Physics-Inspired Algorithms and Phase Transitions in Community Detection

Cristopher Moore, Santa Fe Institute

joint work with Lenka Zdeborová, Florent Krzakala, Aurelien Decelle, Aaron Clauset, Amir Ghasemian, Leto Peel, Pan Zhang, Xiaoran Yan, Yaojia Zhu, and Lise Getoor








Omidyar Fellowship Postdocs (Nov. 1)
REU Summer for Undergrads
Complex Systems Summer School

How can we find patterns in data?

How do we know if the patterns we see are really there?

Statistical inference $\Leftrightarrow$ statistical physics

What is structure?

Structure is that which...

makes data different from noise

helps us compress the data

helps us generalize from data we've seen from data we haven't seen

helps us coarse-grain the dynamics, reducing the number of variables

Statistical inference

Suppose we have a network (a graph with nodes and links)

Imagine that it is created by a *generative model*, and fit the parameters of this model to the data

Can gracefully incorporate partial information: e.g. if

- attributes of some nodes are known
- some links are known, others not observed yet (e.g. food webs)
- some links are false positives (e.g. gene regulatory networks, protein interactions)

Use the model to generalize from what we know to what we don't

The stochastic block model

nodes have discrete labels: k “groups” or types of nodes

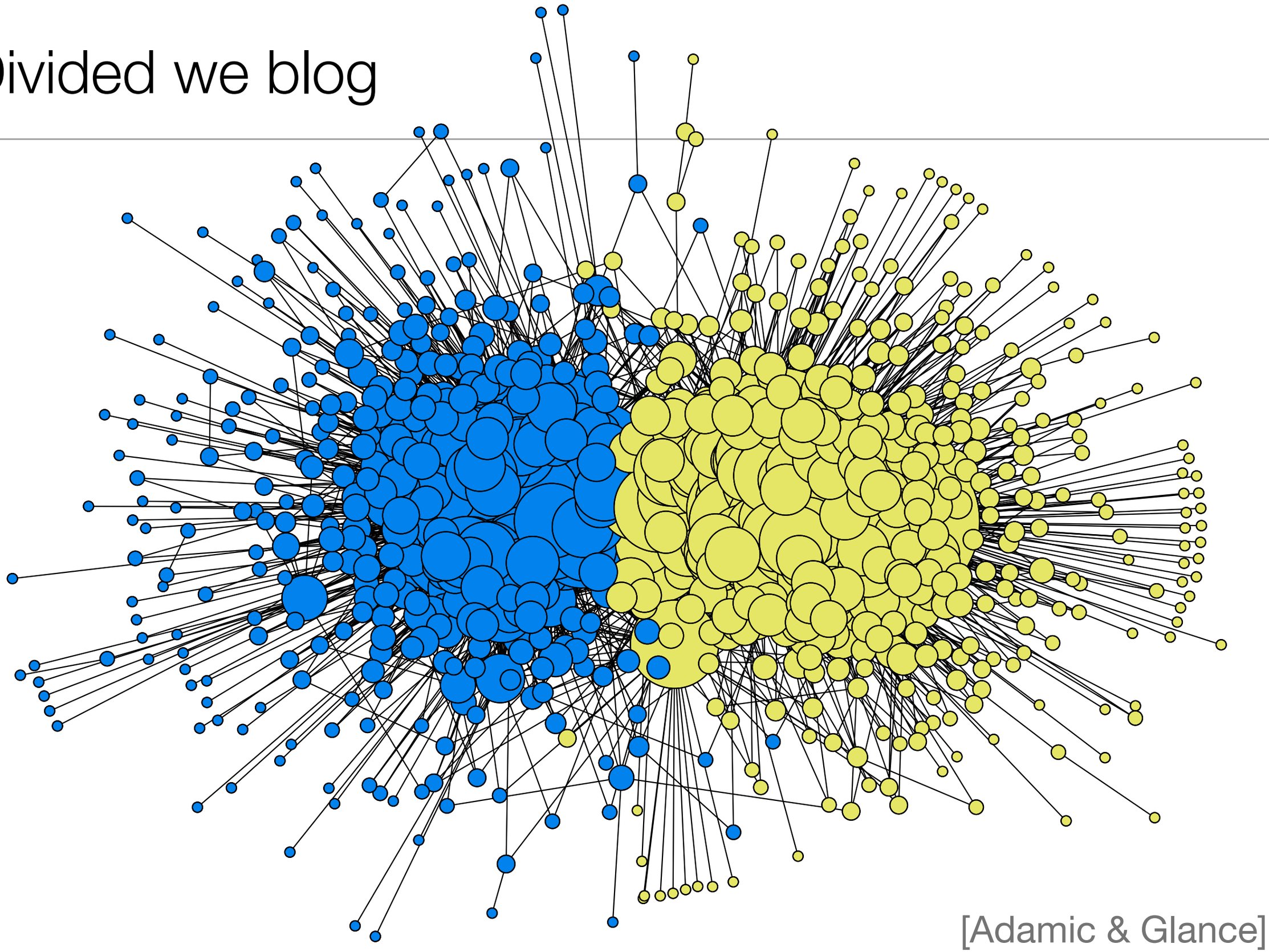
$k \times k$ matrix p of connection probabilities

if i is type r and j is type s, there is a link $i \to j$ with probability p_{rs}

p is not necessarily symmetric, and we don’t assume that $p_{rr} > p_{rs}$

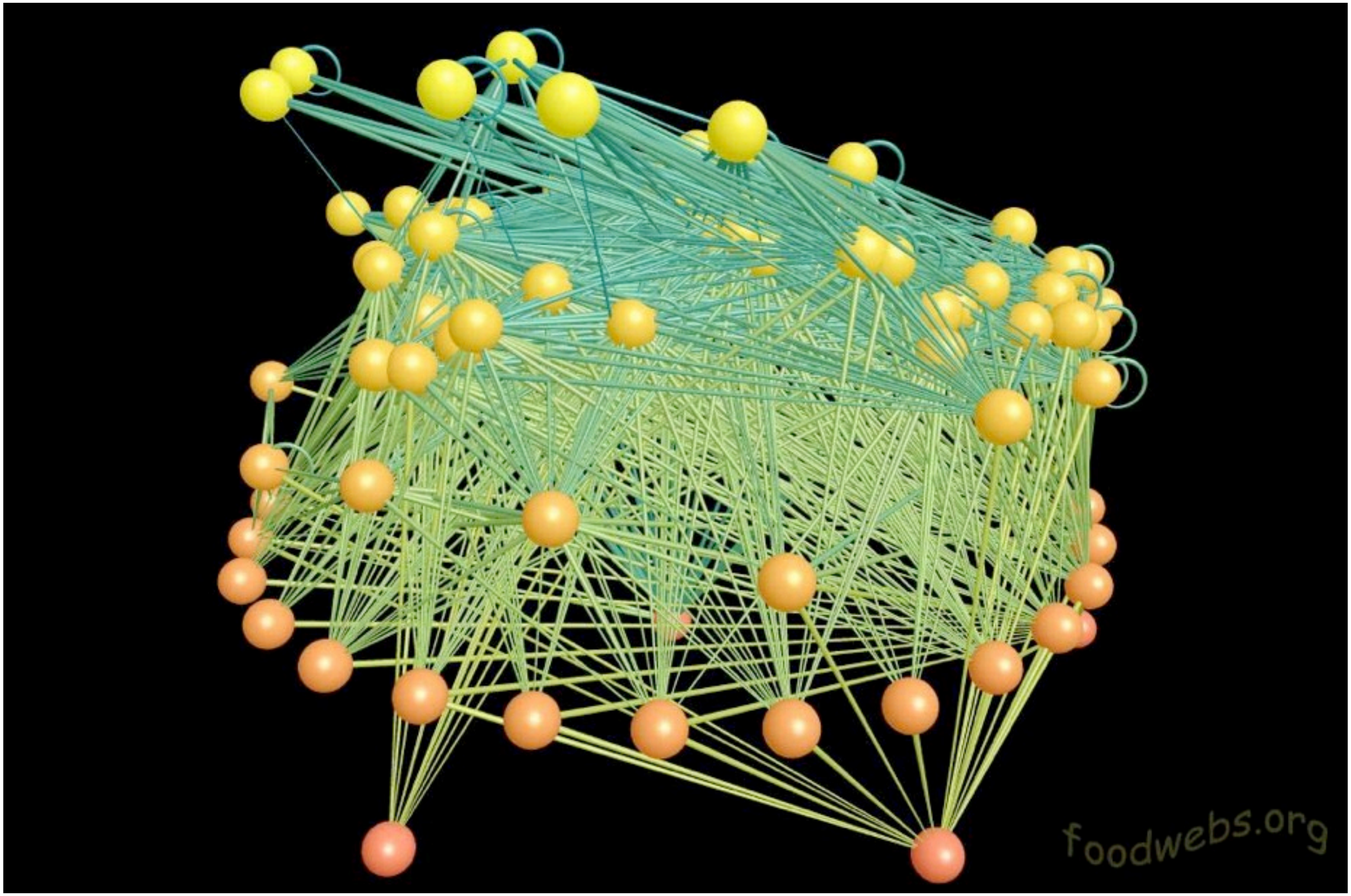
given the graph G, find the labels!

Divided we blog

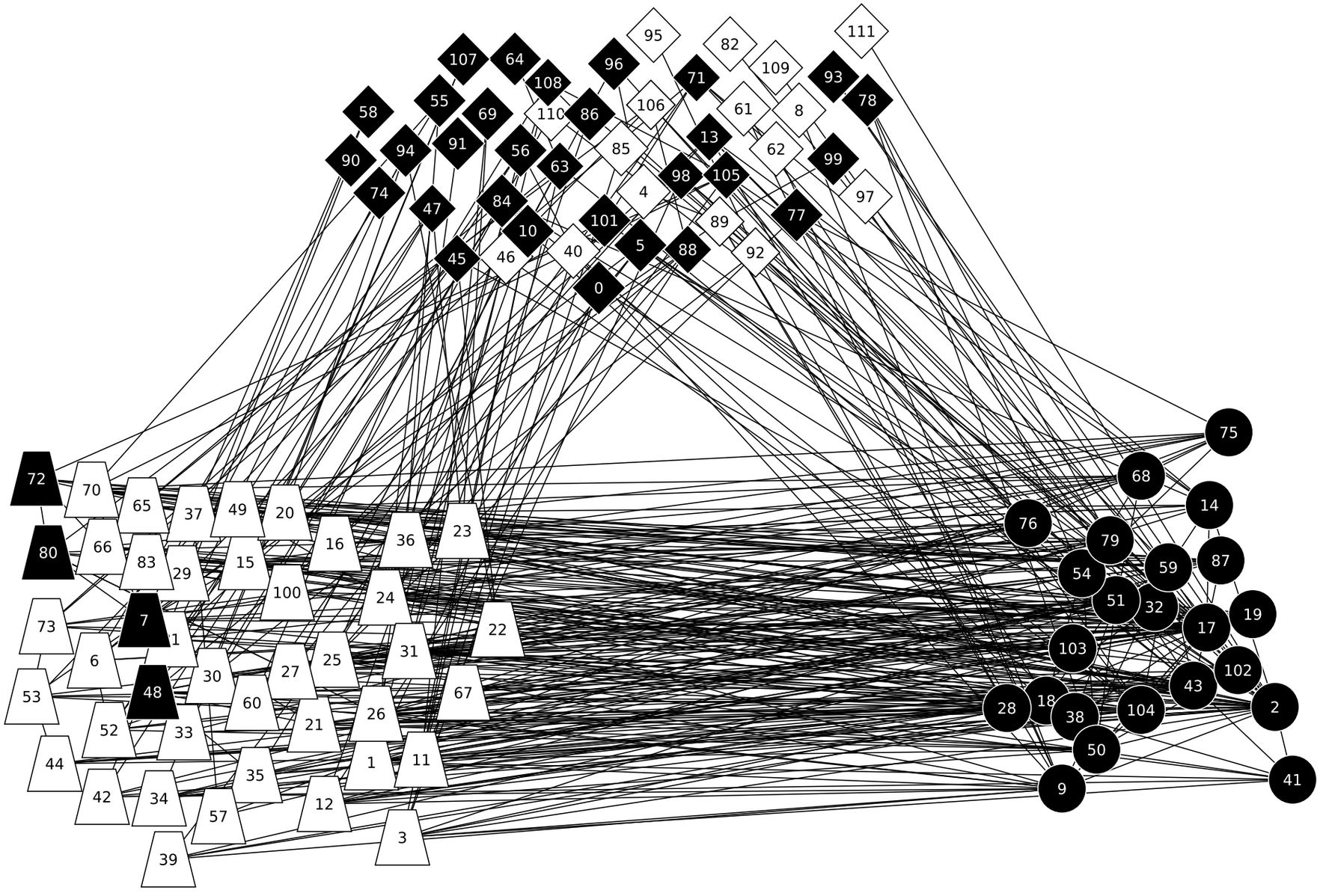


[Adamic & Glance]

Who eats whom



I record that I was born on a Friday



Some cases of interest

$$p = \frac{1}{n}\begin{pmatrix} c_{\text{in}} & c_{\text{out}} \\ c_{\text{out}} & c_{\text{in}} \end{pmatrix}$$

planted partitioning:
$c_{\text{in}} > c_{\text{out}}$ assortative
$c_{\text{in}} < c_{\text{out}}$ disassortative

$$p = \frac{1}{n}\begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

core-periphery:
$a > b > c$

$$p = \frac{c}{n}\frac{k}{k+1}\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

planted graph coloring:
k colors,
average degree c

Inferring the block model scalably

in the worst case, fitting the block model to a graph is NP-hard

in practice, there are now several scalable methods:

- belief propagation [Decelle, Krzakala, Moore, Zdeborová]
- pseudolikelihood [Amini, Chen, Bickel, Levina]
- stochastic optimization using subsampling [Gopalan, Blei, et al.]
- exact EM algorithms [Ball, Karrer, Newman]
- spectral methods

belief propagation (BP) lets us build analogies with statistical physics,
gives natural measures of statistical significance,
and reveals phase transitions in the detectability of community structure

Likelihood and energy

the probability of *G* given the types *t* is a product over edges and non-edges:

$$P(G \mid t) = \prod_{(i,j)\in E} p_{t_i,t_j} \prod_{(i,j)\notin E} (1 - p_{t_i,t_j})$$

using $P \sim e^{-\beta E}$ where $\beta = 1/T$ (Boltzmann) and E is the energy,

$$E(t) = -\log P(G \mid t) = -\sum_{(i,j)\in E} \log p_{t_i,t_j} - \sum_{(i,j)\notin E} \log(1 - p_{t_i,t_j})$$

like Ising model, but with interactions on both edges and non-edges

in the sparse case $p = O(1/n)$, interactions on non-edges are weak

Analogies with statistical physics: a glossary

probability of *G* given *t*	$P(G \mid t)$	$\mathrm{e}^{-\beta E(t)}$	(β=1 for now)
	$-\log P(G \mid t)$	$E(t)$	energy
most likely labeling (MAP)	$\operatorname{argmax}_t P(G \mid t)$	$\operatorname{argmin}_t E(t)$	ground state
total probability of *G* in this model	$\sum_{t \in \{1,\ldots,k\}^n} P(G,t)$	Z	partition function
Gibbs distribution	$P(t \mid G) = \frac{P(G \mid t)}{\sum_{t'} P(G \mid t')}$	$P(t) = \frac{\mathrm{e}^{-E(t)}}{Z}$	Gibbs distribution
	$-\log \sum_t P(G \mid t)$	$F = -\log Z$	free energy

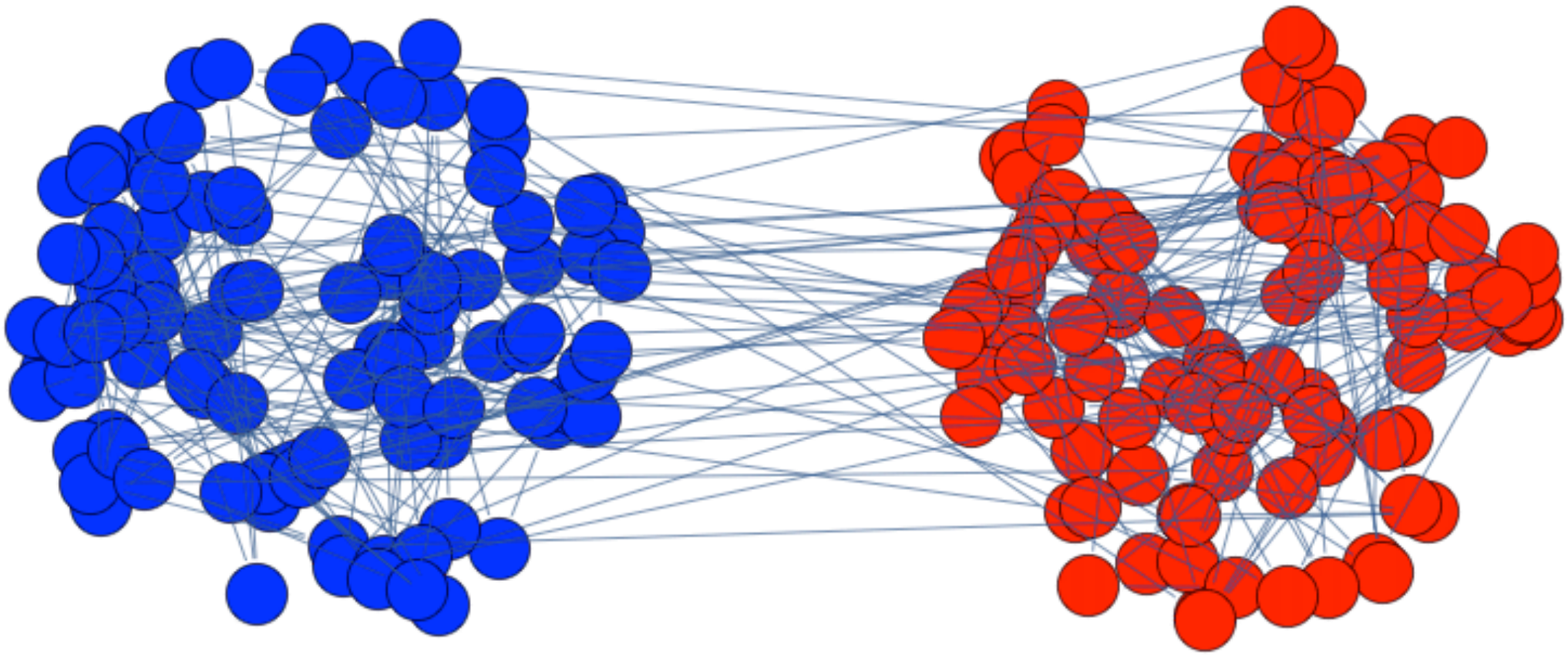
Ground states and illusions

the most likely labeling, or MAP estimate, is the *ground state*: it maximizes $P(G|t)$

but even random 3-regular graphs have labelings with only 11% of the edges crossing the cut [Zdeborová & Boettcher]

many labelings, about as good as each other, with nothing in common!

this is a sign there aren't actually communities at all...

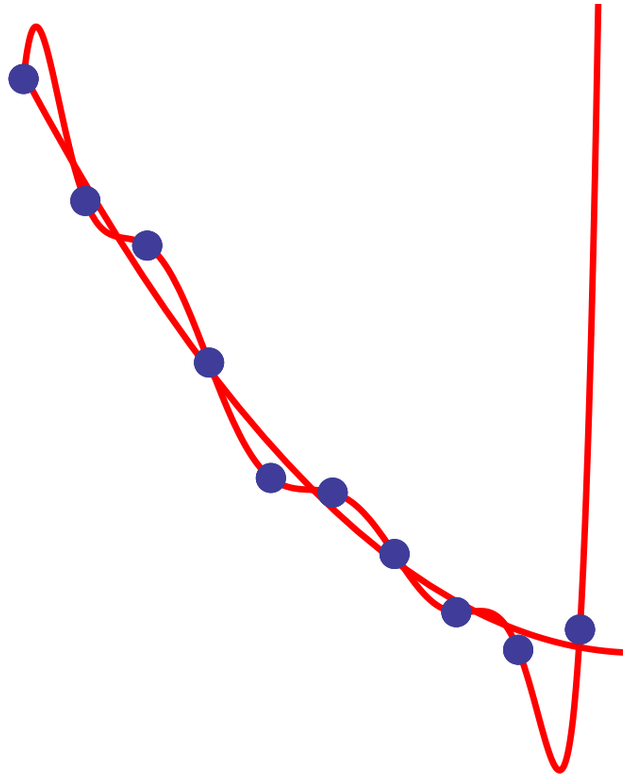


Statistical significance vs. overfitting

we don’t just want the best fit!

random graphs have illusory communities, that only exist because of noise

sometimes the patterns we find aren’t really there:



we want to understand the coin, not the coin flips

What's the best labeling, redux

for each node, compute its *marginal distribution*, the probability that it belongs to each group

assign each node to its most-likely label

achieves a higher "overlap" with the true labeling than the ground state: maximizes the expected fraction of nodes labeled correctly

marginals represent clusters of many solutions that agree on most nodes...

the consensus of many likely solutions is better than the most-likely one

Model selection and free energy

let θ denote the parameters of the model, e.g. factions vs. core-periphery

best model: maximize *total* probability of G, summed over all possible labelings:

$$P(G \mid \theta) = \sum_{t \in \{1,\ldots,k\}^n} P(G, t \mid \theta)$$

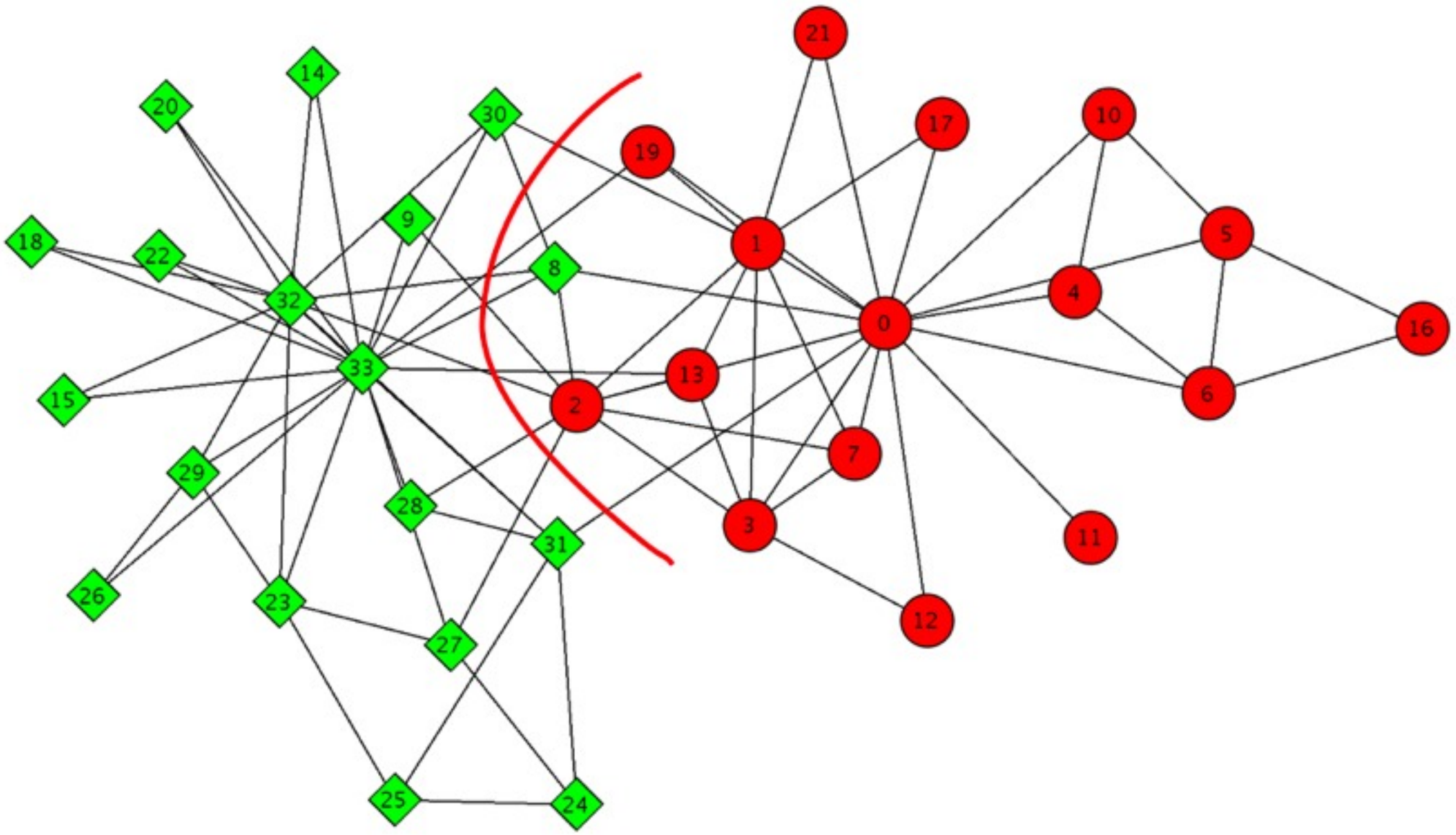
this is the partition function Z and $F = -\log P(G|\theta)$ is a free energy

thermodynamically, $F = E - TS$

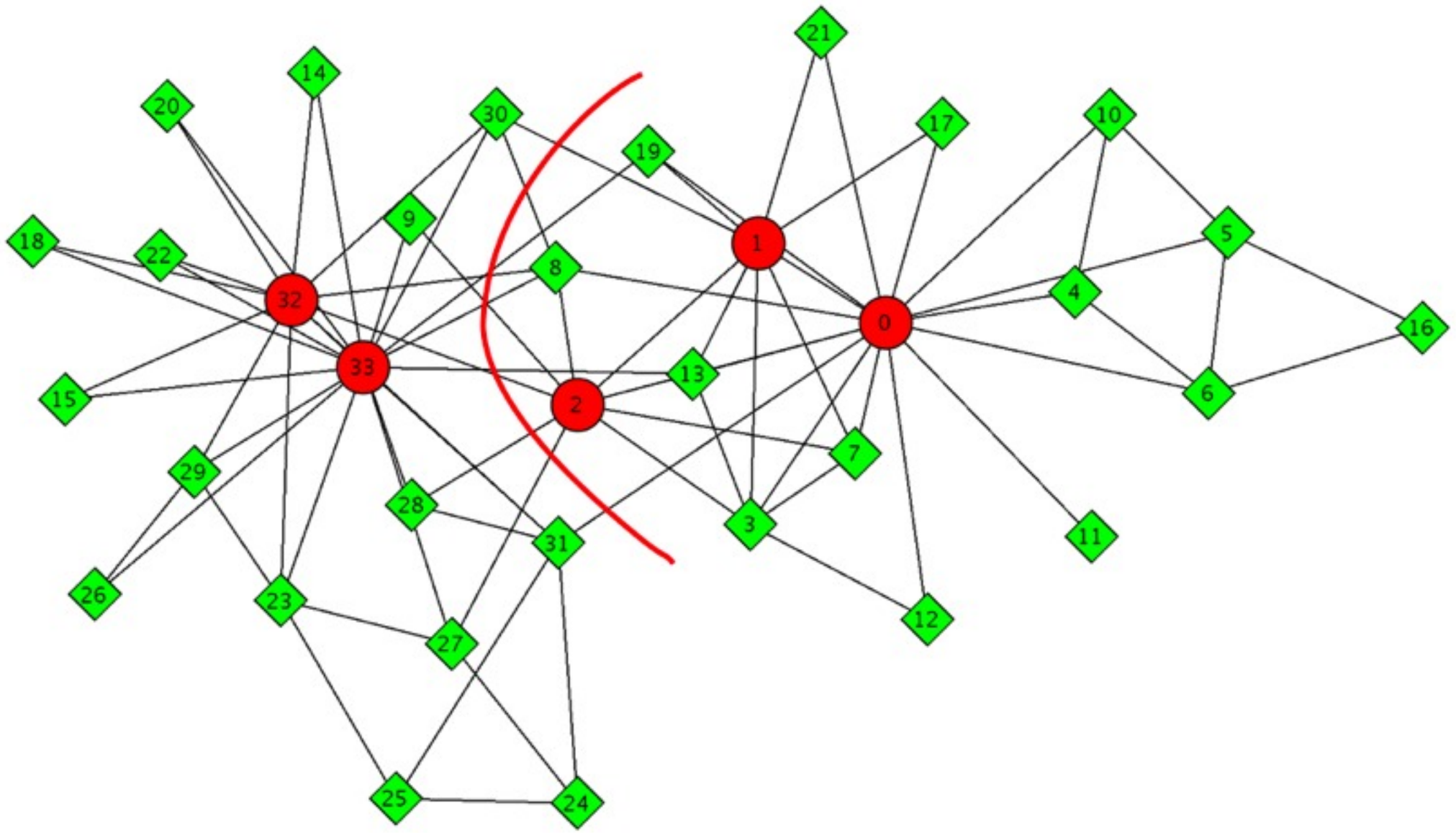
minimizing F = low energy (high probability) + high entropy (many good solutions)

a good model fits the data robustly, with many values of the hidden variables

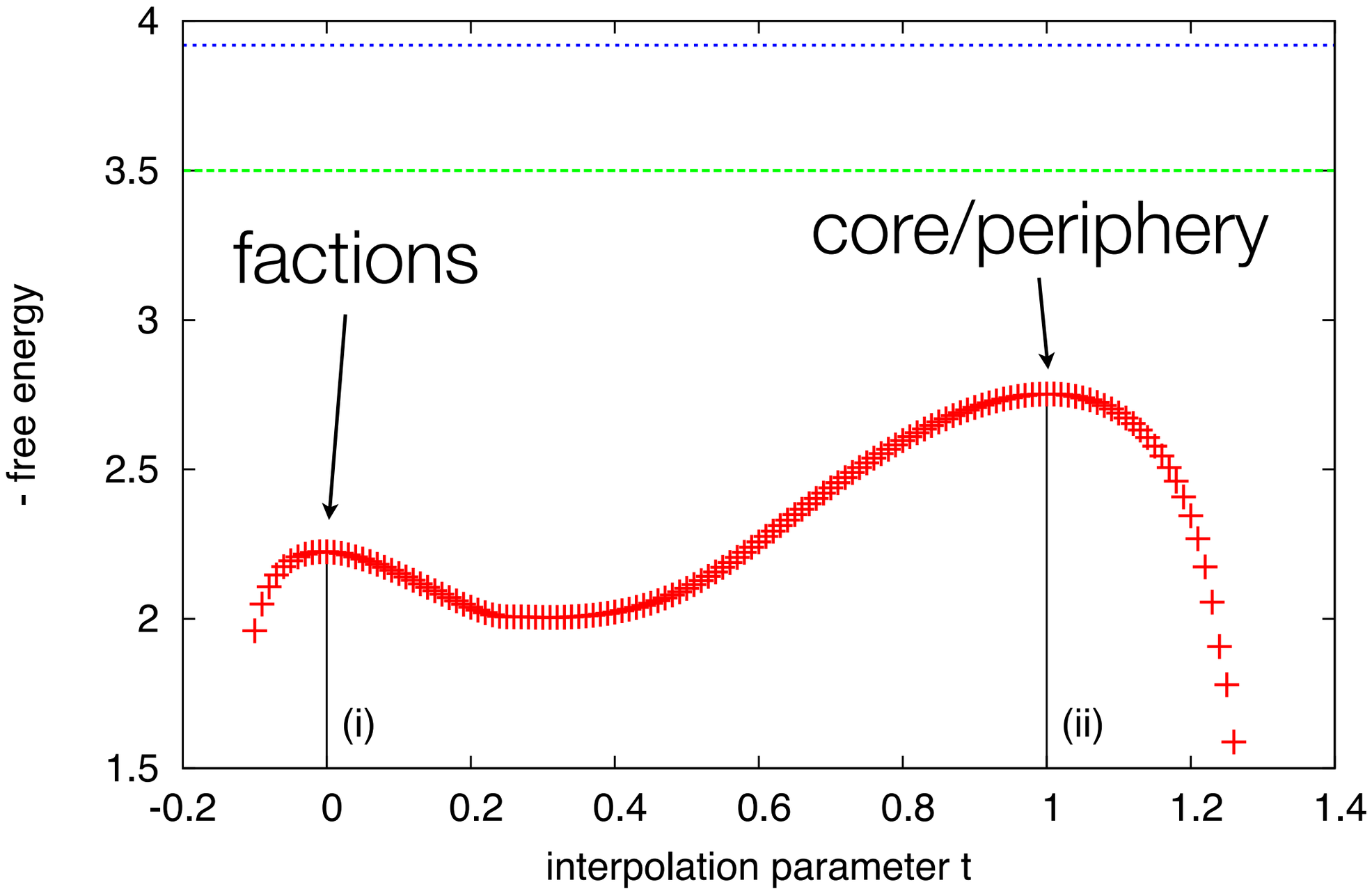
Zachary's Karate Club: Two factions



Zachary's Karate Club: Core-periphery



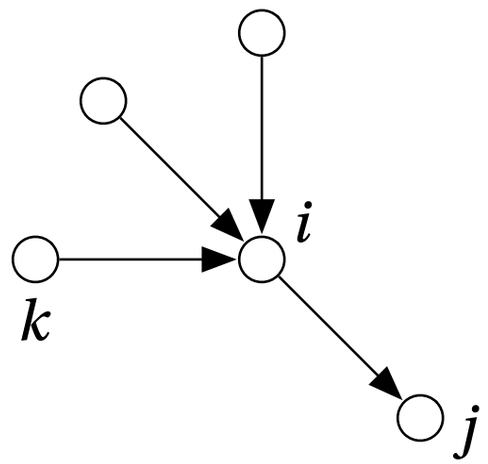
Two local optima in free energy



But how can we compute marginals and free energies?

Monte Carlo is too slow!

Belief propagation (a.k.a. the cavity method)



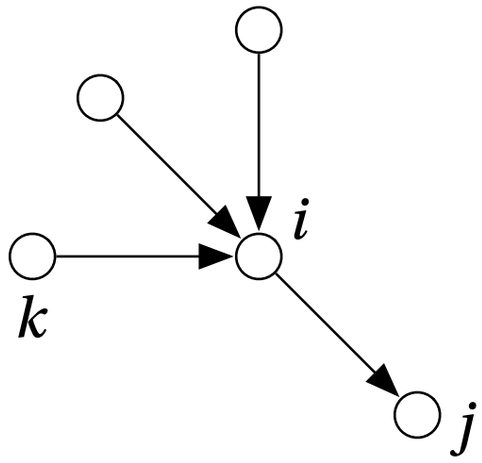
each node *i* sends a “message” to each of its neighbors *j*, giving *i*’s marginal distribution based on its other neighbors *k*

avoids an “echo chamber” between pairs of nodes

update until we reach a fixed point (how many iterations? does it converge?)

fixed point returns estimated marginals and the Bethe free energy

Updating the beliefs



WARNING: EXACT ONLY ON TREES

conditional independence

$$\mu_s^{i\to j} = \frac{1}{Z^{i\to j}} q_s \prod_{\substack{k\neq j \\ (i,k)\in E}} \sum_r \mu_r^{k\to i} p_{rs} \times \prod_{\substack{k\neq j \\ (i,k)\notin E}} \sum_r \mu_r^{k\to i}(1-p_{rs})$$

a complete graph of messages: takes O(n^2) time to update. Not scalable!

sparse case: can simplify by assuming that $\mu_r^{k\to i} = \mu_r^k$ for all non-neighbors *i*

then update takes O(*n*+*m*) time: scalable!

Approximating the free energy: variational trick

$$\log P(G\,|\,\theta) = \log \sum_t P(G\,|\,t,\theta)$$

$$= \log \mathop{\mathbb{E}}_{t\sim Q} \frac{P(G\,|\,t,\theta)}{Q(t)}$$

$$\geq \mathop{\mathbb{E}}_{t\sim Q} \log \frac{P(G\,|\,t,\theta)}{Q(t)}$$

$$= \mathop{\mathbb{E}}_{t\sim Q} \log P(G\,|\,t,\theta) + S(Q)$$

$$-\beta F = \log Z = \log \sum_t \mathrm{e}^{-\beta E(t)}$$

$$= \log \mathop{\mathbb{E}}_{t\sim Q} \frac{\mathrm{e}^{-\beta E(t)}}{Q(t)}$$

$$\geq -\beta \mathop{\mathbb{E}}_{t\sim Q} E(t) + S(Q)$$

$$= -\beta \langle E \rangle + S(Q)$$

where $S(Q) = -\sum_t Q(t) \log Q(t)$ or $F = E - TS$

holds with equality when Q(*t*) is the Gibbs distribution

variational approach: find the best Q(*t*) (with the lowest free energy) in a family of distributions with poly(*n*) parameters

each family gives a lower bound on *P*(G|θ), upper bound on free energy

The Bethe free energy

average energy depends just on 1- and 2-point marginals,

$$\langle E \rangle = \sum_{(i,j)\in E} \sum_{r,s=1}^{k} \mu_{rs}^{ij} \log p_{rs} + \sum_{(i,j)\notin E} \sum_{r,s=1}^{k} \mu_{rs}^{ij} \log(1-p_{rs})$$

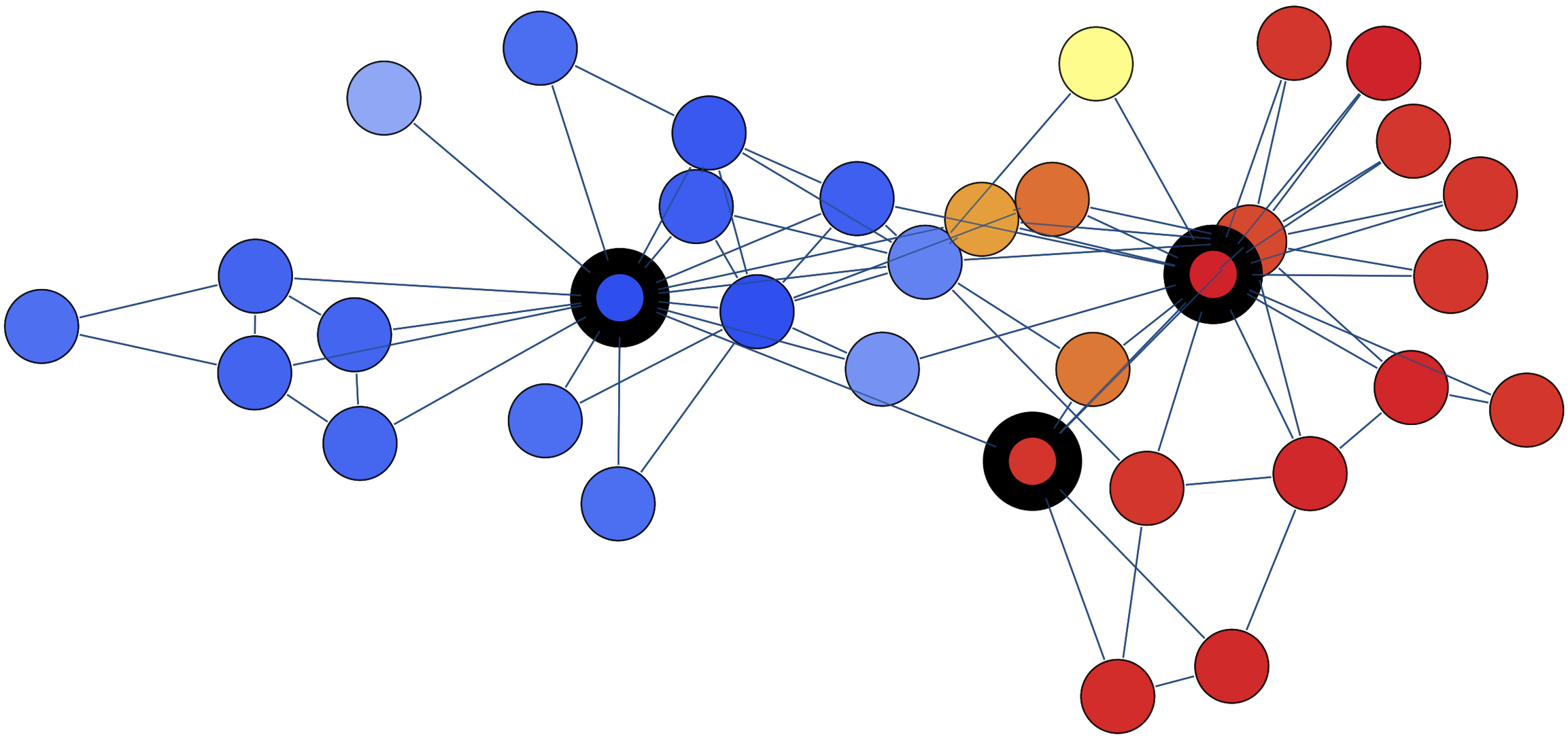
but the entropy is more complicated... so approximate the Gibbs distribution with a form that depends just on 1- and 2-point marginals:

$$Q(\{t_i\}) = \frac{\prod_{(i,j)\in E} \mu_{t_i,t_j}^{ij}}{\prod_i \left(\mu_{t_i}^{i}\right)^{d_i-1}}$$

exact for trees, but pretty good even for graphs with loops

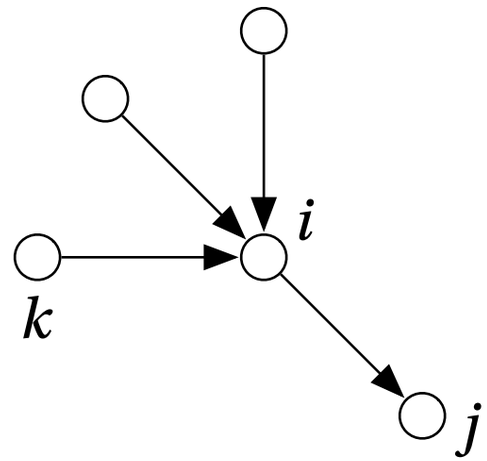
BP fixed points are local optima of the Bethe free energy [Yedidia]

Active learning:
update the model as we learn more



[Moore, Yan, Zhu, Rouquier, Lane, *KDD* 2011]

The double life of Belief Propagation



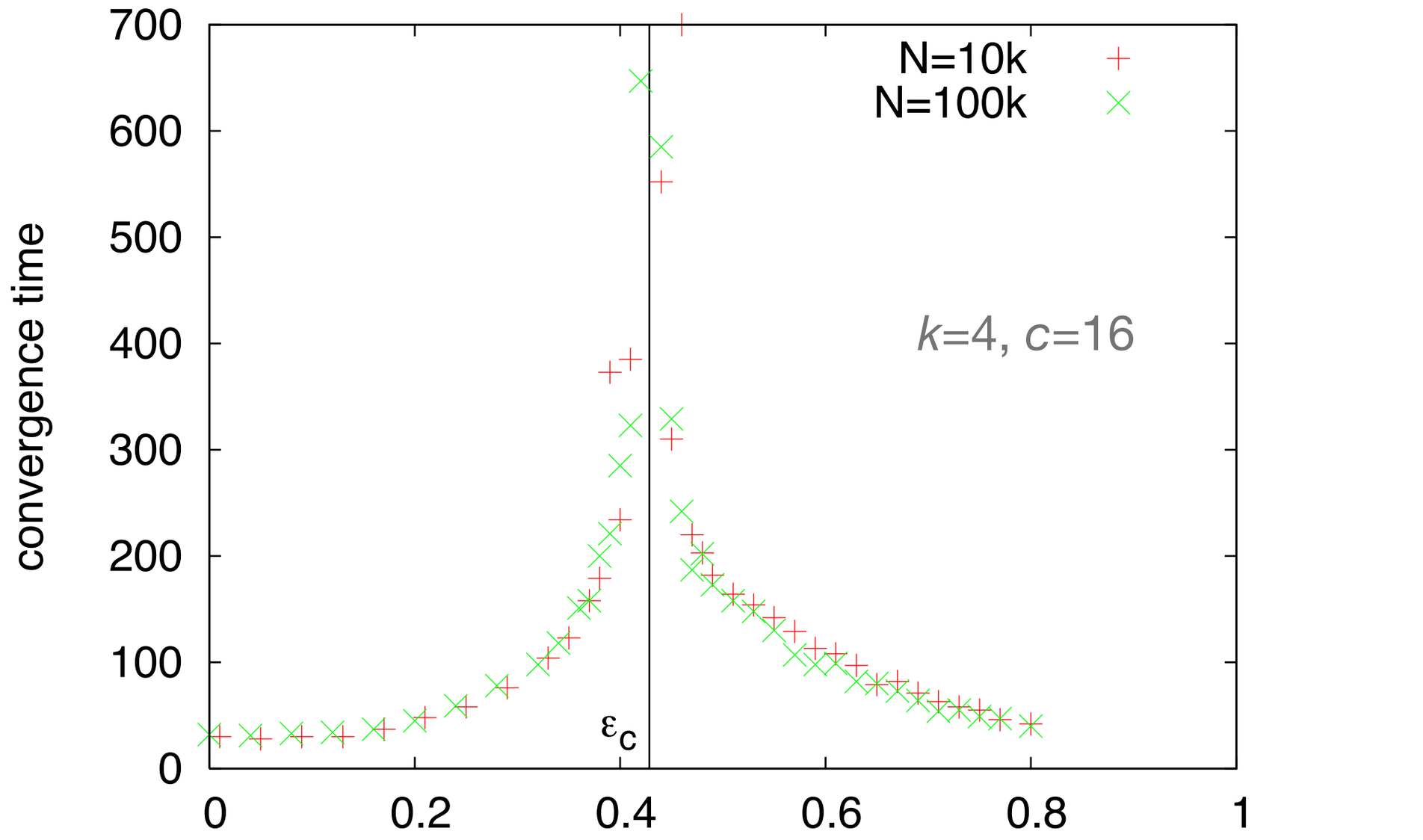
BP is a fast algorithm we can run on real networks...

but it's also a framework for analytic calculations on ensembles of graphs (e.g. the stochastic block model) in the large-*n* limit

analyze fixed points of the messages, their basins of attraction, their stability

BP convergence: nearly size-independent, but with critical slowing down at a phase transition

$p_{ss}=c_{in}/n$, $p_{rs}=c_{out}/n$



A phase transition: detectable to undetectable communities

when c_{out}/c_{in} is small enough, BP can find the communities

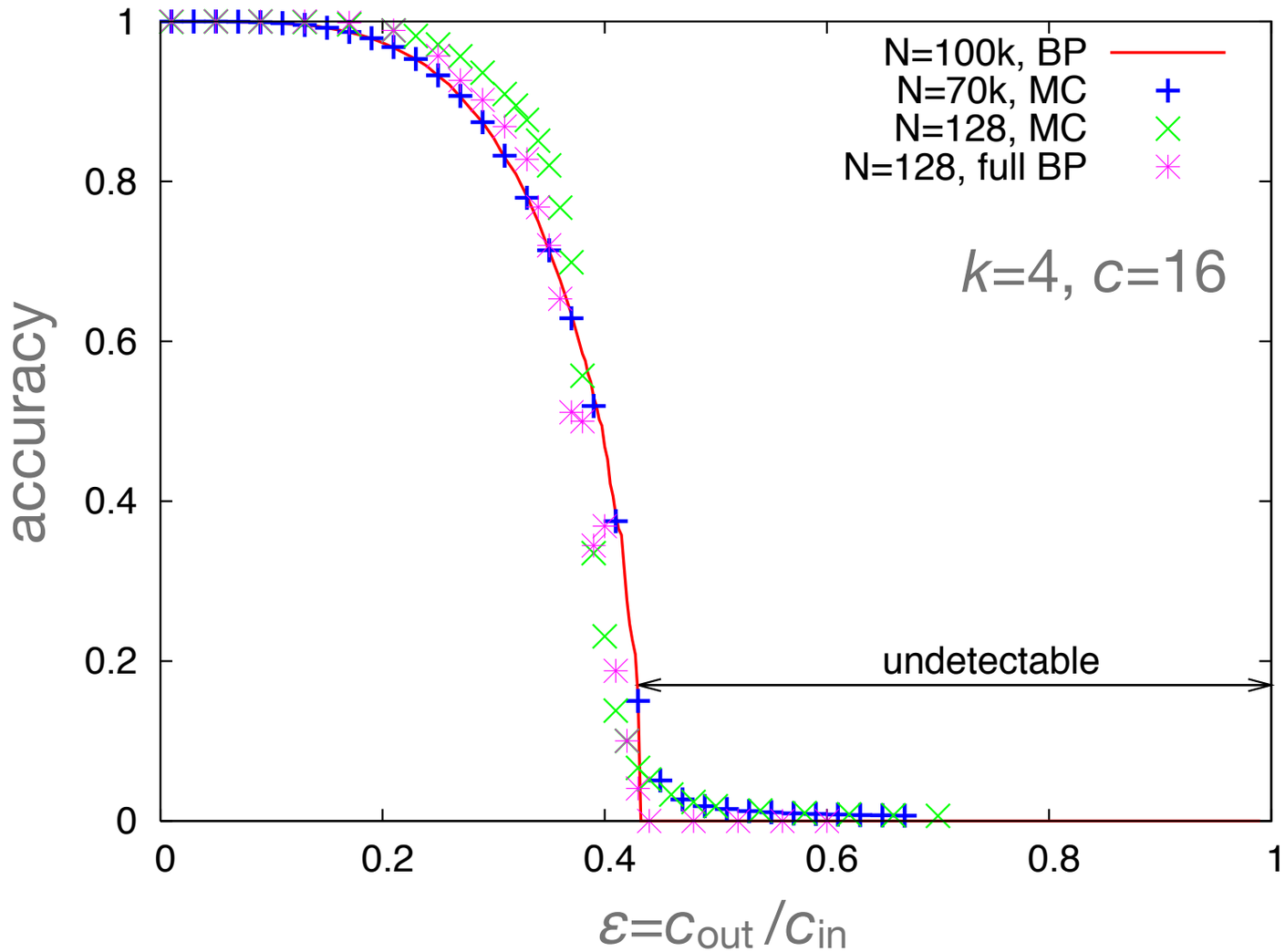
there is a regime where it can't, and no algorithm can!

for 2 groups, the threshold is at

$$|c_{\text{in}} - c_{\text{out}}| = 2\sqrt{c}$$

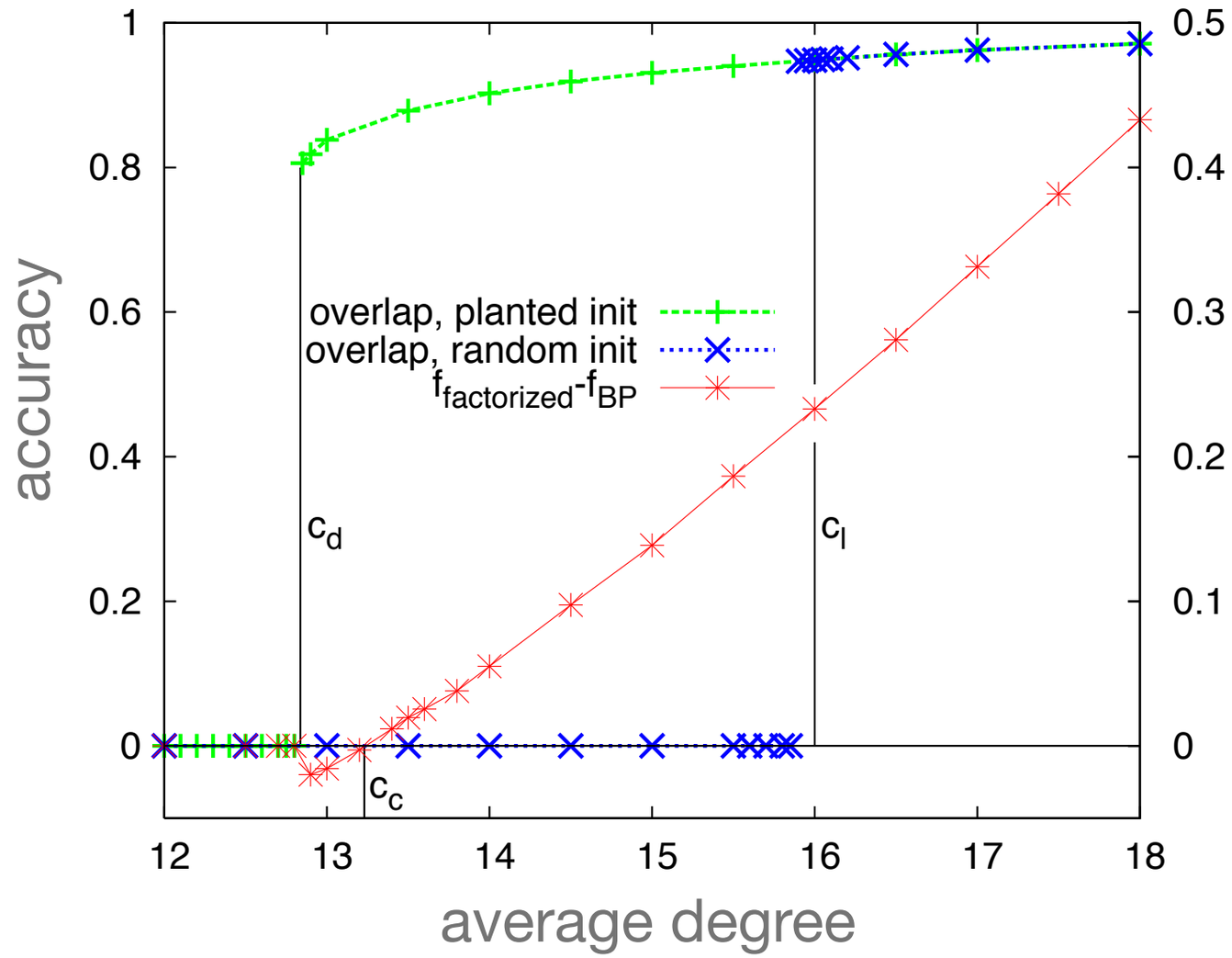
there is a fixed point where all nodes have uniform marginals...

at the transition, it becomes stable



conjectured by [Decelle, Krzakala, Moore, Zdeborová, '11]
proved by [Mossel, Neeman, Sly, `13; Massoulié '13]
for $k>2$ groups, not much is known rigorously...

Another regime: detectable but hard



find the 5-coloring!

in the hard region, BP has two fixed points: the trivial one and an accurate one

but we need some initial help to find the accurate one...

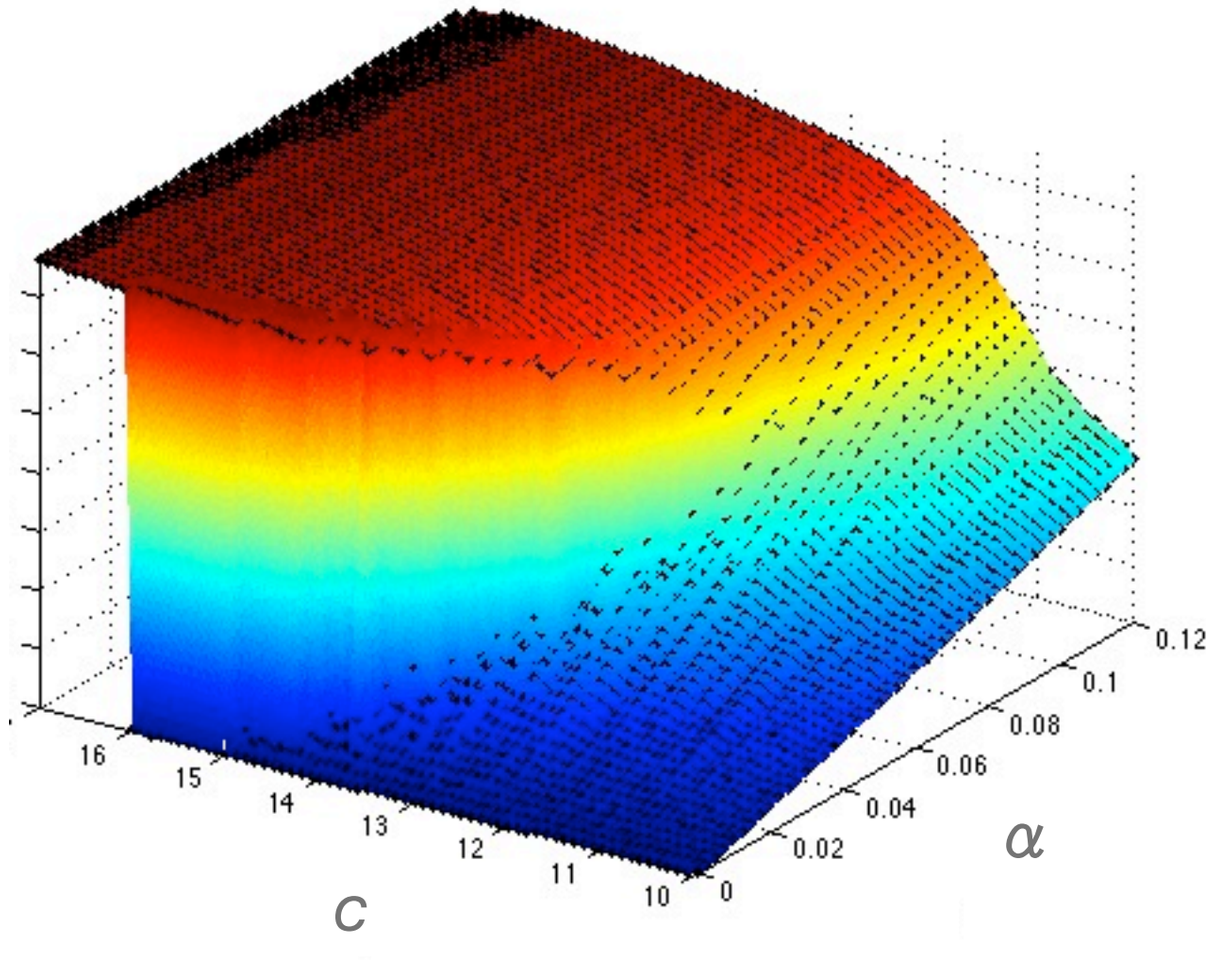
Phase transitions with metadata: what if we know some labels?

suppose we are given the correct labels for αn nodes for free

can we extend this information to the rest of the graph?

when α is large enough, knowledge percolates from the known nodes to the rest of the network

a line of discontinuities in the (c,α) plane, ending at a critical point

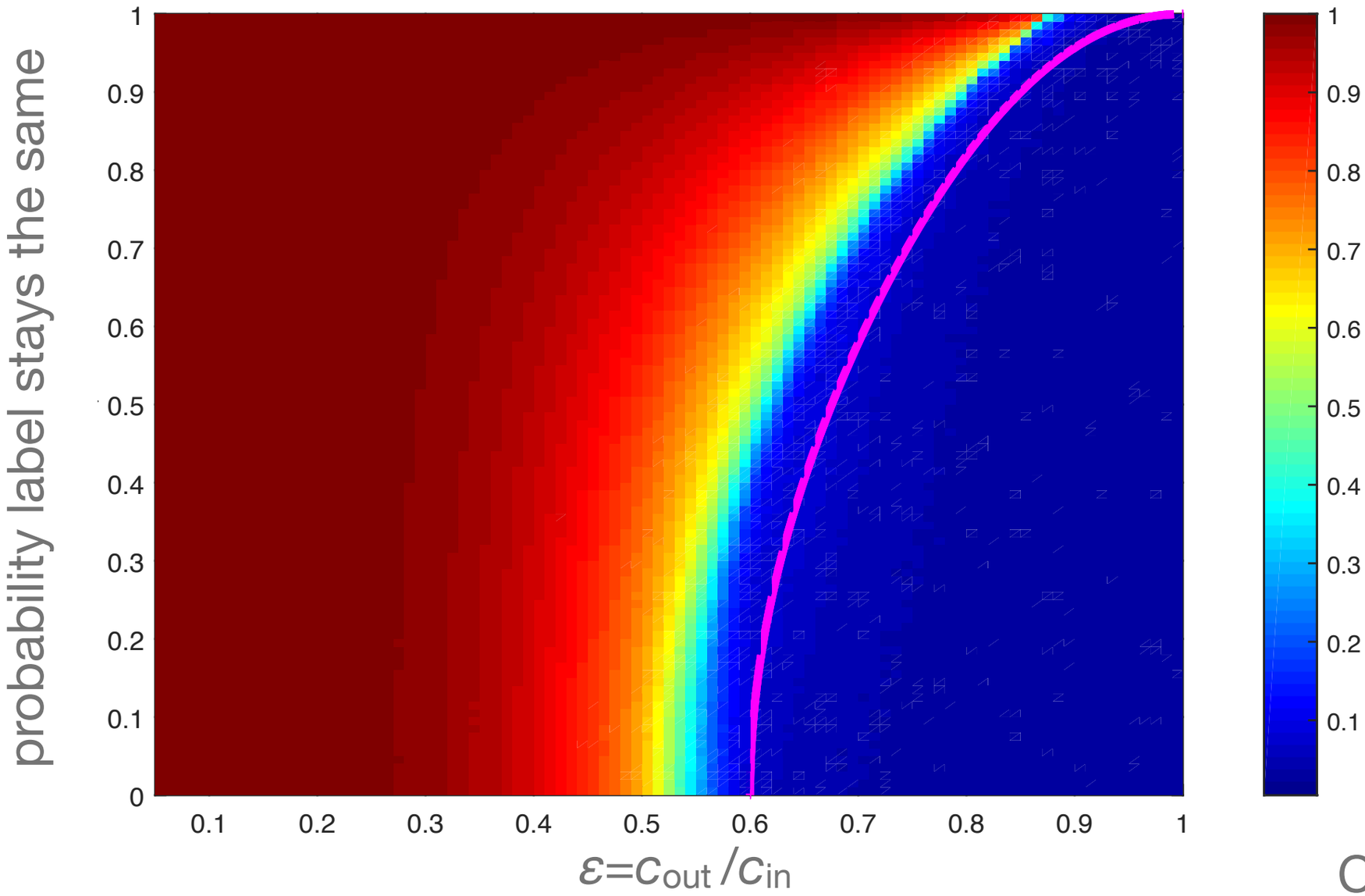


[Zhang, Moore, Zdeborová ’14]

Dynamic networks

what if nodes change their label, moving from group to group over time?

tradeoff between persistence of labels and the strength of the communities



[Ghasemian, Zhang, Clauset, Moore, Peel]

Extensions to richer data

can add metadata to nodes and edges: signed or weighted edges, nodes with social status, location, content...

for networks of documents, a model that combines overlapping communities with standard models of word frequencies

a network of 1,000 microprocessor patents (joint work with Sergi Valverde):

arithmetic	testing	power		
multiplexer	debugging	reset	protection	branching
buses	emulator	frequencies	transparent	prediction
microinstructions	error	pulses	security	concurrence
microprograms	traces	voltages	multi-tasking	speculation
	embedding	sensing	encryption	reordering
	jumps	driving	restricting	
	halting	oscillators		

using both text and links does better than either one alone

[Zhu, Yan, Getoor, Moore, *KDD* 2013]

Statistical significance and the energy landscape

Statistical significance and the temperature

recall the Boltzmann distribution: $P \sim e^{-\beta E}$ where $\beta = 1/T$

higher β = lower temperature = greedier algorithm = stronger structure

what happens if we look for more structure than is really there?

if we insist on pushing towards absolute zero, and the absolute optimum...

we find lots of near-optima, with nothing in common

BP bounces around them, never settling down

even if you could find the true optimum, would you care?

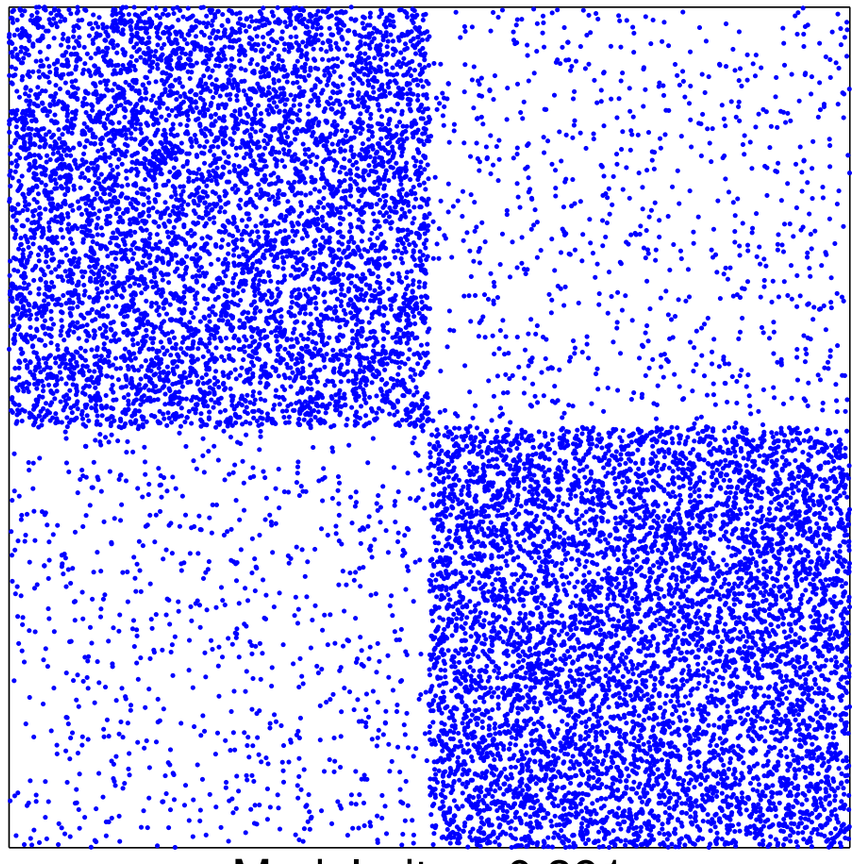
Statistical significance and the temperature

modularity Q = # within-group edges – expected number [Newman & Girvan]

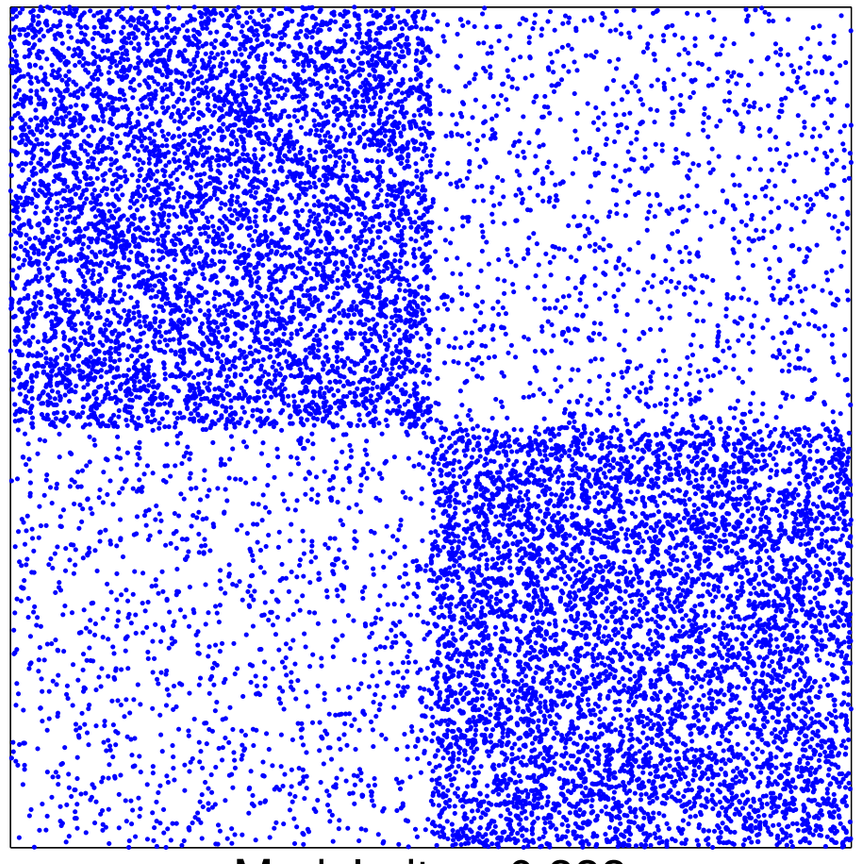
can be large even for random graphs (e.g. Guimera, Sales-Pardo, Amaral)

and yield inconsistent results in real ones (Good, Montjoye, Clauset)

quick! which graph has communities?



Modularity = 0.391



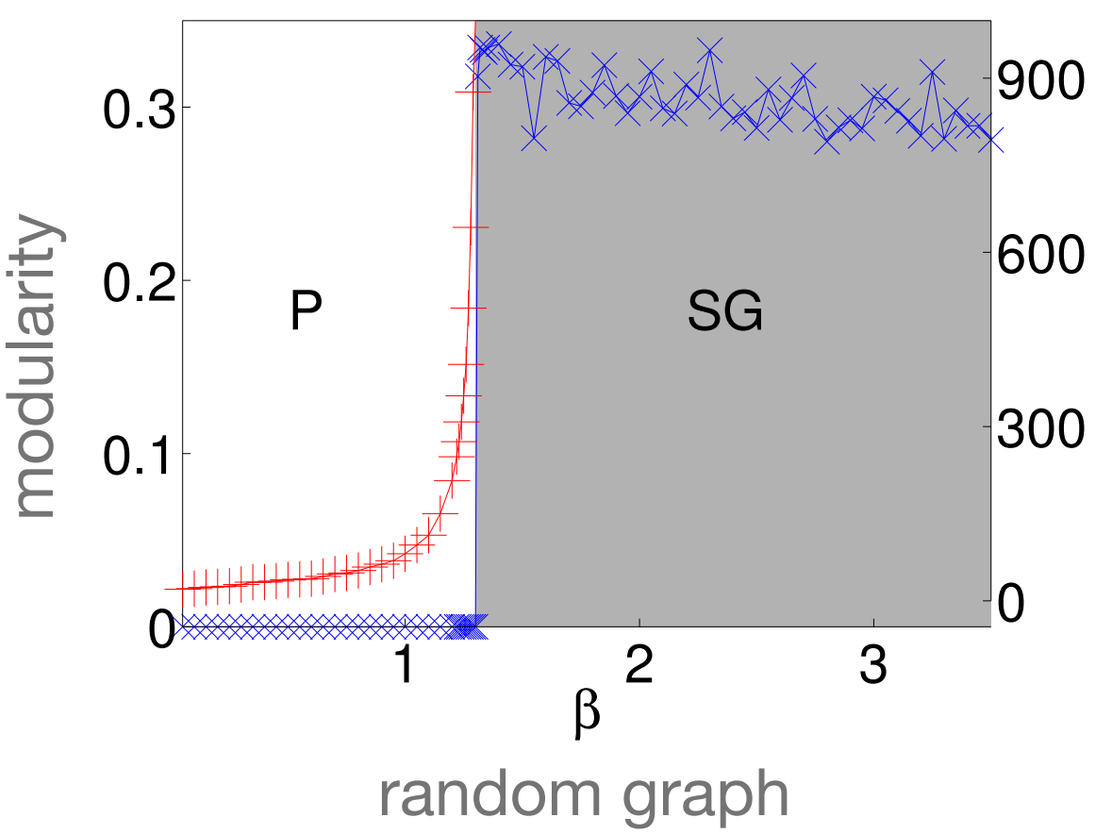
Modularity = 0.333

Real structure or glassy illusion?

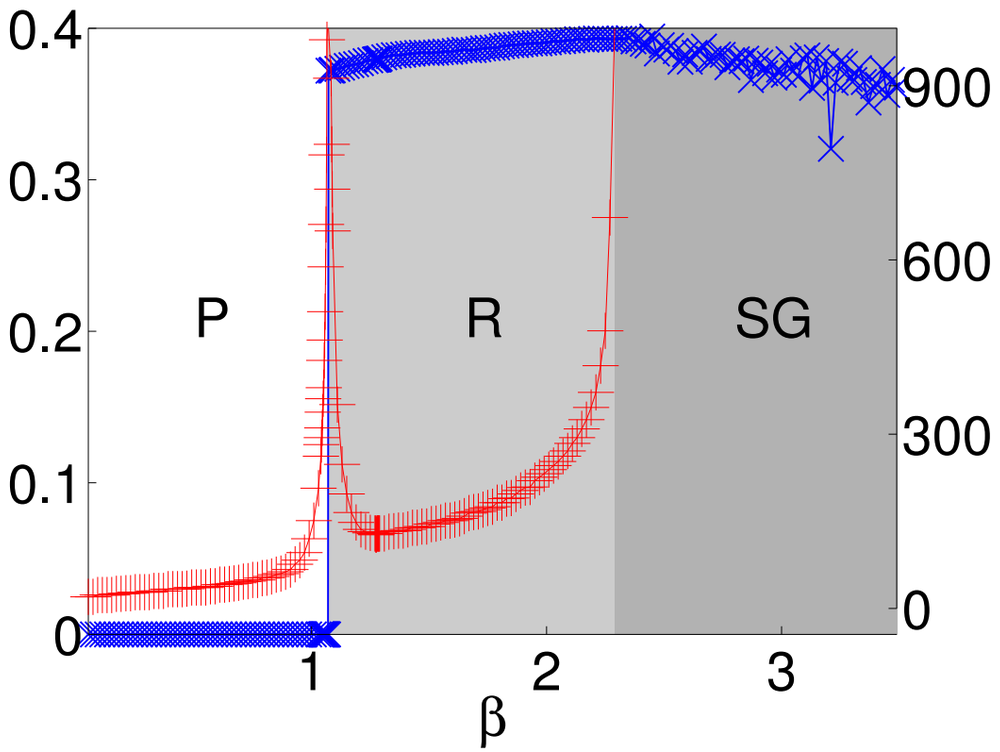
at low β (high T) the trivial fixed point is stable, BP finds zero modularity

if β is too large we're too greedy: a "spin glass" where BP fails to converge, wandering on a bumpy landscape of uncorrelated local optima

if there is real structure, there is a range of β where BP converges, and the consensus partition has high modularity

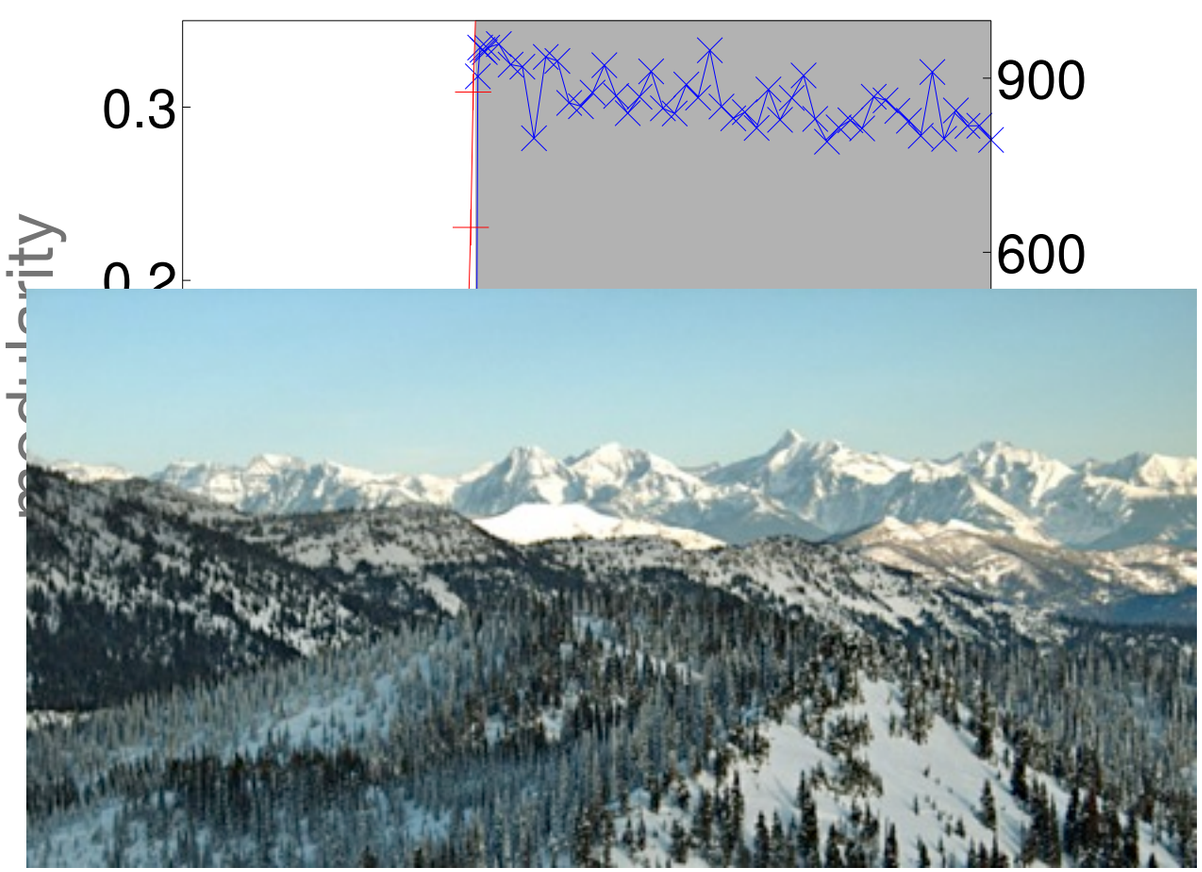


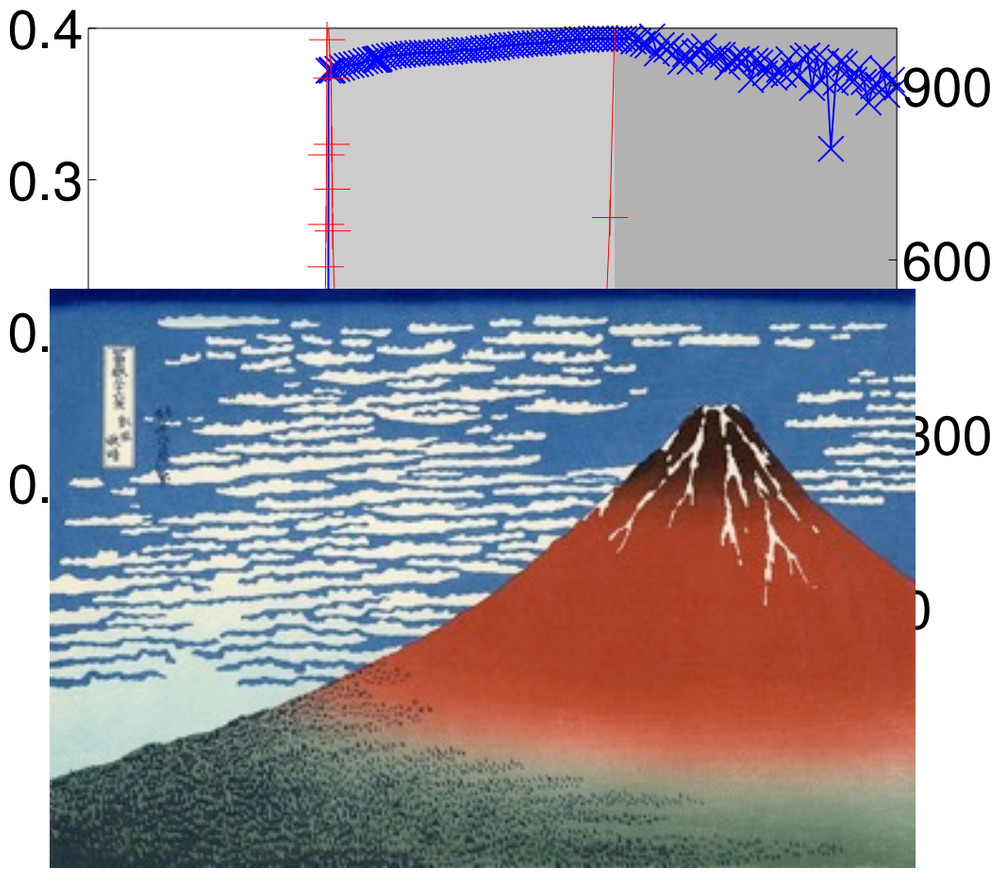
random graph



real block model

Real structure or glassy illusion?



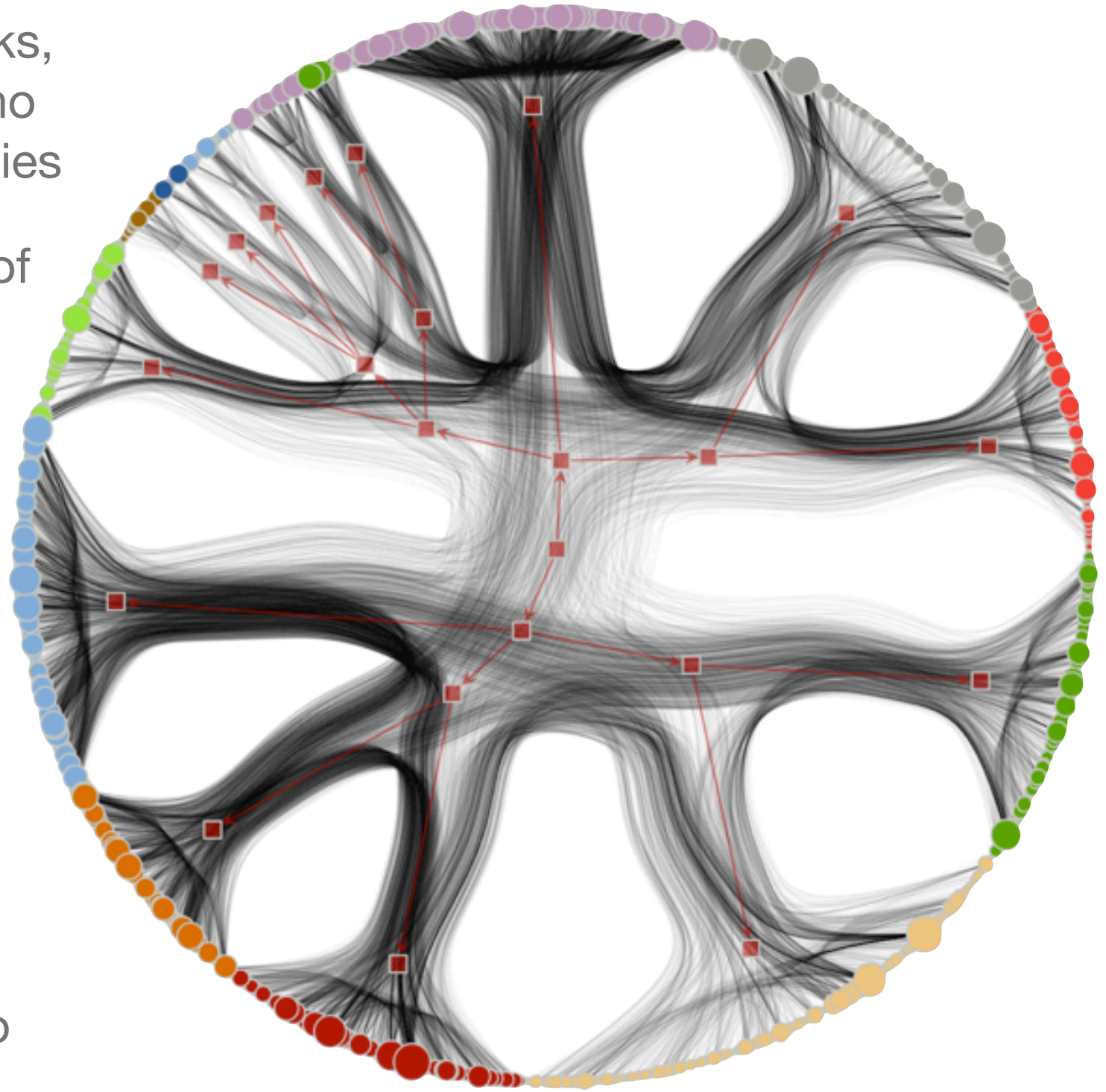


Hierarchical clustering

divide a network into subnetworks, until the remaining pieces have no statistically significant communities

reveals substructure in network of political blogs

don’t maximize modularity!
the consensus of many high-modularity structures is better than the “best” one



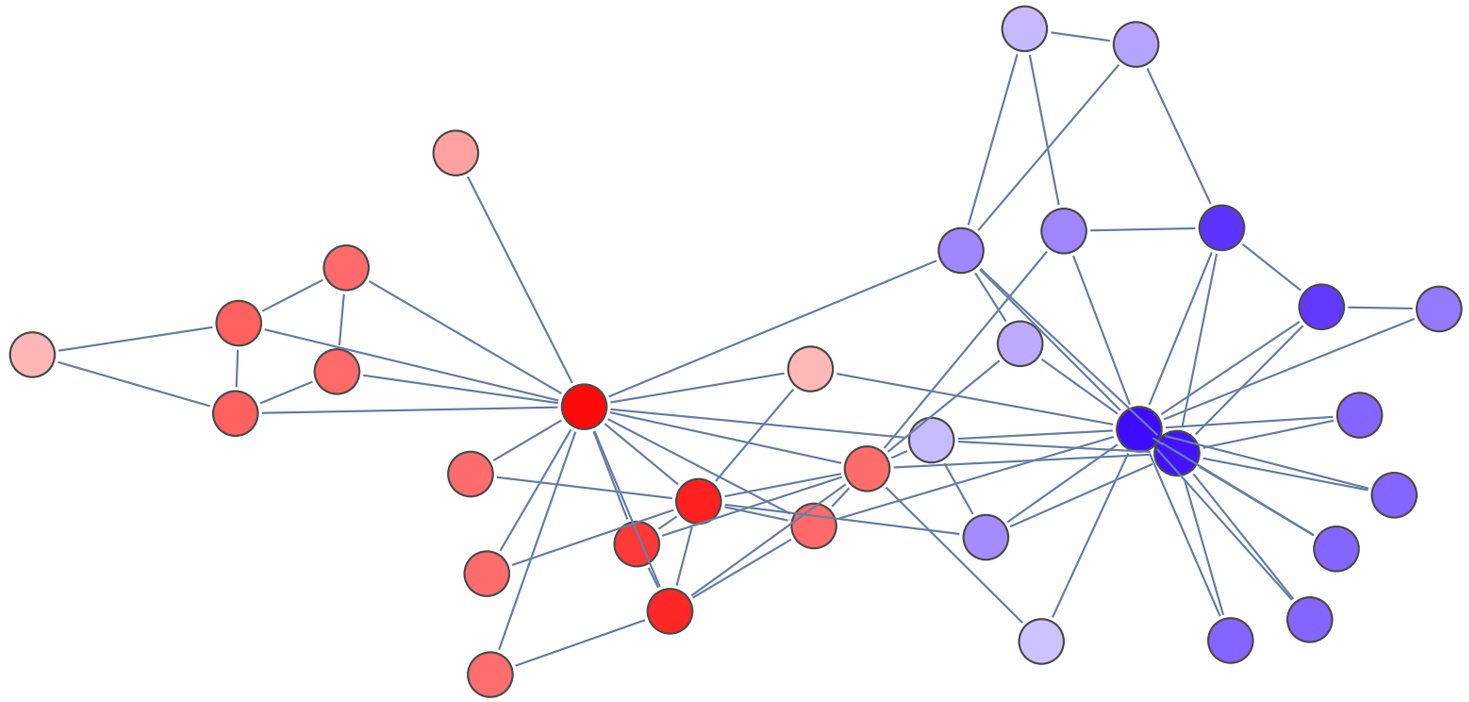
[Zhang and Moore, *PNAS* 2014]
image by Tiago de Paula Peixoto

Spectral methods and their redemption

Clustering nodes with eigenvalues

linear operators associated a graph: adjacency matrix, Laplacian, etc.

if there are 2 groups, label nodes according to the sign of the 2nd eigenvector

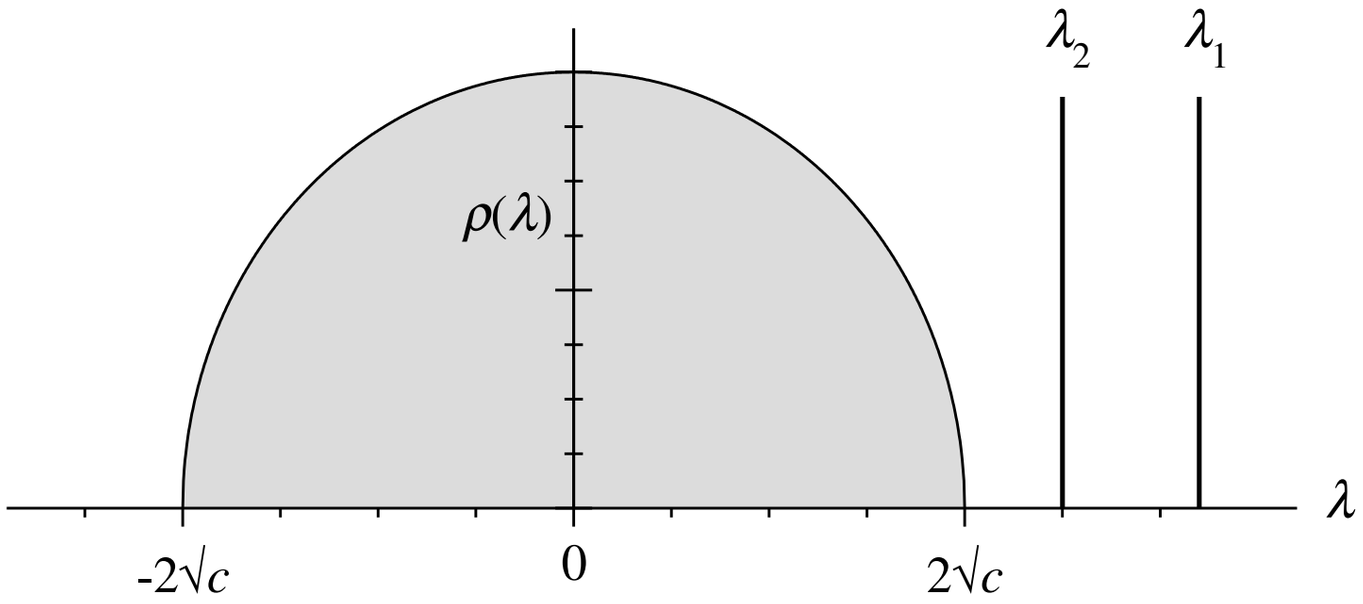


if there are *k* groups, look at the first *k* eigenvectors, and use your favorite clustering algorithm in R^k

When does this work?

using random matrix theory, can compute the typical spectrum of a graph generated by the stochastic block model

“bulk” follows the Wigner semicircle law



communities are detectable as long as λ_2 lies outside this bulk...

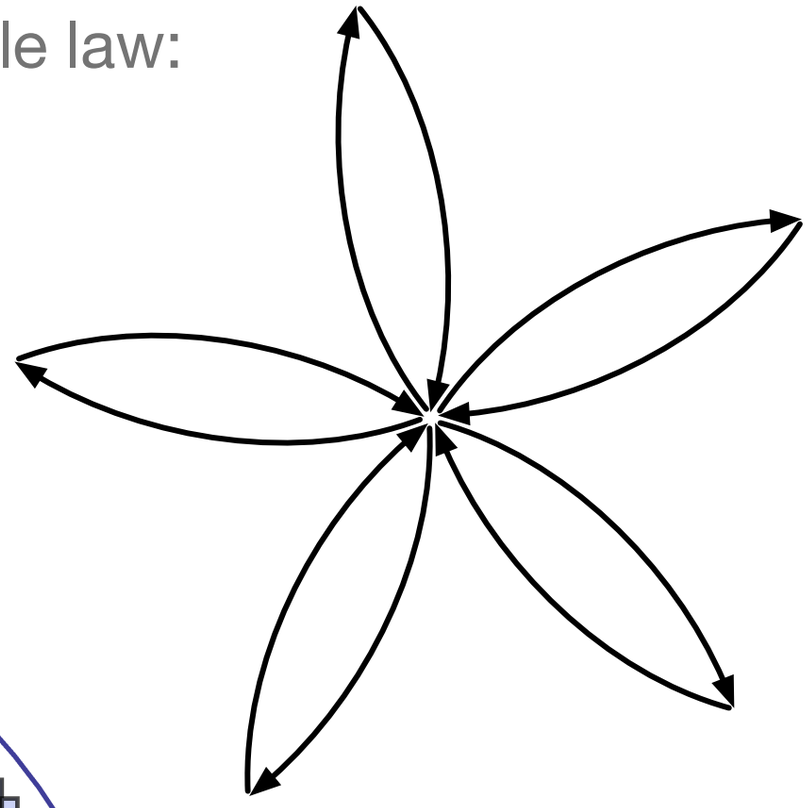
crosses at the detectability transition... if the graph is dense enough

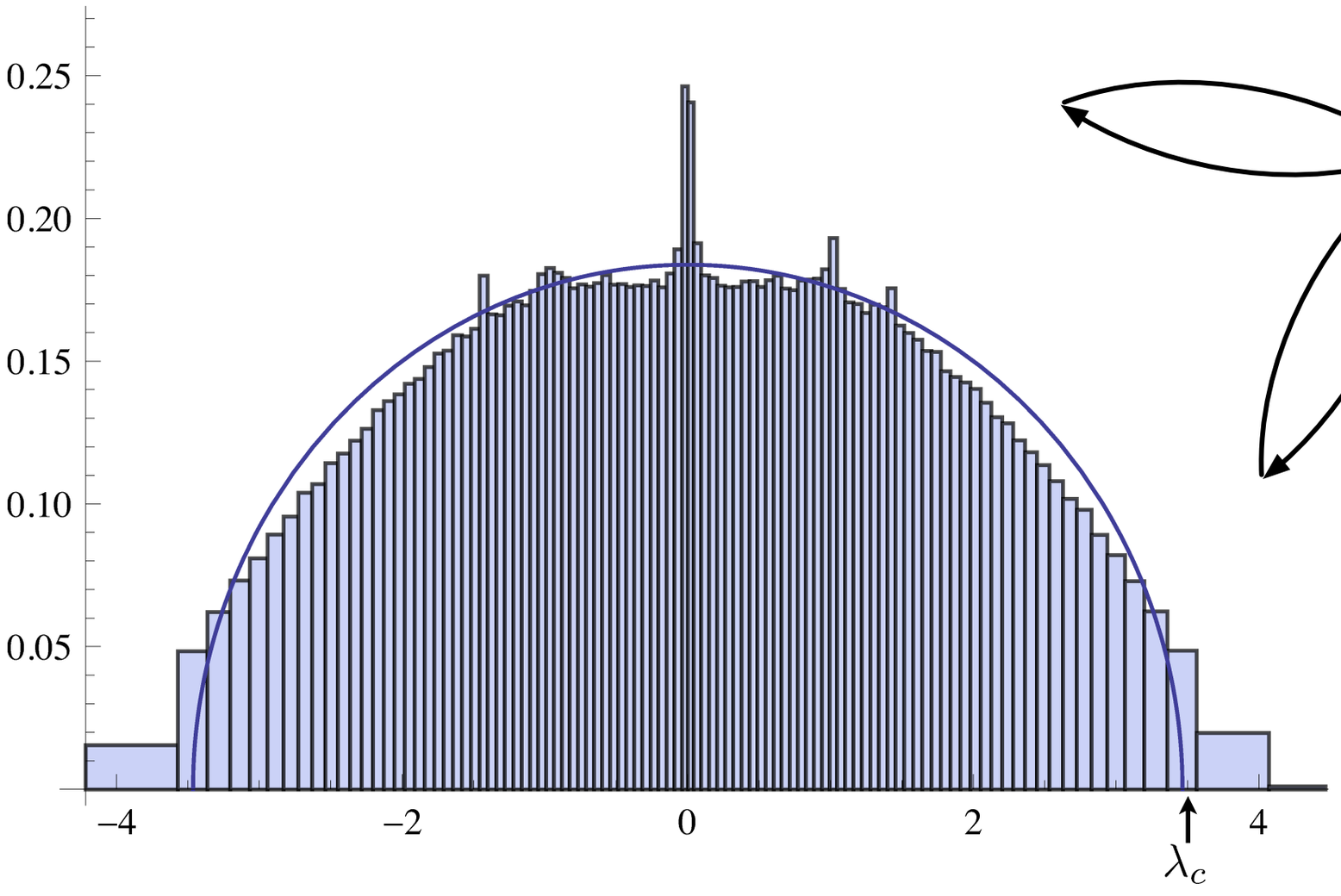
[Nadakuditi and Newman, `12]

But in the sparse case...

if v has degree d, applying A^2 has d ways to return to v
thus A has an eigenvector with an eigenvalue at least $\sqrt{d}$

these localized eigenvalues deviate from the semicircle law:
informative eigenvectors get lost in the bulk

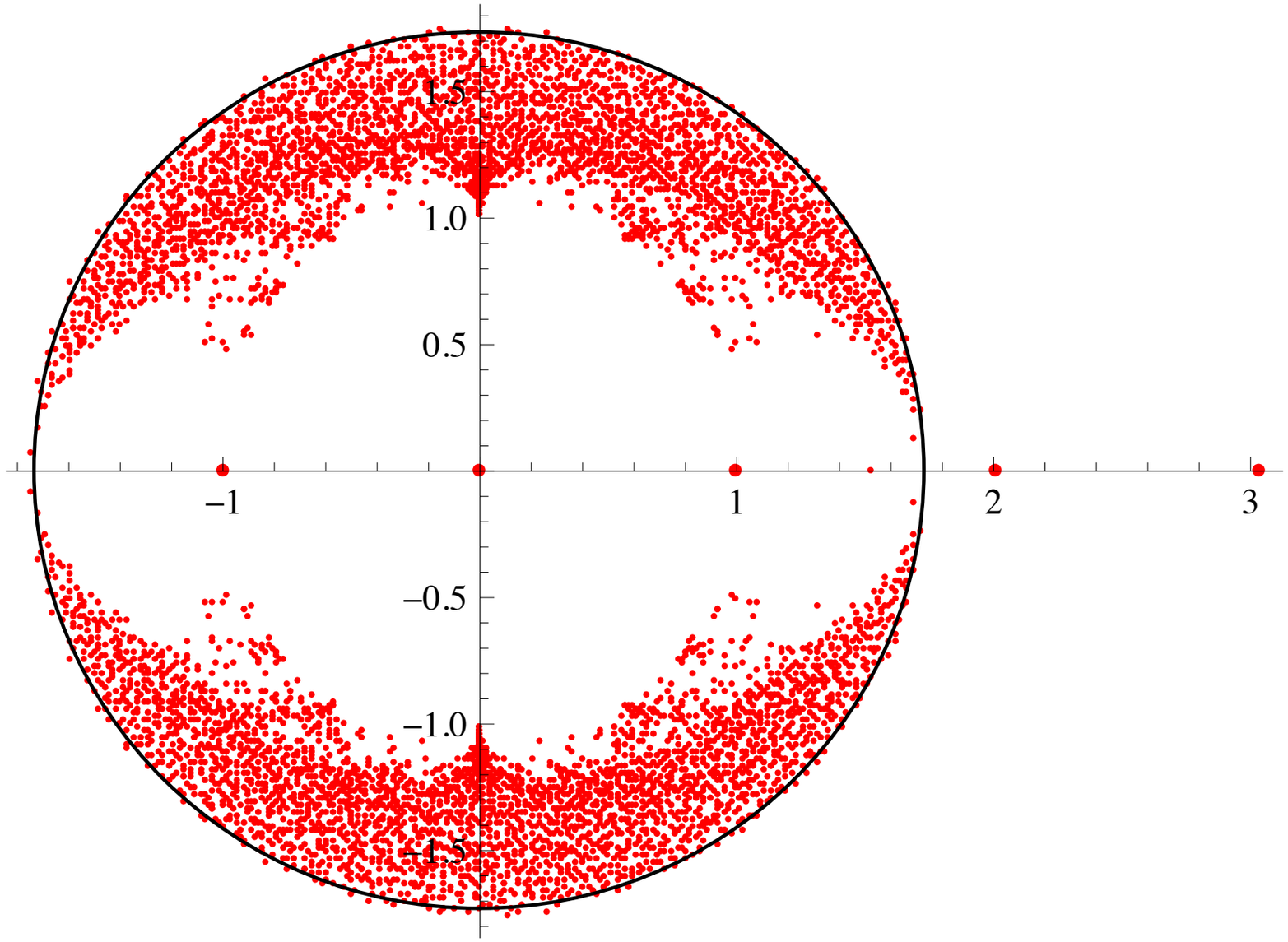


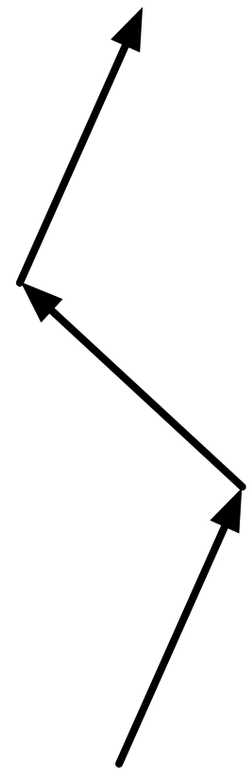


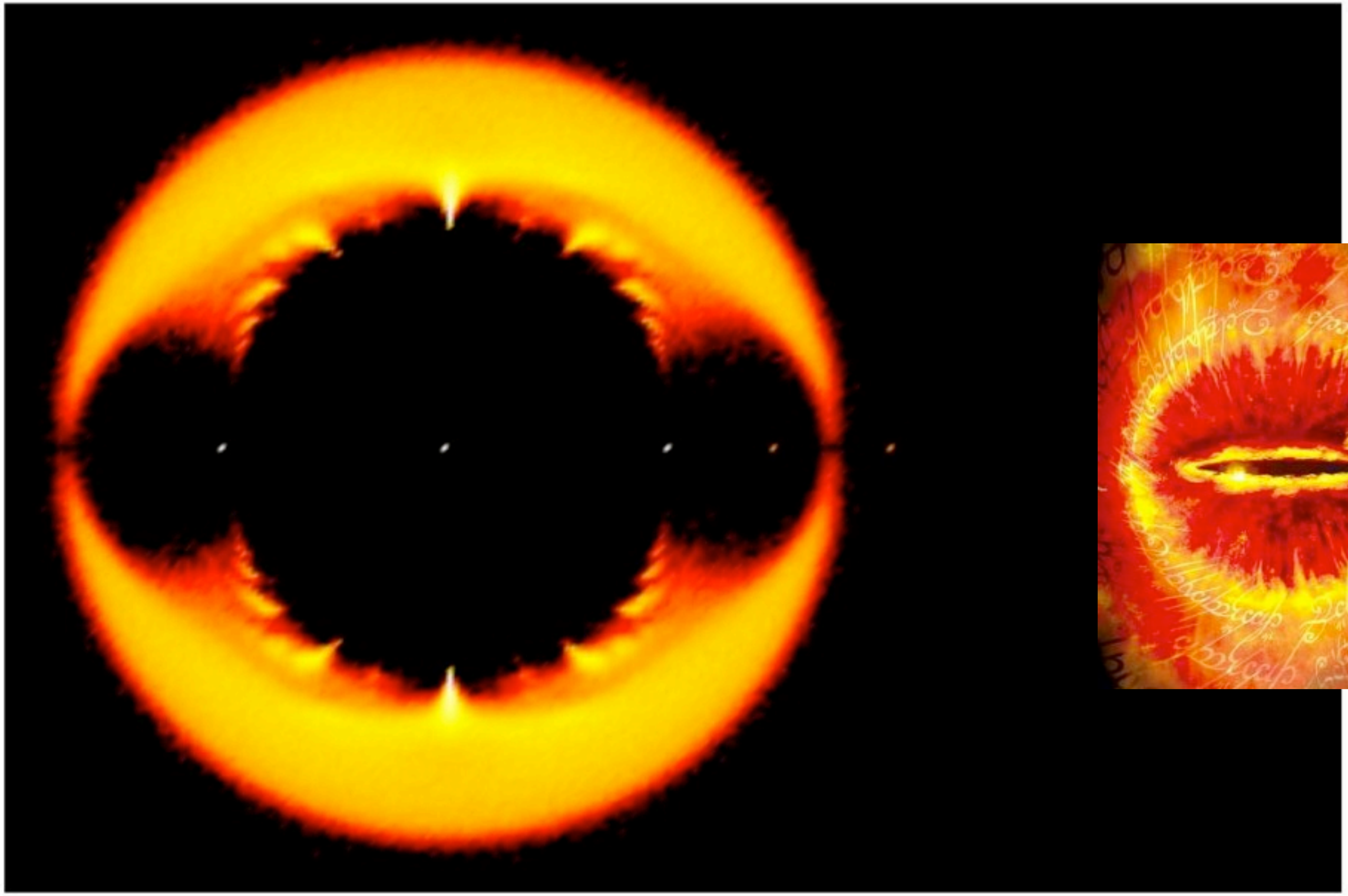
The non-backtracking operator

B is a walk on directed edges, with backtracking prohibited:
prevents paths from returning to a high-degree vertex, or getting stuck in trees

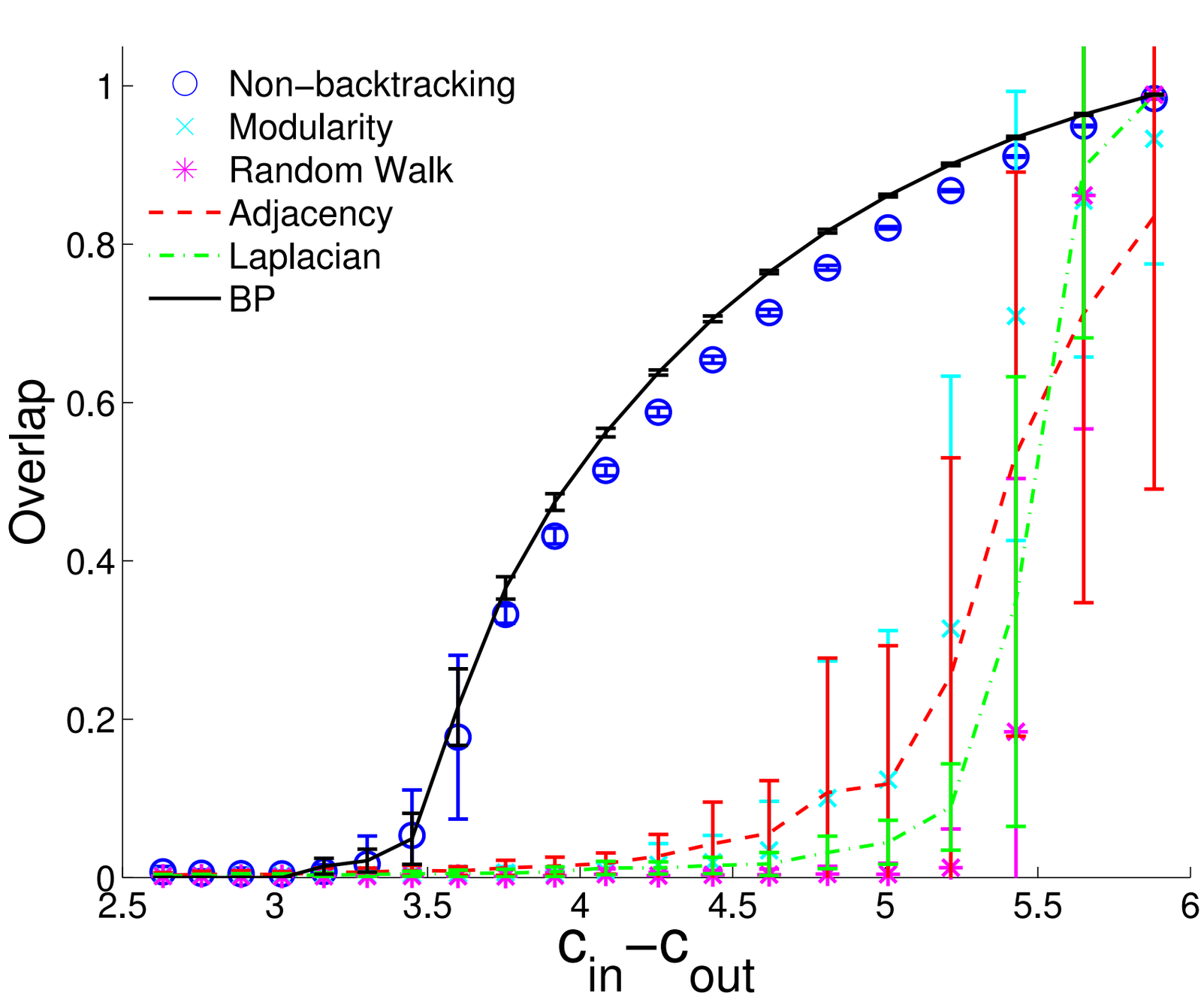
bulk of *B*'s spectrum is confined to a disk of radius $\sqrt{c}$, even in the sparse case







Comparing with standard spectral methods

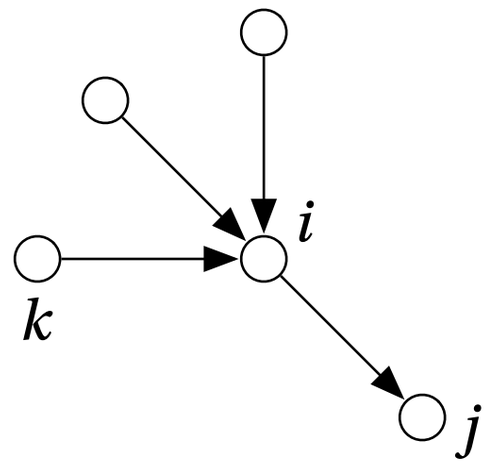


You may ask yourself,
Well, how did we get here?

---

expand the BP equations around the trivial fixed point to first order:

the matrix of derivatives is a tensor product of B with a $k \times k$ matrix



no echo chamber = non-backtracking

bulk confined = works all the way down to the detectability transition

[Krzakala, Moore, Mossel, Neeman, Sly, Zdeborová, Zhang, *PNAS* 2013]

[Bordenave, Lelarge, Massoulié, 2015]

Morals

the most likely model often overfits!

even if you can find the best fit, you might not want to...

...if it's just one of many local optima that have nothing in common

the consensus of many likely fits is a better judge of structure

Belief Propagation finds this consensus in nearly-linear time (when it works)...

...gives us an analytic framework for finding phase transitions

...and linearization yields new spectral algorithms

Finally, the shape of the energy landscape can tell us whether structures are statistically significant

Physics culture meets machine learning

"as simple as possible (but no simpler)"

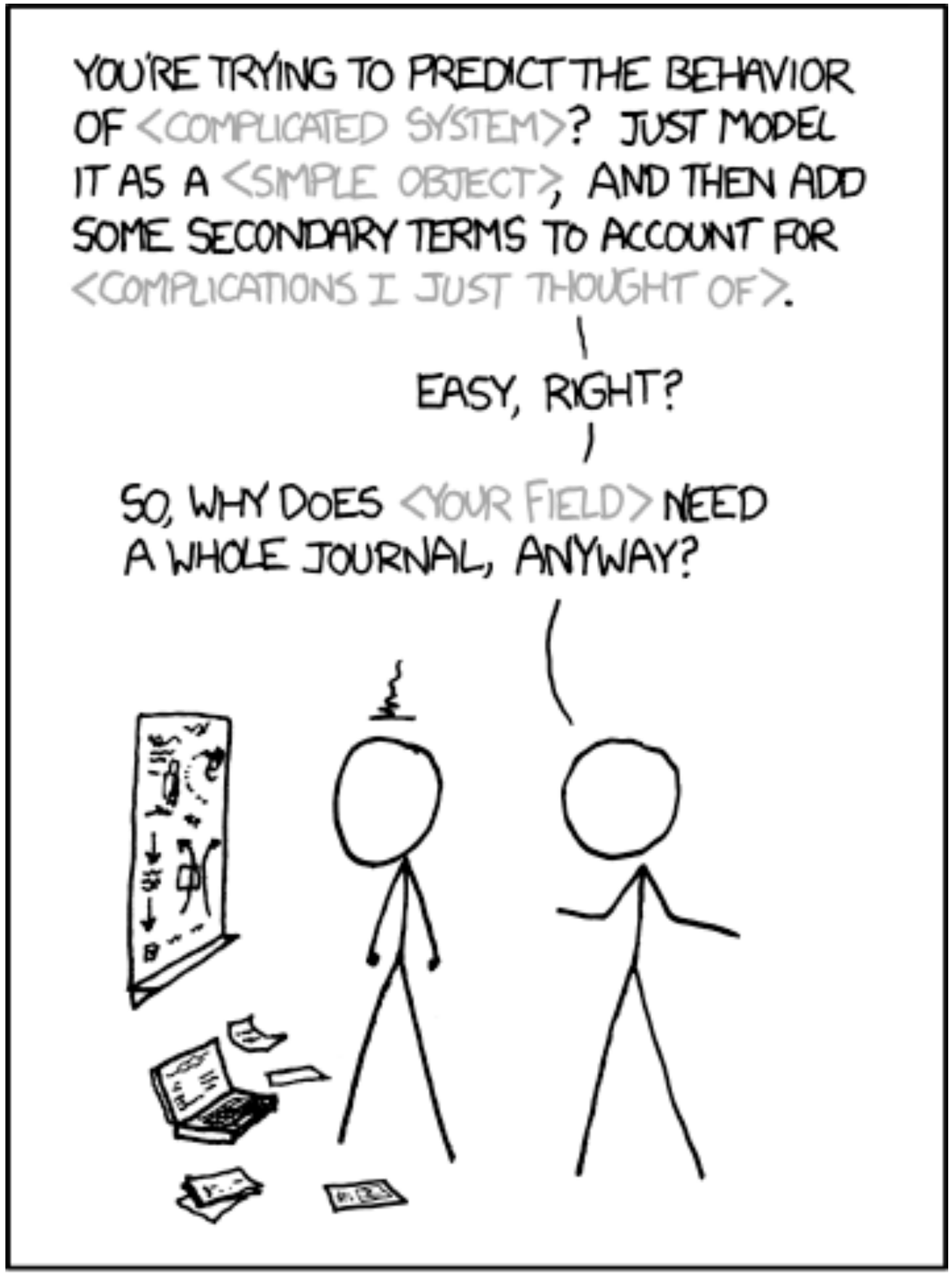
mathematical elegance and tractability pays off, even with real data

in practice: simpler models are easier to optimize, giving faster algorithms

in theory: analytic results on the strengths and weaknesses of these algorithms, fundamental limits on when (and how well) these problems can be solved

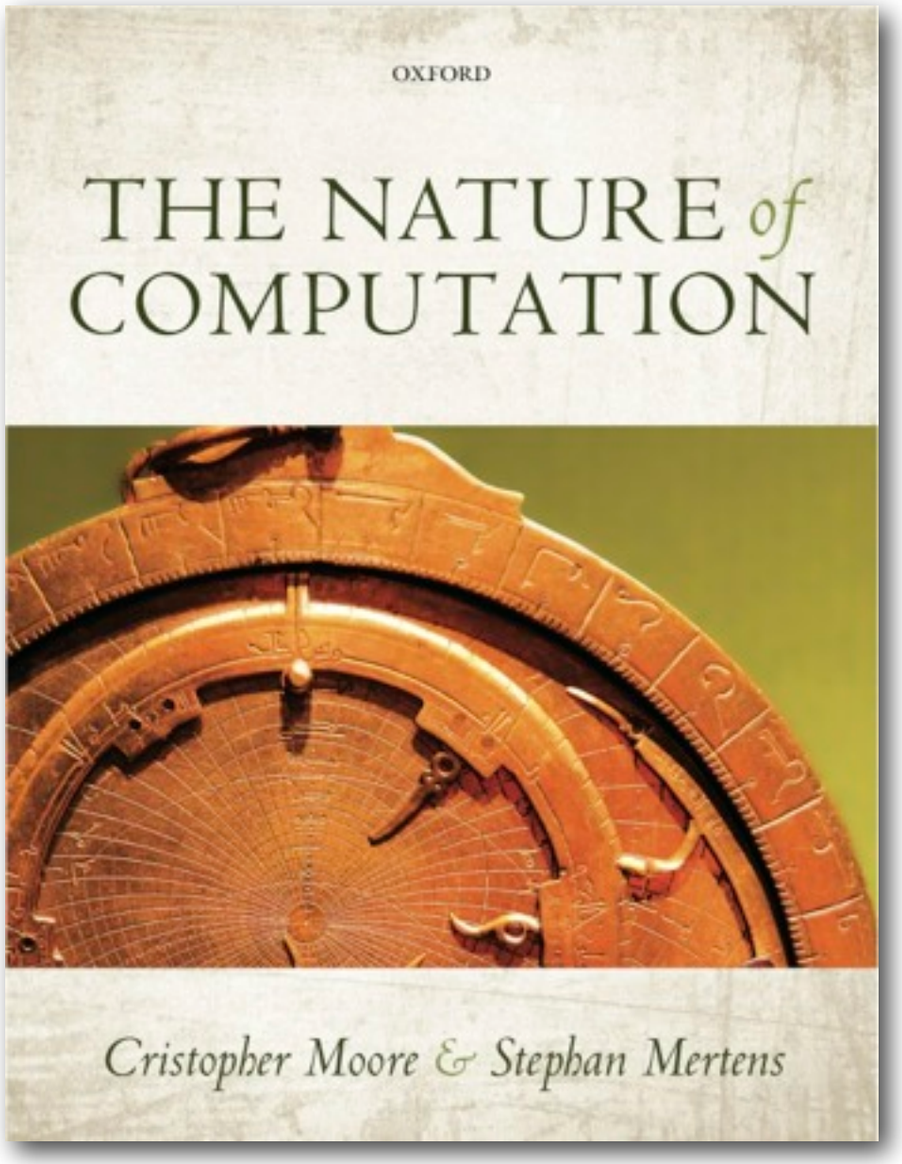
insights are better than small improvements in accuracy:
although big improvements in accuracy usually come from insights

the physics picture of the "energy landscape" can help us decide whether the structures we find are really there



LIBERAL-ARTS MAJORS MAY BE ANNOYING SOMETIMES, BUT THERE'S *NOTHING* MORE OBNOXIOUS THAN A PHYSICIST FIRST ENCOUNTERING A NEW SUBJECT.

Shameless Plug



www.nature-of-computation.org

To put it bluntly: this book rocks! It somehow manages to combine the fun of a popular book with the intellectual heft of a textbook.

Scott Aaronson, MIT

This is, simply put, the best-written book on the theory of computation I have ever read; one of the best-written mathematical books I have ever read, period.

Cosma Shalizi, Carnegie Mellon