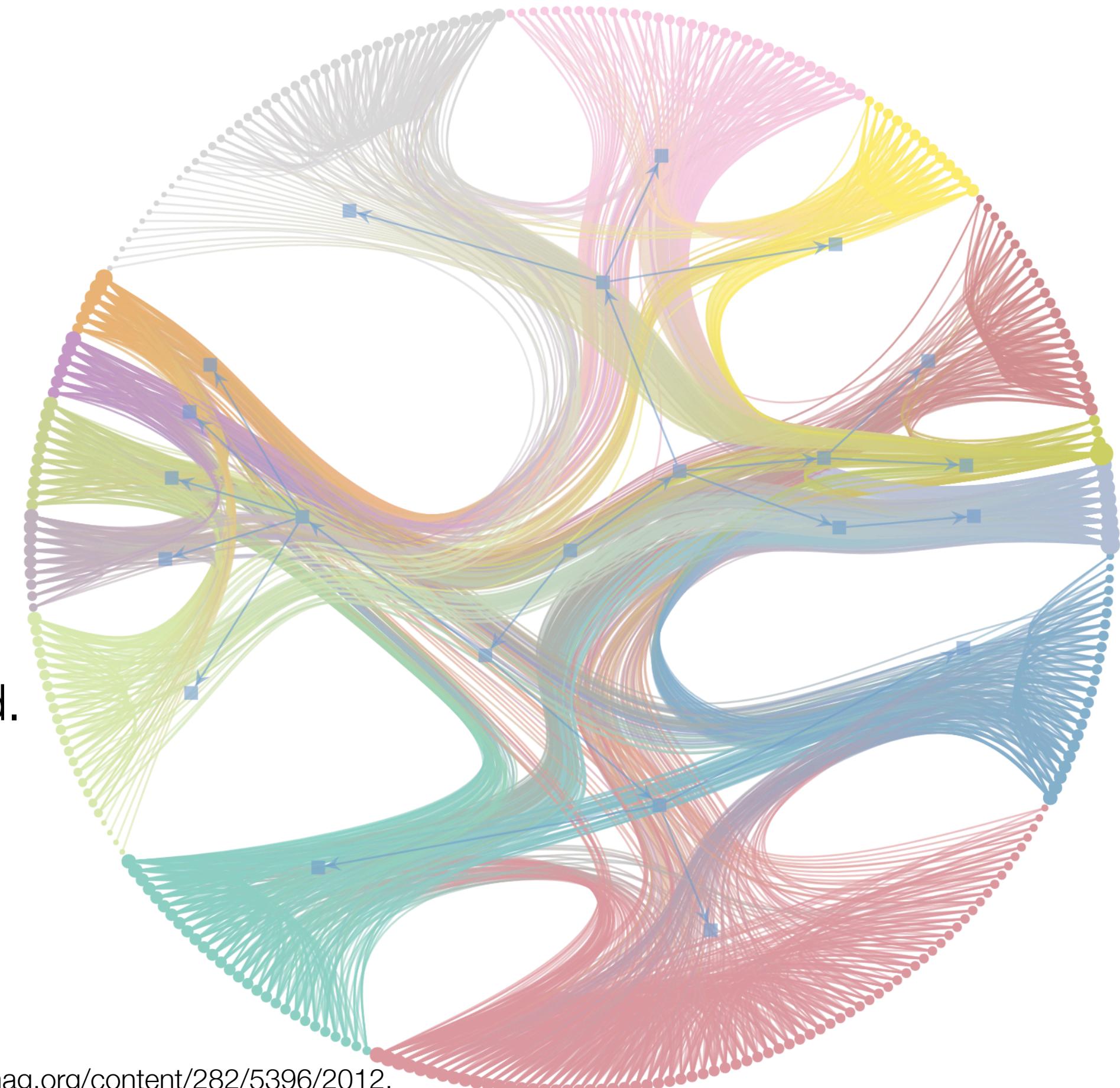
C. elegans neuronal network.



297 neurons, completely mapped.

The neurons do not fire action potentials, and do not express any voltage-gated ion channels.

Note the different layout...



A good alternative? cross-validation via link prediction

Select B by choosing the model that makes the best predictions.

Perform k -fold cross validation:

1. Divide the edges of the network into k groups, called folds.
2. Hide one of the folds (the “test set”)
3. Fit each SBM to the remaining $k-1$ folds (the “training set”), varying B .
4. Test the ability of the fitted models to predict the hidden test data.
5. Switch which fold is “test” and which are “training” and repeat.

Choose the B with the highest performance on link prediction over all k folds.

Note: this is exactly how we write [math] exams, minus the “repeat”.

Degree Correction

It turns out that I hid something from you before about the stochastic block model.

Rather than generate block-wise random matrices, we want to be able to do so but also choose the degrees of the nodes.

Generative model perspective: if we can't generate communities/networks with broad degree distributions, we can't find those communities in real data.

What can be done? **Correct for (or control for) the degrees of the nodes.**

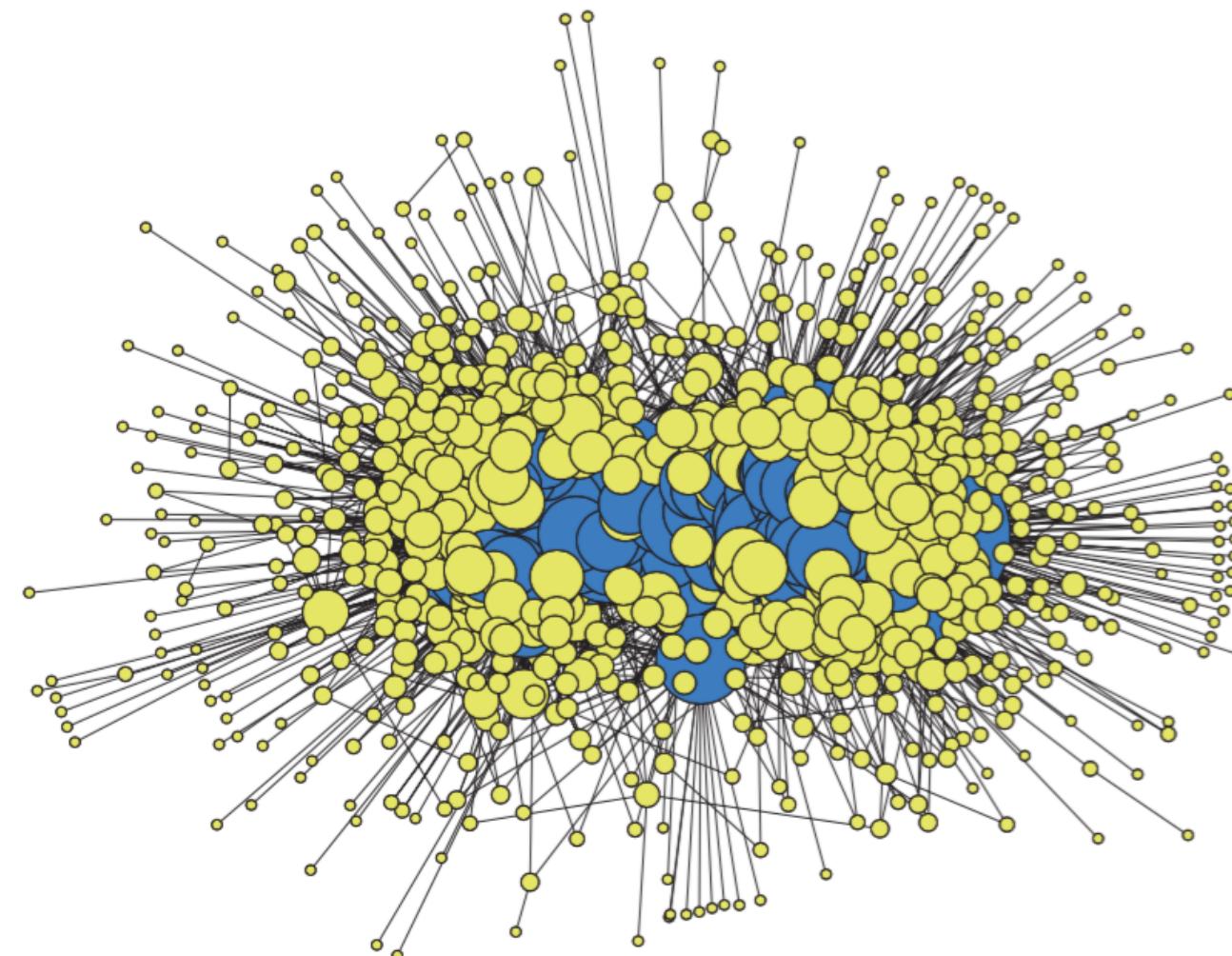
Degree Correction

no more math on slides 😭

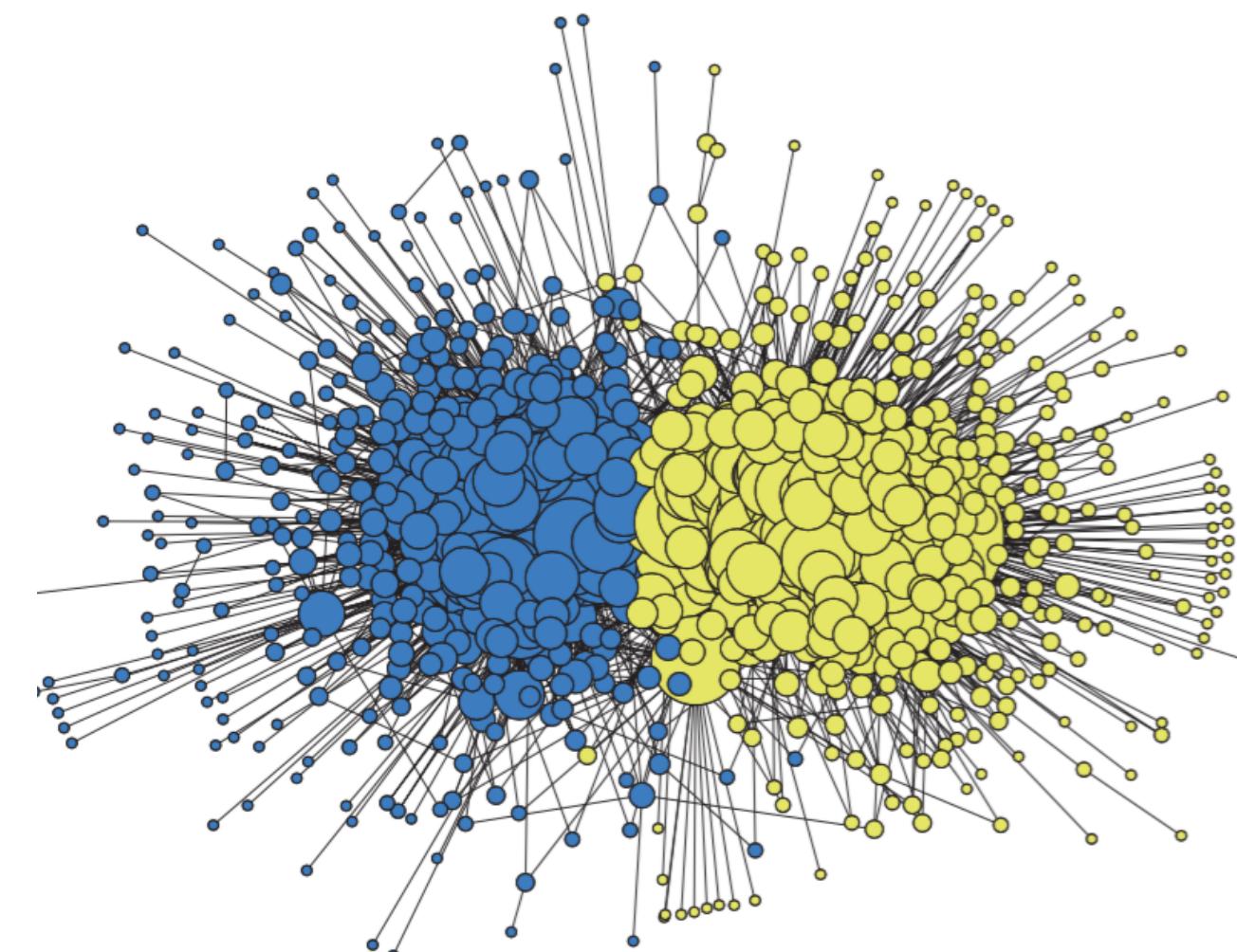
(boringness prevention intervention)

Karrer, Newman. Stochastic blockmodels and community structure in networks.
Phys. Rev. E 83, 016107 (2011).

SBM vs “DC” SBM



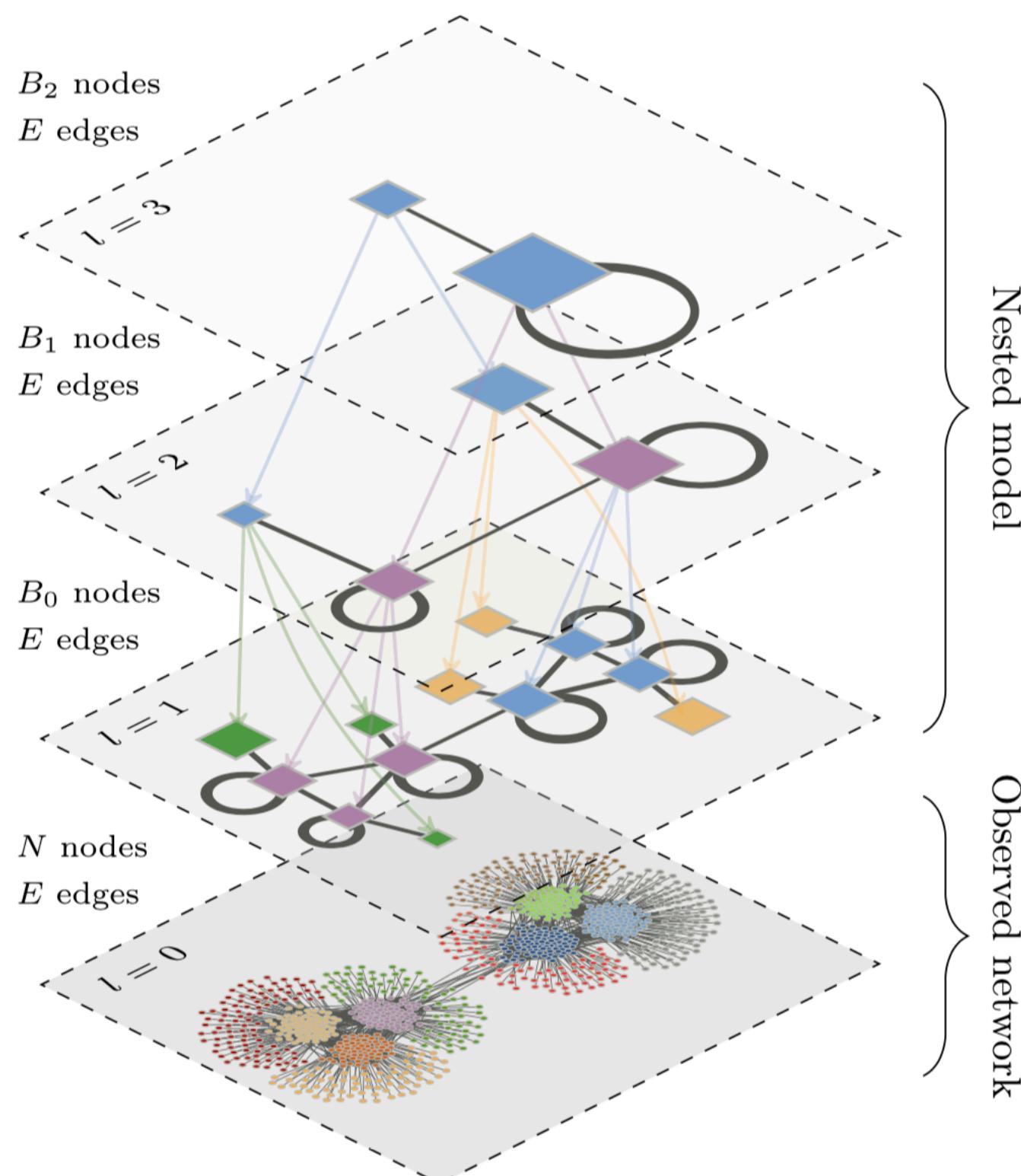
(a) Without degree-correction



(b) With degree-correction

The degree-corrected stochastic block model (right) finds communities which correspond to known metadata.

Advanced topic 1: hierarchical communities



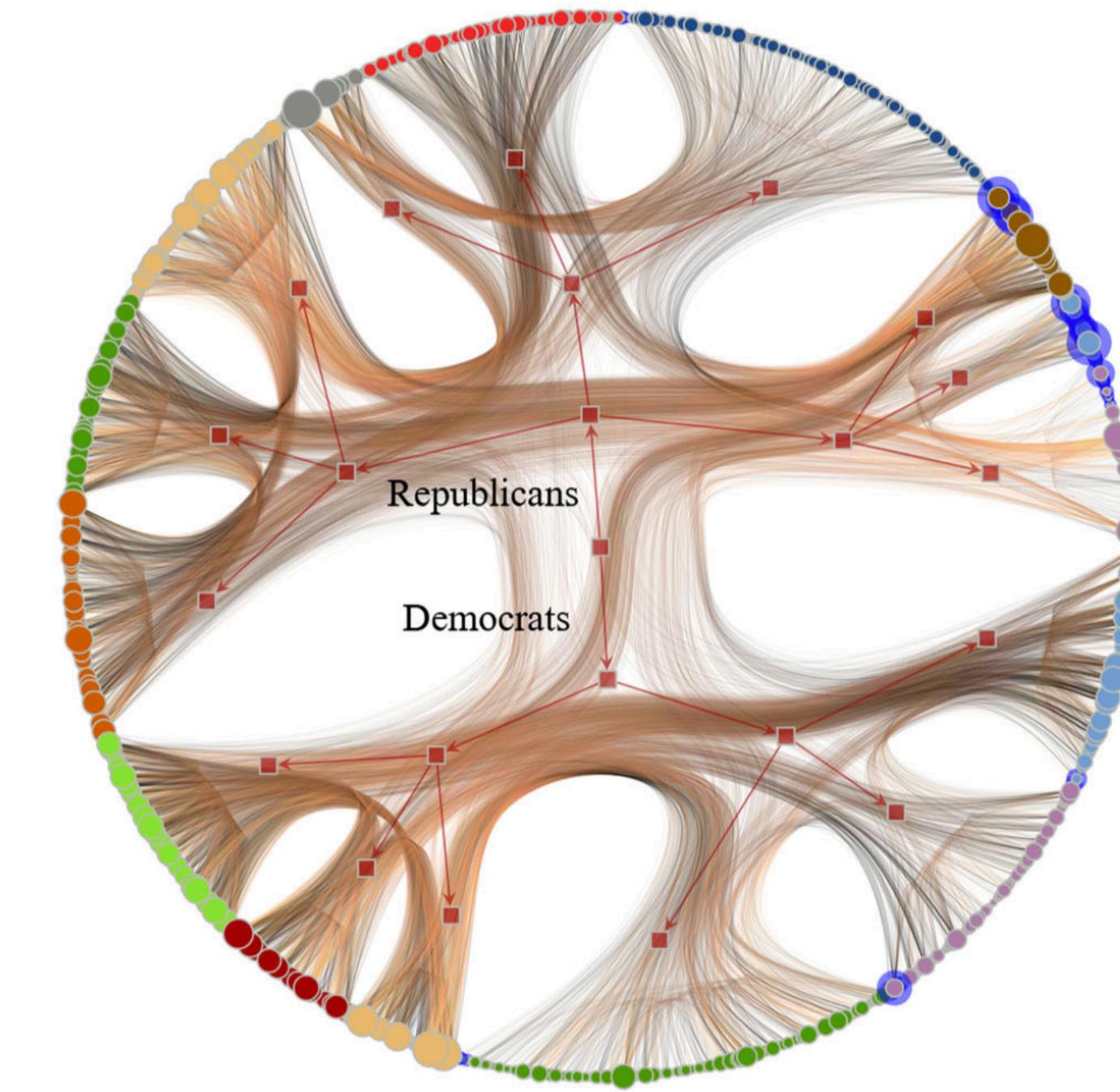
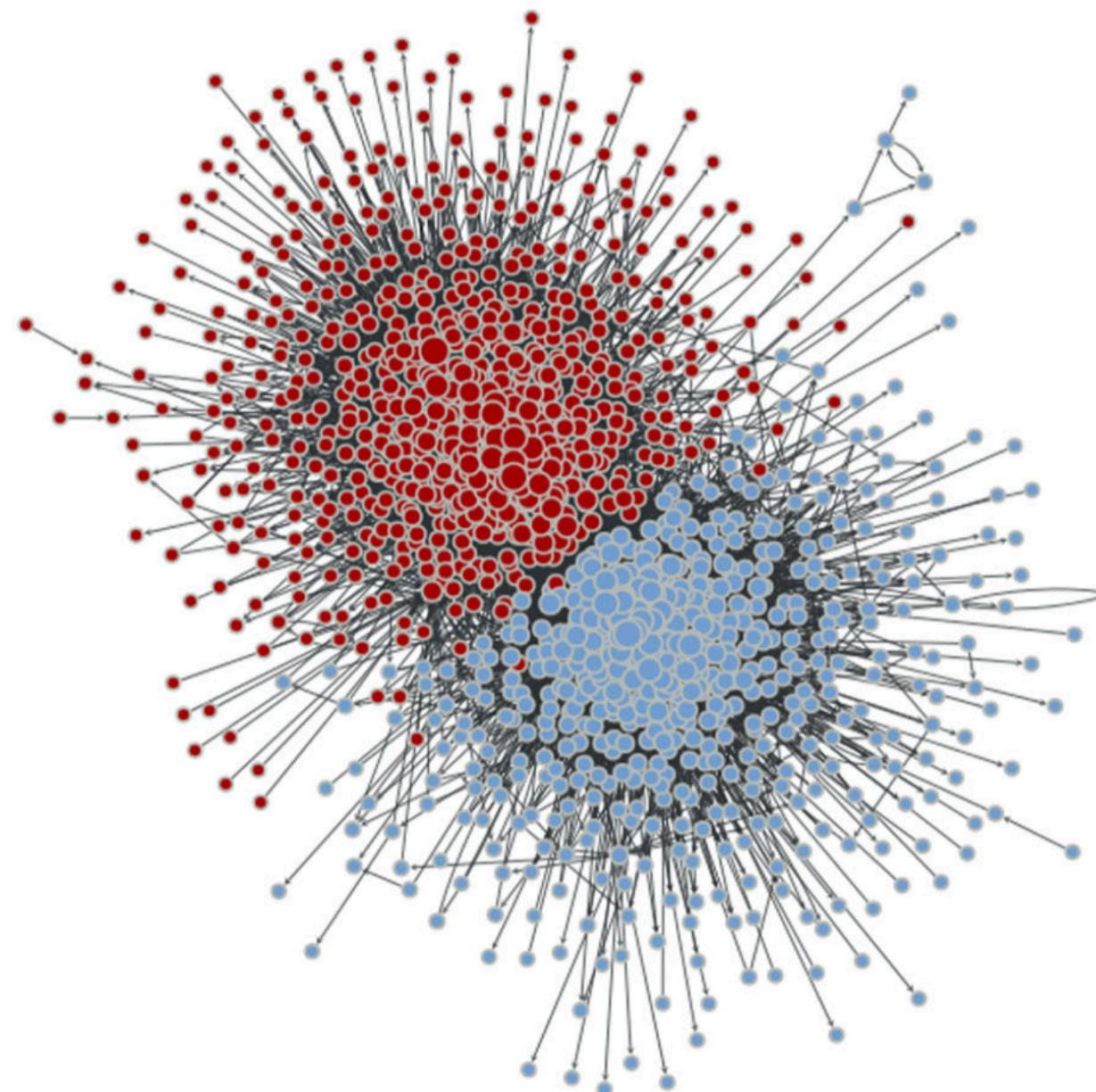
Don't minimize the description length
of the data and the model...

Model the model as well. Why?

If we compress the model, we can
afford a bigger model, but a lower
overall cost.

Except now, the description length
includes two models. Or three? Or?

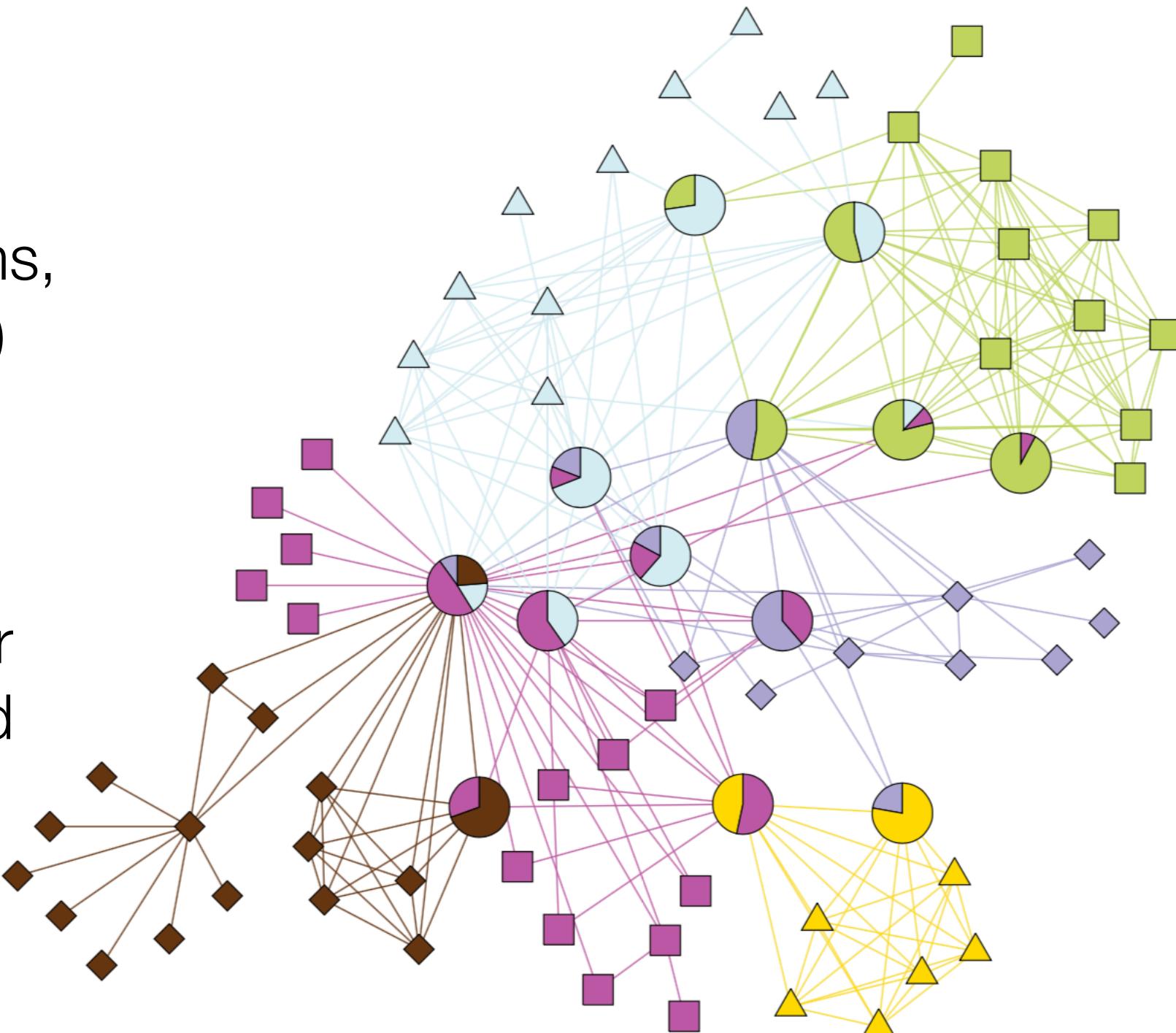
Advanced topic 1: hierarchical communities



Advanced topic 2: mixed-membership

Nodes are often pulled between communities. (Or in real social systems, individuals belong to multiple groups.)

“Mixed membership” models allow for that, by assigning *links* to groups, and assigning nodes to groups based on their links.



Advanced Topic 3: multilayer networks

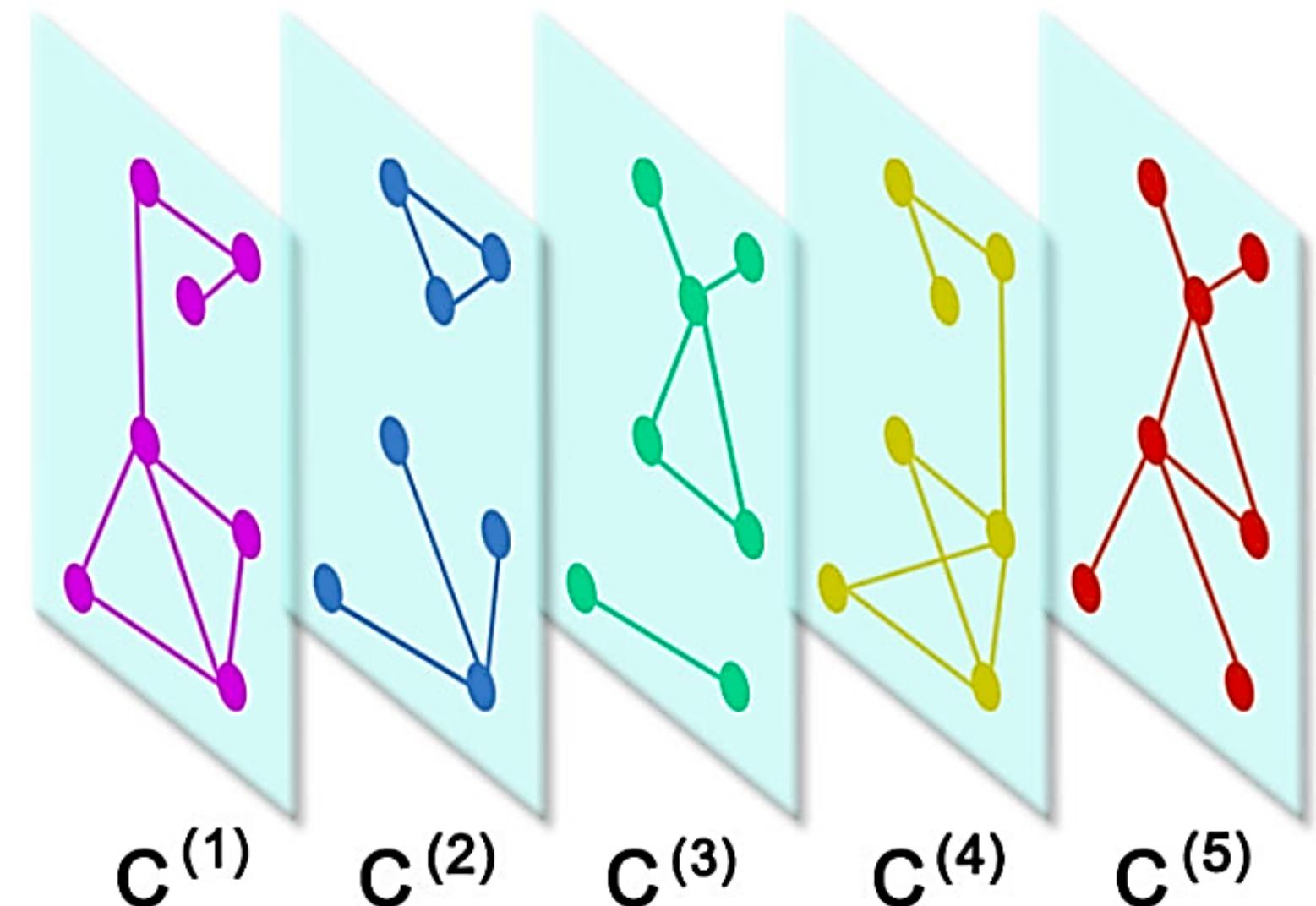
In single-layer networks:
nodes and edges

In multilayer networks:
nodes, edges, and layers

edges: different types of relationships

layers: each layer contains all edges of one type

nodes: same nodes in each layer



Multilayer network: air travel



Ryanair

Lufthansa

Vueling

British airways

Aggregate



traditional: booking with airline



disrupted: booking with kayak, expedia, etc

Multilayer network: community structure?

three key approaches:

1. **Non-generative**: modularity maximization; vary inter-layer strength.

Mucha et al *Science* 2010. <http://science.sciencemag.org/content/328/5980/876>

2. **Generative**: SBM for each layer, but jointly model layers whenever their structures are sufficiently similar.

Peixoto, T. P. *Phys. Rev. E* 92, 042807–15 (2015).

3. **Generative**: SBM for each layer, and model all layers simultaneously with same community structure, but allow relationships between groups to vary.

De Bacco Power Larremore Moore. *Phys. Rev. E* 95, 1981–10 (2017).

1 is preferred if nodes appear/disappear over time.

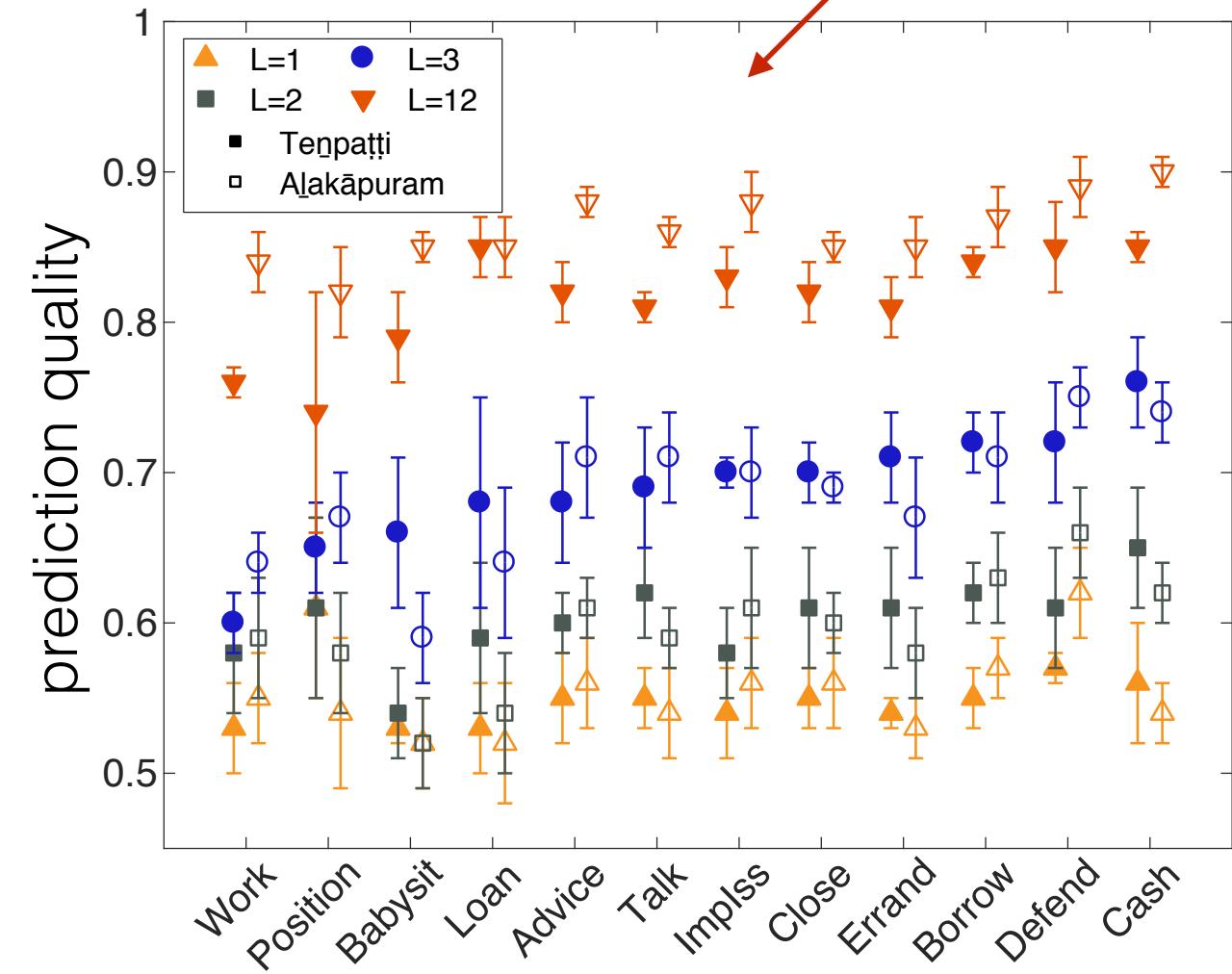
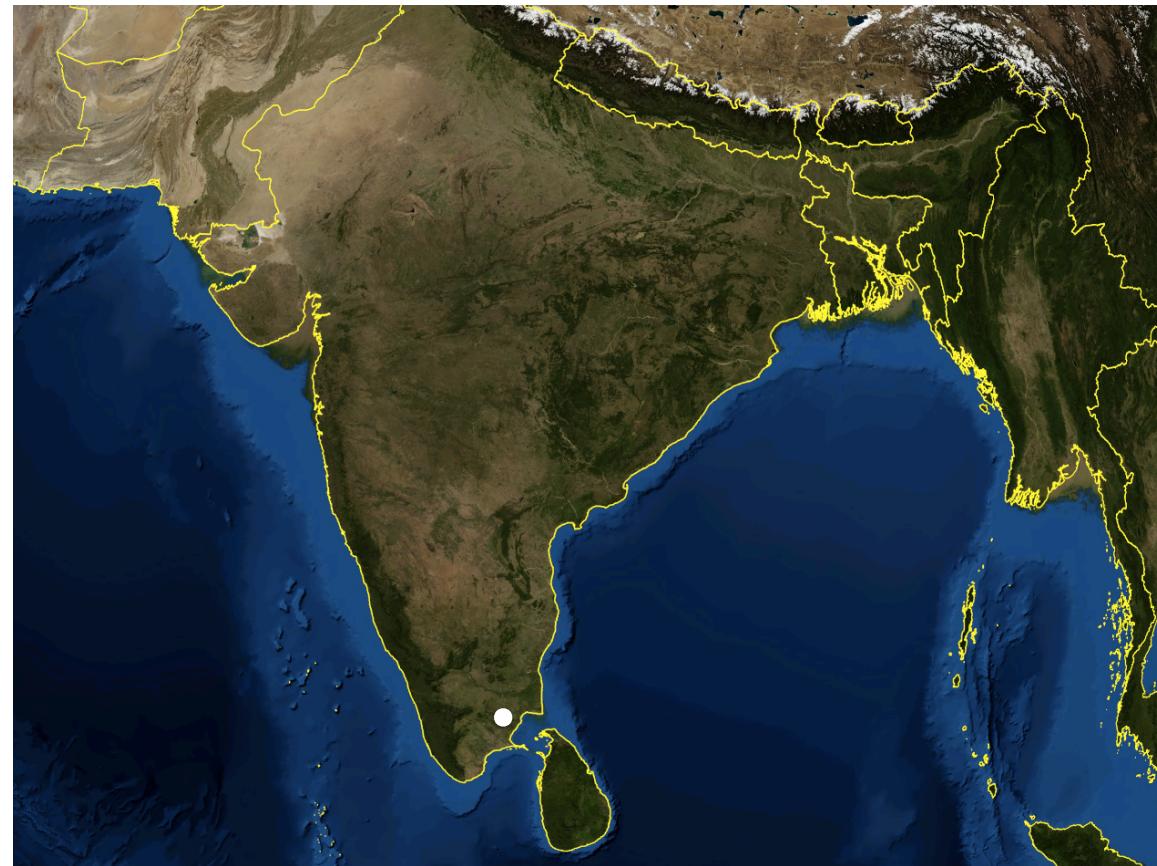
2,3 are preferred to solve the *layer interdependence problem*

Layer interdependence

more layers = better performance
(layer structure generated by same social mech.)

Are layers structurally similar? Complementary? Neither?

“Learn” a SBM from m layers; try to predict links of $m+1$.

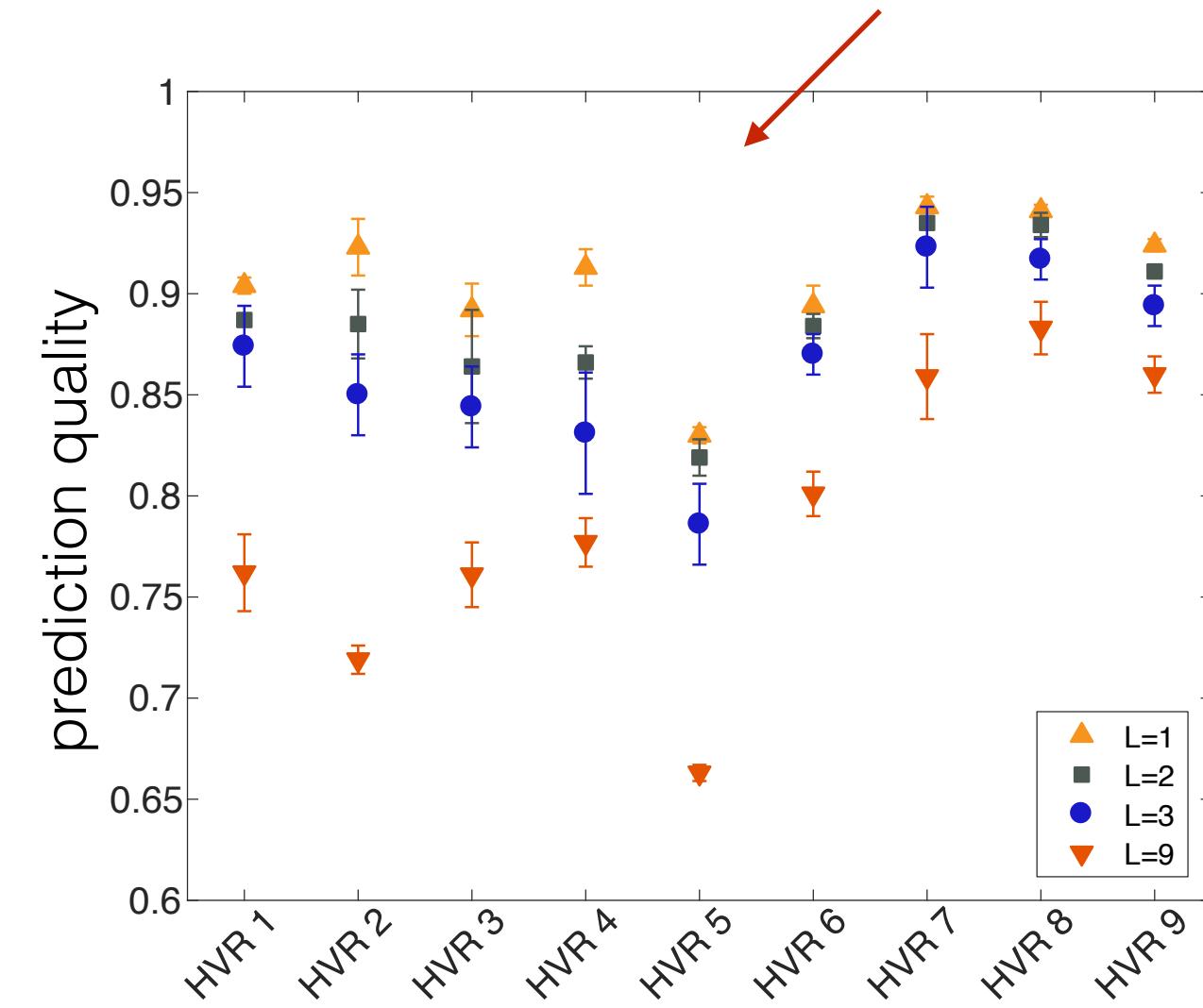


12 layer social support network across 2 villages in South India.

Layer interdependence - malaria



more layers = worse performance
(layer structure generated by different biol. mech.)



cannot predict the structure of one region in the immune-evasion genes
by using other regions; layers are unrelated!

Advanced topic 4: metadata+communities

What are metadata?

How well do metadata explain the network structure? “BESTest”

Peel*, Larremore*, Clauset. *Science Advances* 3(5) e1602548. (2017) <http://advances.sciencemag.org/content/3/5/e1602548.full>

How do metadata relate to network structure? “neoSBM”

Peel*, Larremore*, Clauset. *Science Advances* 3(5) e1602548. (2017) <http://advances.sciencemag.org/content/3/5/e1602548.full>

Can we use metadata to guide community detection? “metadata assisted SBM”

Newman, Clauset. *Nature communications* 7 (2016). <https://www.nature.com/ncomms/2016/160616/ncomms11863/full/ncomms11863.html>

Can we find patterns in the metadata itself? Apply multilayer SBM

Peixoto. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. *Phys. Rev. X* 4, 011047 (2014).

Blockmodel entropy significance test

How well do the metadata explain the network?

randomly assigned metadata
→ model gives no explanation, high H

metadata correlated with communities
→ model gives good explanation, low H

1. Divide the network G into groups according to metadata labels M .
2. Fit the maximum likelihood parameters of an *a posteriori* SBM and compute the entropy $H(G,M)$ of the corresponding ensemble.
3. Compare the entropy of this SBM ensemble to distribution of entropies from SBMs partitioned using shuffled metadata \underline{M} .

$$\text{p-value} = \Pr[H(G,\{\underline{M}\}) \leq H(G,M)]$$

Multiple network layers; multiple metadata attributes

Network	Status	Gender	Office	Practice	Law School
Friendship	$< 10^{-6}$	0.034	$< 10^{-6}$	0.033	0.134
Cowork	$< 10^{-3}$	0.094	$< 10^{-6}$	$< 10^{-6}$	0.922
Advice	$< 10^{-6}$	0.010	$< 10^{-6}$	$< 10^{-6}$	0.205

model = SBM

Multiple sets of metadata **significantly explain** multiple networks.
[Should one particular set of metadata be ground truth?]

BESTest accommodates many models of group structure

Network	Model	
	SBM	DCSBM
Malaria 1	0.566	0.066
Malaria 2	0.064	0.126
Malaria 3	0.536	0.415
Malaria 4	0.588	0.570
Malaria 5	0.382	0.097
Malaria 6	0.275	0.817
Malaria 7	0.020	0.437
Malaria 8	0.464	0.143
Malaria 9	0.115	0.104

metadata = parasite origin

A negative result: parasite origin is irrelevant to genetic substring-sharing.

Malaria parasites *do not* have a strong strain structure, with implications for diversifying selection among parasites.

neoSBM

Choose between the **SBM partition** and the **metadata partition**.

$$\mathcal{L}_{\text{neoSBM}} = \mathcal{L}_{\text{SBM}} + f(\theta)$$

neoSBM log likelihood SBM log likelihood cost

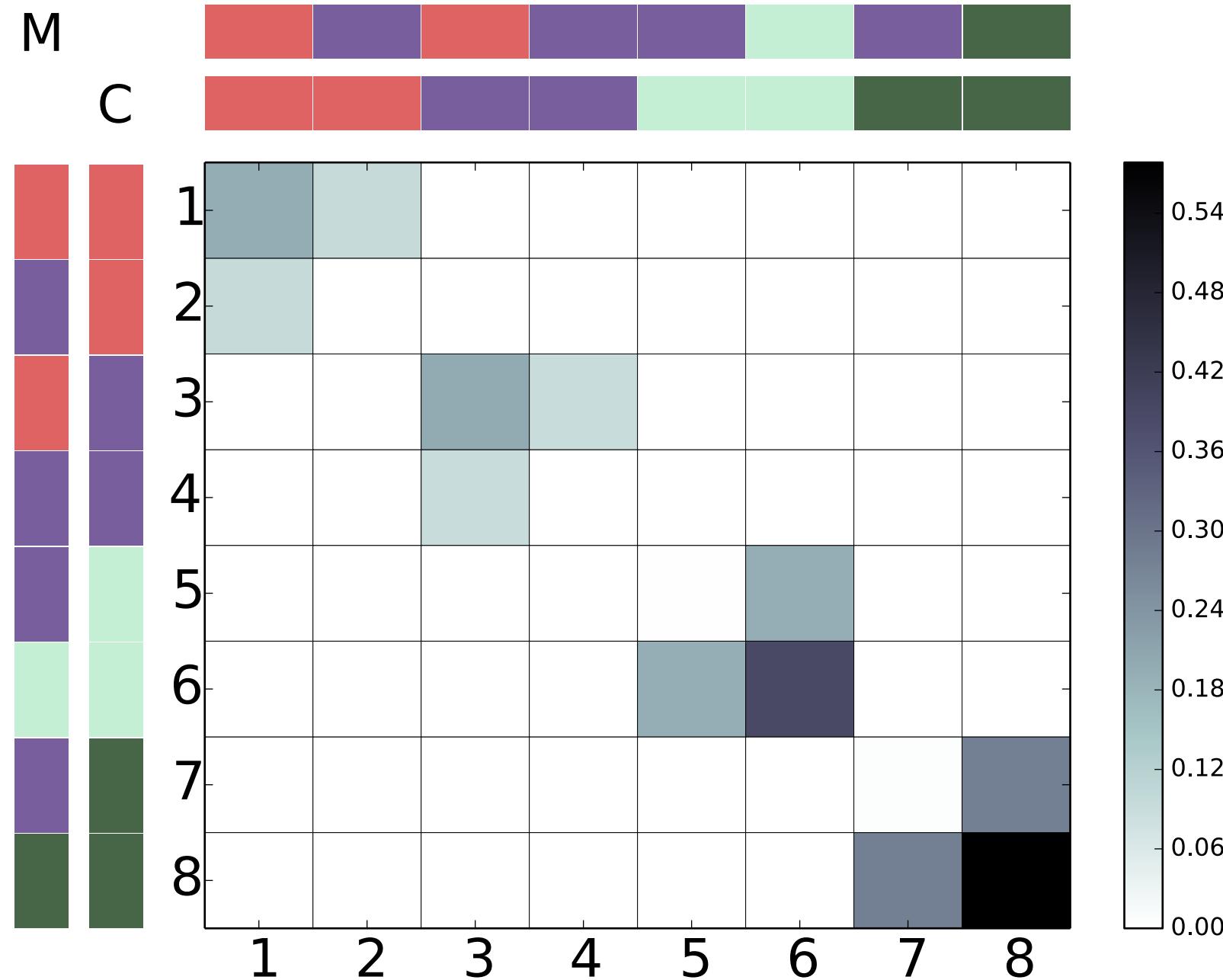
Log likelihood with parameterized prior:

θ is the parameter of a Bernoulli prior on whether the node is **free to choose its own community** or held **fixed at its metadata label**.

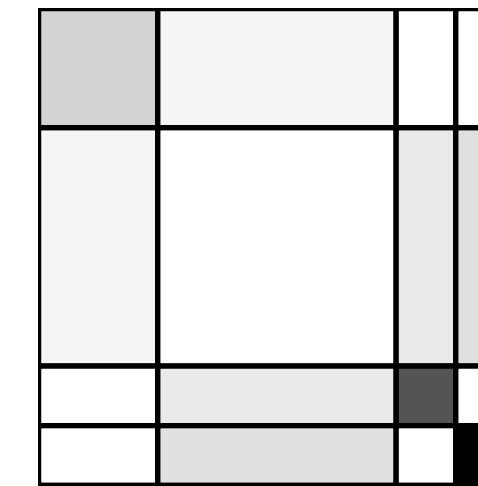
As θ increases, the cost of freeing a node decreases.

Varying θ in the unit interval **explores the space of partitions** between M and C .

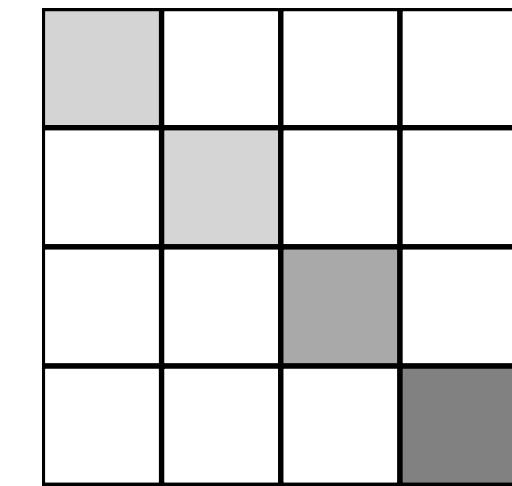
Plant two different *kinds* of structure in a network



SBM with 8 groups and
two interesting 4-group partitions:

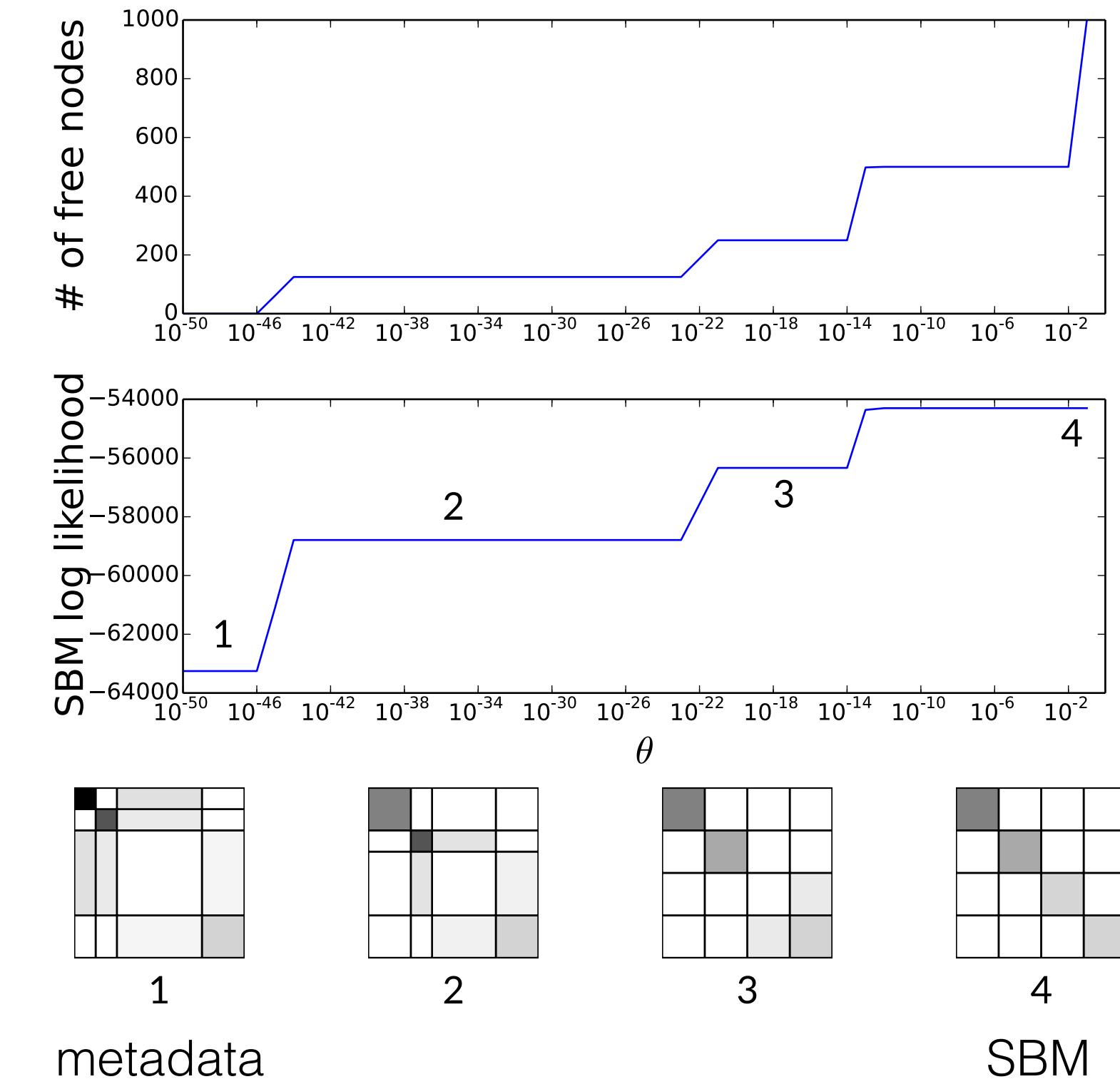
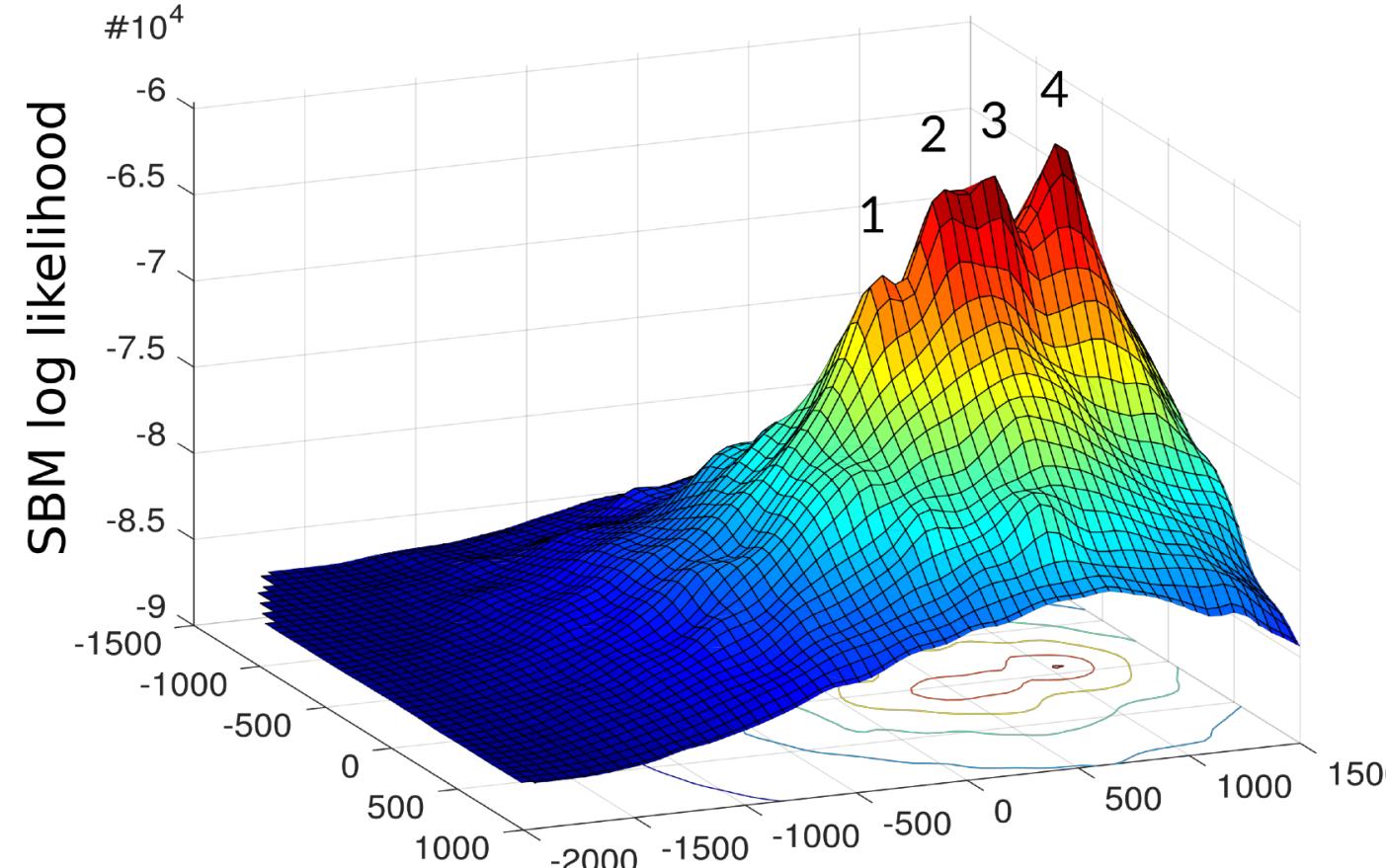


i. core-periphery

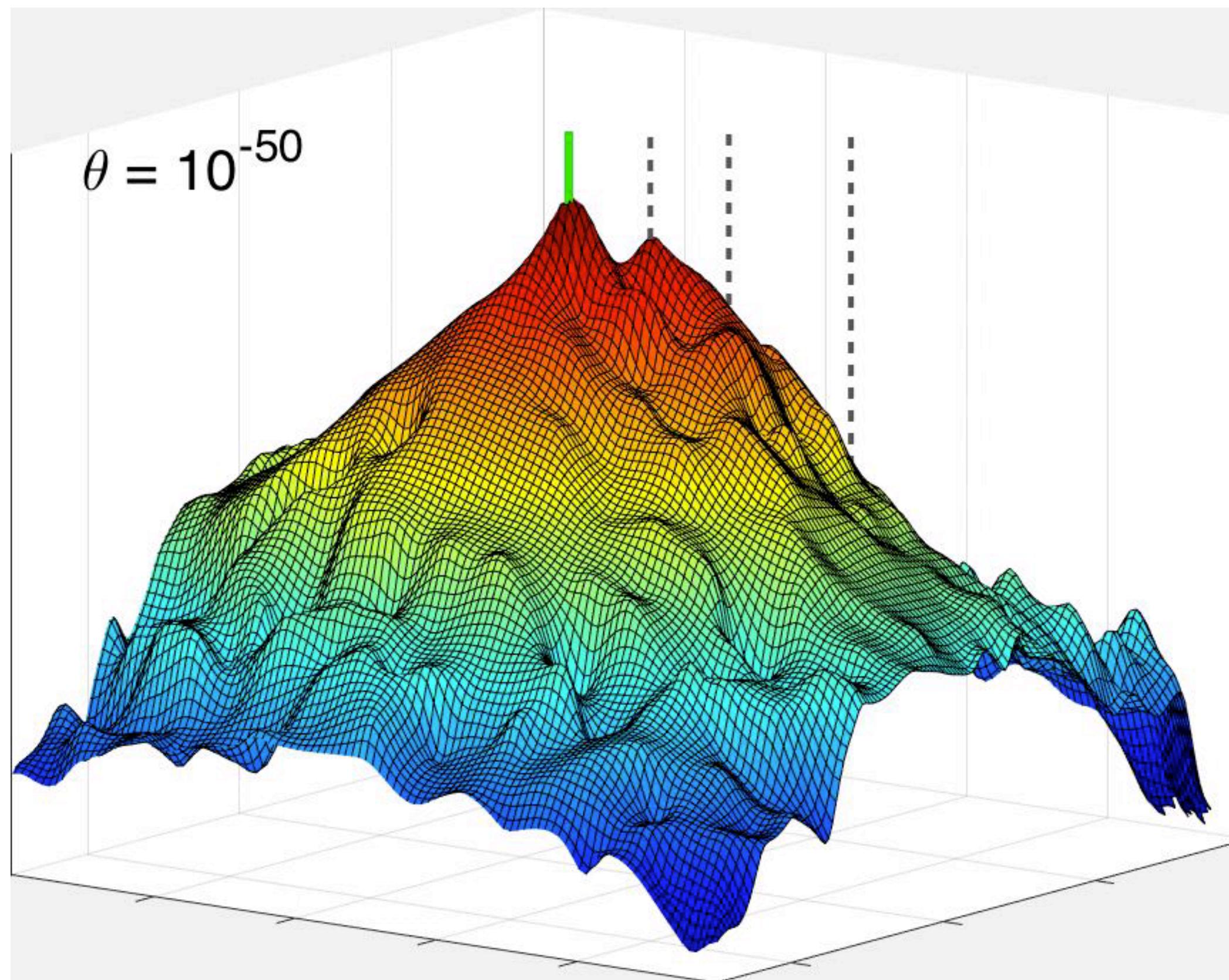


ii. assortative

The neoSBM identifies four interesting partitions



The prior parameter changes the likelihood surface



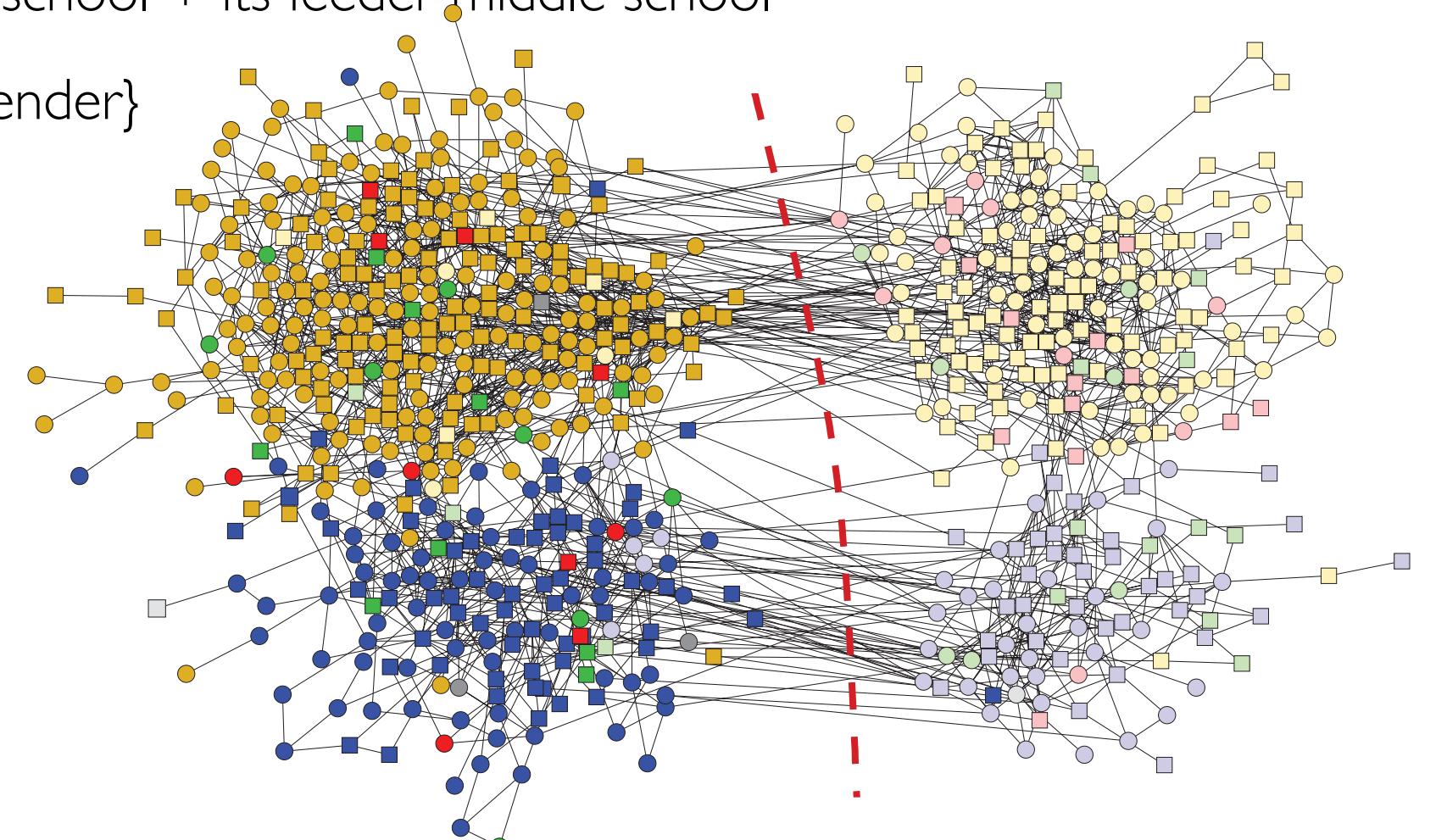
Metadata-aware SBM

high school social network



795 students from an American high school + its feeder middle school

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



Male



Female

White

Middle

High

Black

High

High

Hispanic

High

High

Other

High

High

Missing

High

High

Metadata-aware SBM

high school social network



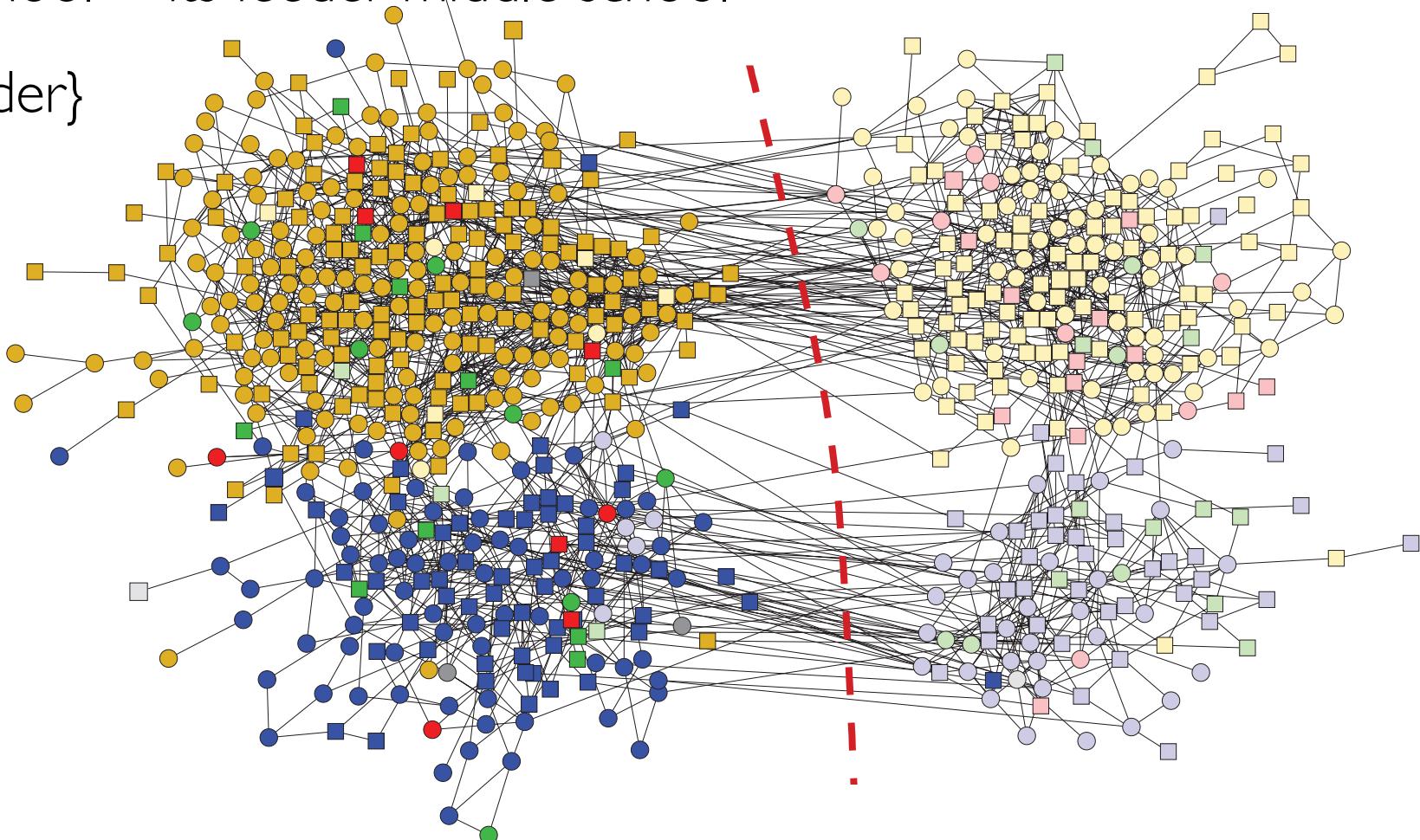
795 students from an American high school + its feeder middle school

- $\mathbf{x} = \{\text{grade 7-12, ethnicity, gender}\}$
- method finds a good partition between high-school and middle-school

$$\text{NMI} = 0.881$$

- without metadata:

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



	White	Black	Hispanic	Other	Missing	Male
Middle	○	○	○	○	○	○
High	●	●	●	●	●	●

Newman & Clauset, *Nat. Comms.* 7, 11863 (2016)

Add Health network data, designed by Udry, Bearman & Harris

<http://www.santafe.edu/~aarong/>

Metadata-aware SBM

high school social network



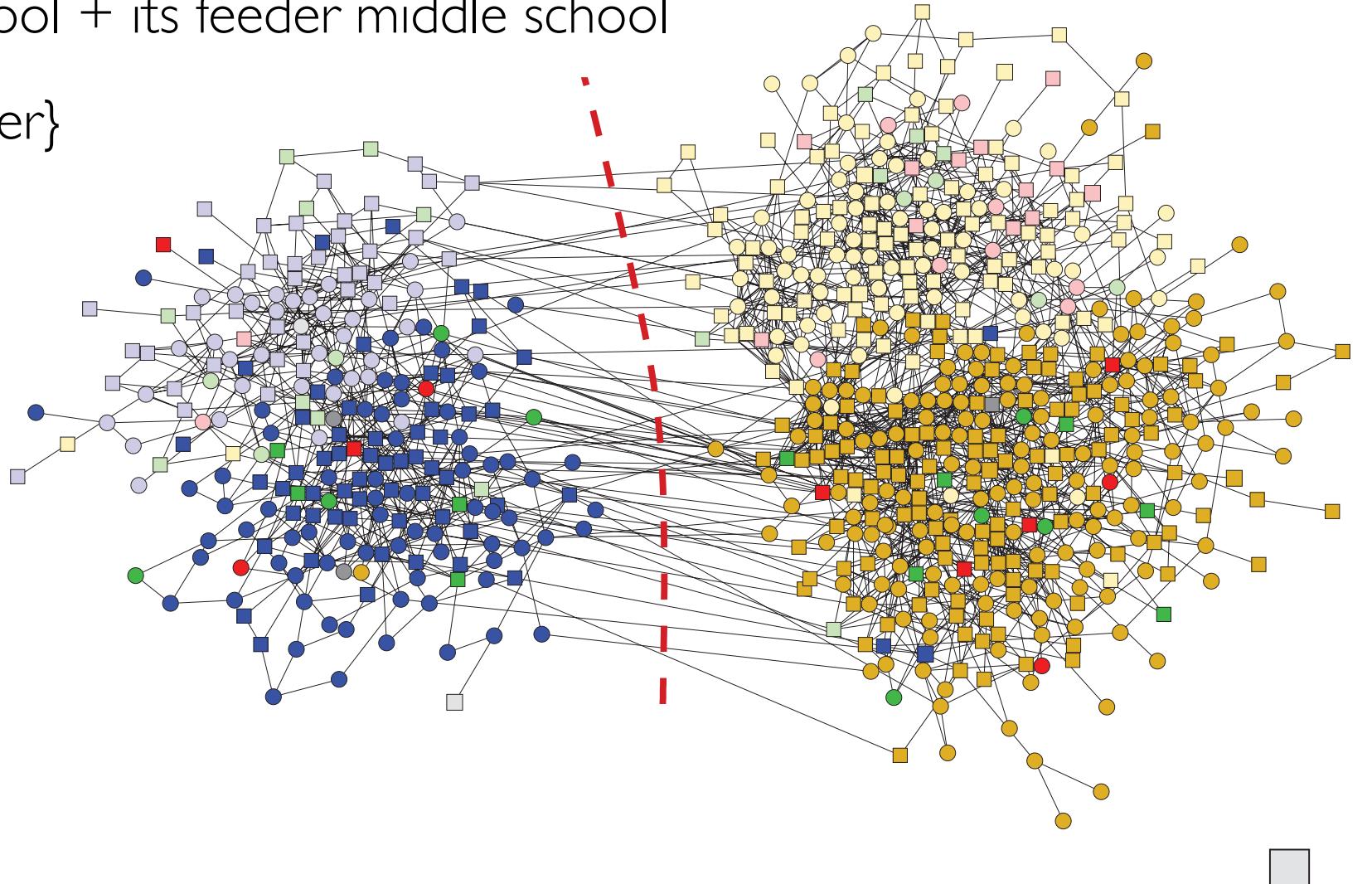
795 students from an American high school + its feeder middle school

- $\mathbf{x} = \{\text{grade 7-12, ethnicity, gender}\}$
- method finds a good partition between blacks and whites (with others scattered among)

NMI = 0.820

- without metadata:

NMI $\in [0.120, 0.239]$



	White	Black	Hispanic	Other	Missing	Male
Middle	○	○	○	○	○	○
High	○	○	○	○	○	○

Newman & Clauset, *Nat. Comms.* 7, 11863 (2016)

Add Health network data, designed by Udry, Bearman & Harris

<http://www.santafe.edu/~aaronc/>

Metadata-aware SBM

high school social network



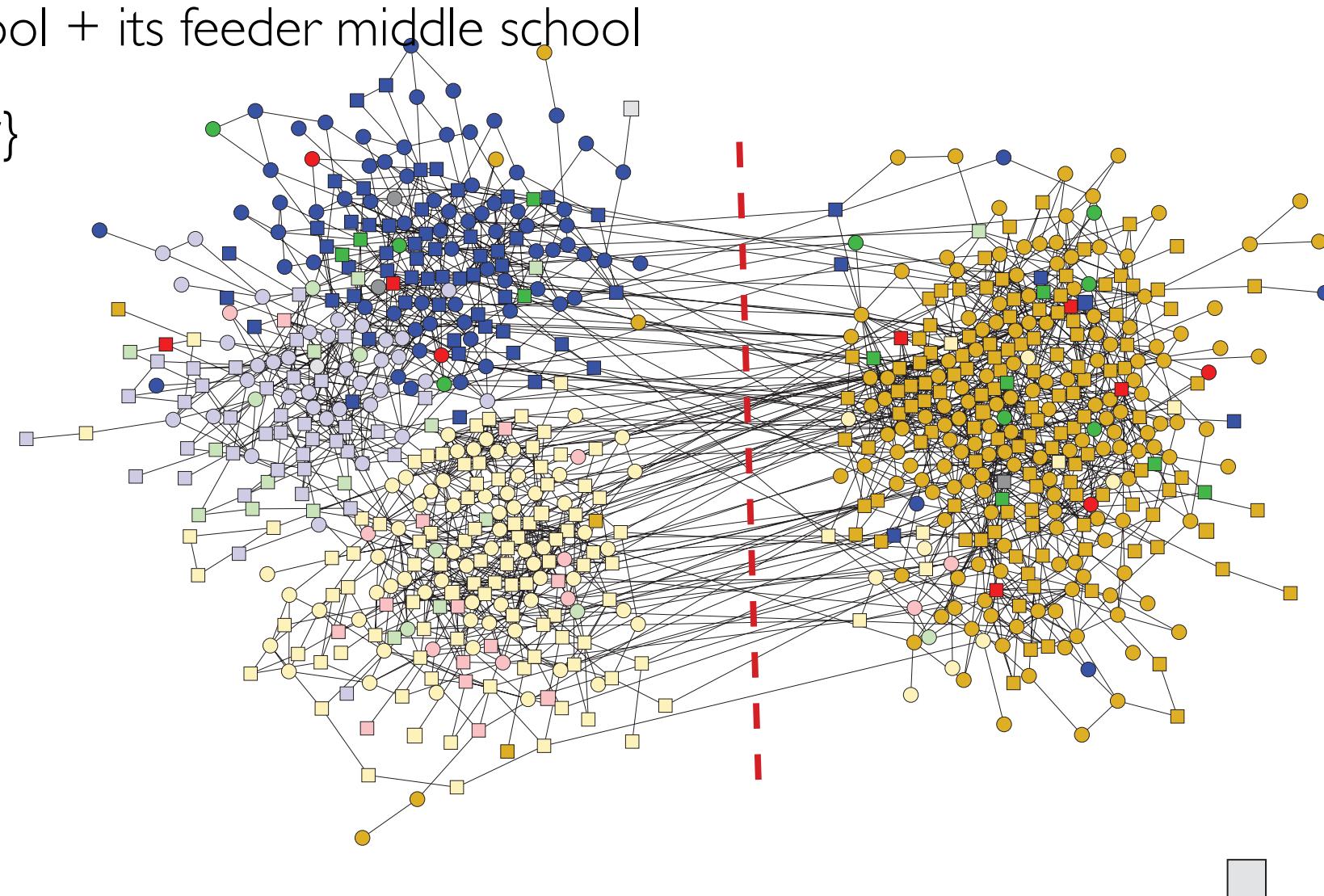
795 students from an American high school + its feeder middle school

- $\mathbf{x} = \{\text{grade 7-12, ethnicity, gender}\}$
- method finds no good partition between males/females.
instead, chooses a mixture of grade/ethnicity partitions

$$\text{NMI} = 0.003$$

- without metadata:

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



	White	Black	Hispanic	Other	Missing	Male	
Middle	○	○	○	○	○	○	
High	○	○	○	○	○	○	

Newman & Clauset, *Nat. Comms.* 7, 11863 (2016)

Add Health network data, designed by Udry, Bearman & Harris

<http://www.santafe.edu/~aaronc/>

Other things to know about 1: “The Louvain Method”

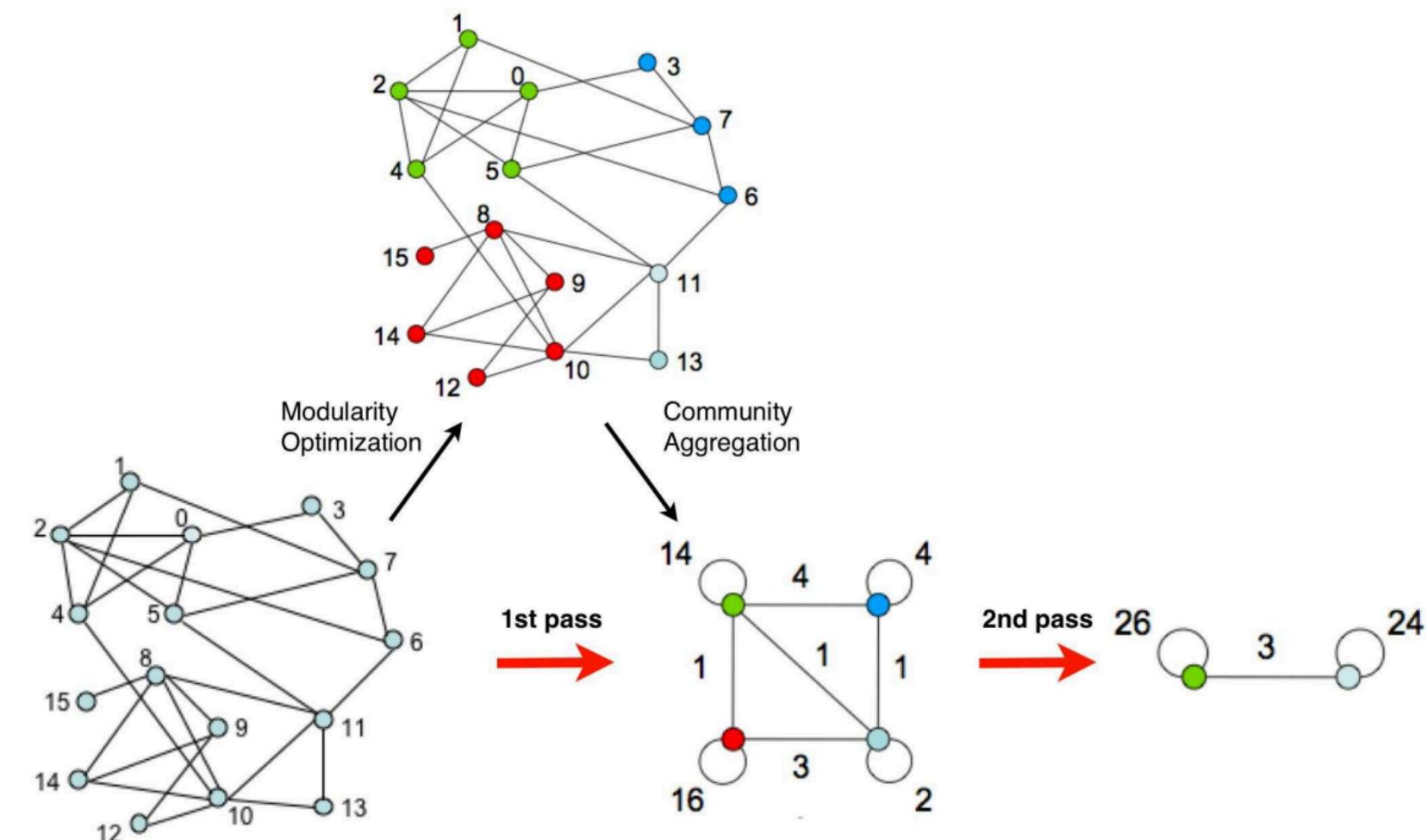
If your network is *really* big. (Millions of nodes, Billions of edges)

Take ClausetNewmanMoore’s approach for greedy Q maximization and find small groups. Run the code again on those groups... And again...

Advantage: fast! big! 

Disadvantage: inherits the assumptions of modularity.
(clustering vs modeling)

6K citations. People like it!

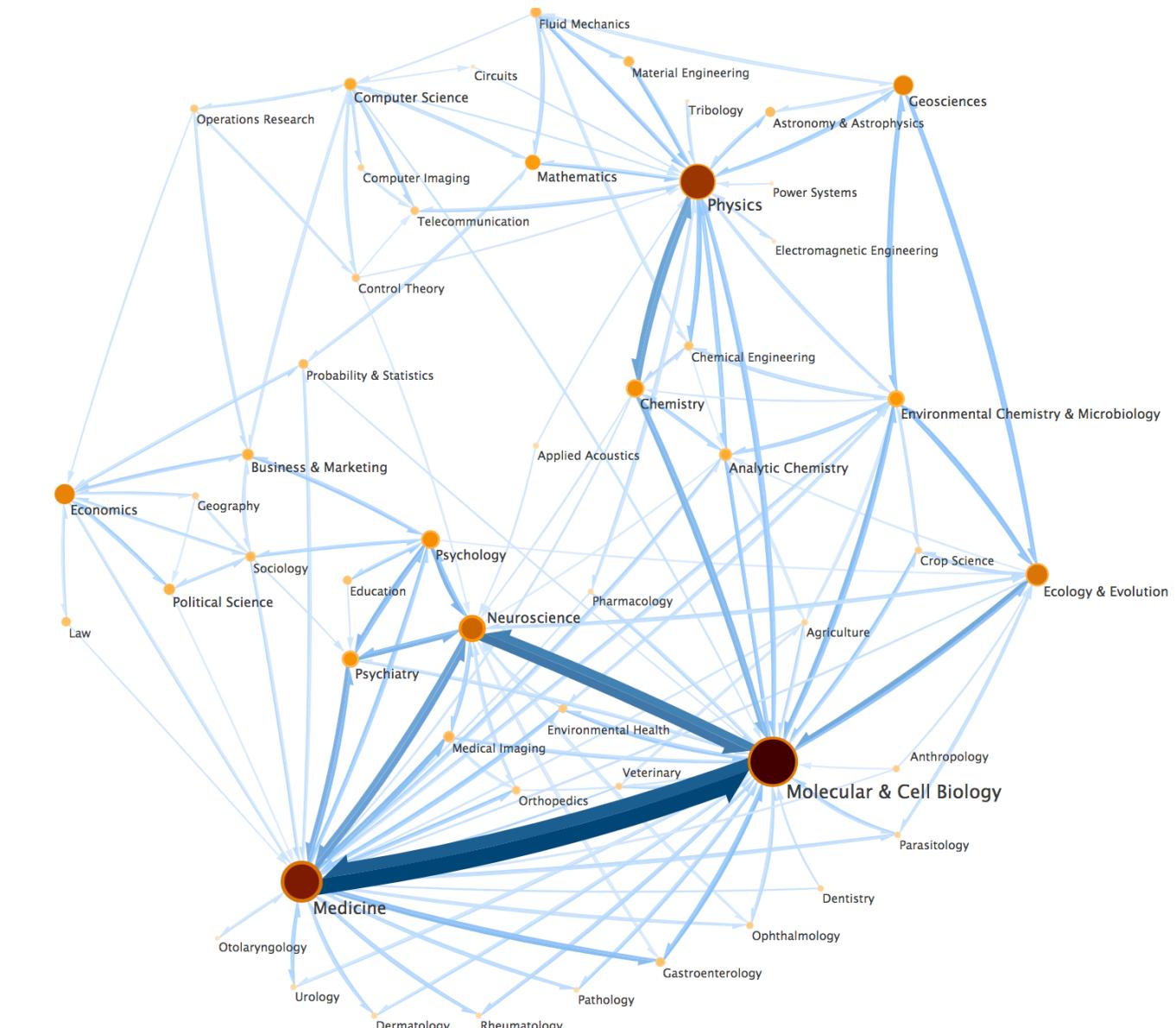


Other things to know about 2: InfoMap

Imagine a random walker on a network.

A description of her walk can be compressed if the network has regions in which the random walker tends to stay for a long time.

Minimizing the “map equation” over all possible network partitions is the same as finding the best codebook.



<http://www.mapecuation.org/apps/MapDemo.html> 

<http://www.mapequation.org/code.html>

Outlook for community detection

Simply put, we have amazingly powerful tools that did not exist 15 years ago.

Many are principled, statistically rigorous, and we learn more all the time. Those that aren't statistically rigorous are really, really fast.

There is no multiple regression for networks.

“Controlling for C, how important is X in predicting Y?”

Tradeoffs between general and bespoke methods are still being explored.

Outside of SBM, Modularity, Louvain, Infomap, it's a wild west.

Methodologists are keen to be challenged by new problem types.

New scientific questions inspire new methods.

Revisit: uses for models of large-scale structure

Treat the network like a system:

Extrapolation. Make predictions for as-yet unseen nodes (in “space” or time).

Interpolation. Identify missing links.

Generalization. Nodes of this type are like others of the same type.

Treat the network like an artifact:

Mechanisms. How did this network arise? What rules governed its assembly?

Explanations. Coarse-graining or compression.

Treat the network like a means to an end; an intermediate data structure:

Useful division. Need groups so that we can assign treatments in an A/B test.

Simplification. Downstream regression model needs ranks or groups.

intuition: compare this list with the list you would write for regression