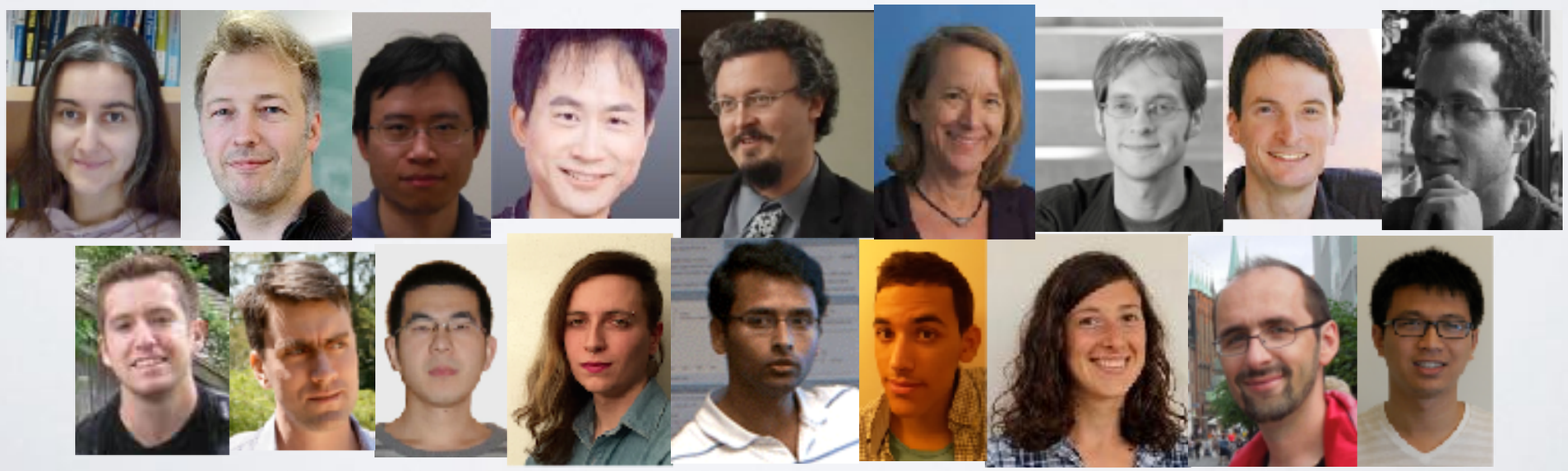
Physics, Phase Transitions in Inference, and Networks

Cristopher Moore, Santa Fe Institute

with Aurelien Decelle, Lenka Zdeborová, Florent Krzakala, Xiaoran Yan, Yaojia Zhu, Cosma Shalizi, Lise Getoor, Aaron Clauset, Mark Newman, Elchanan Mossel, Joe Neeman, Allan Sly, Pan Zhang, Jess Banks, Praneeth Netrapalli, Thibault Lesieur, Caterina de Bacco, Roman Vershynin, and Jiaming Xu



Statistical inference $\Leftrightarrow$ statistical physics

How can we find patterns in noisy data?

Phase transitions

Optimal algorithms

Information vs. computation

Why least squares?

the most common way to fit a line to noisy data

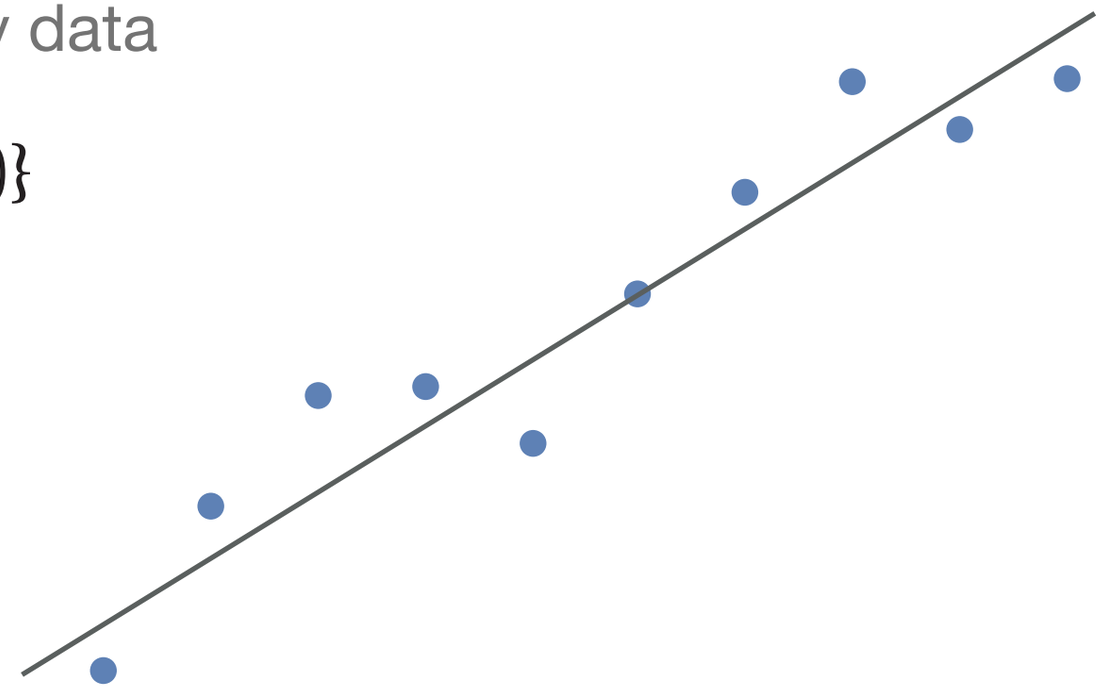
data points $Y = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

model: $y_i = a x_i + b$

find *a,b* that minimize

$$\sum_i (y_i - (a x_i + b))^2$$

but why?



A model of noise

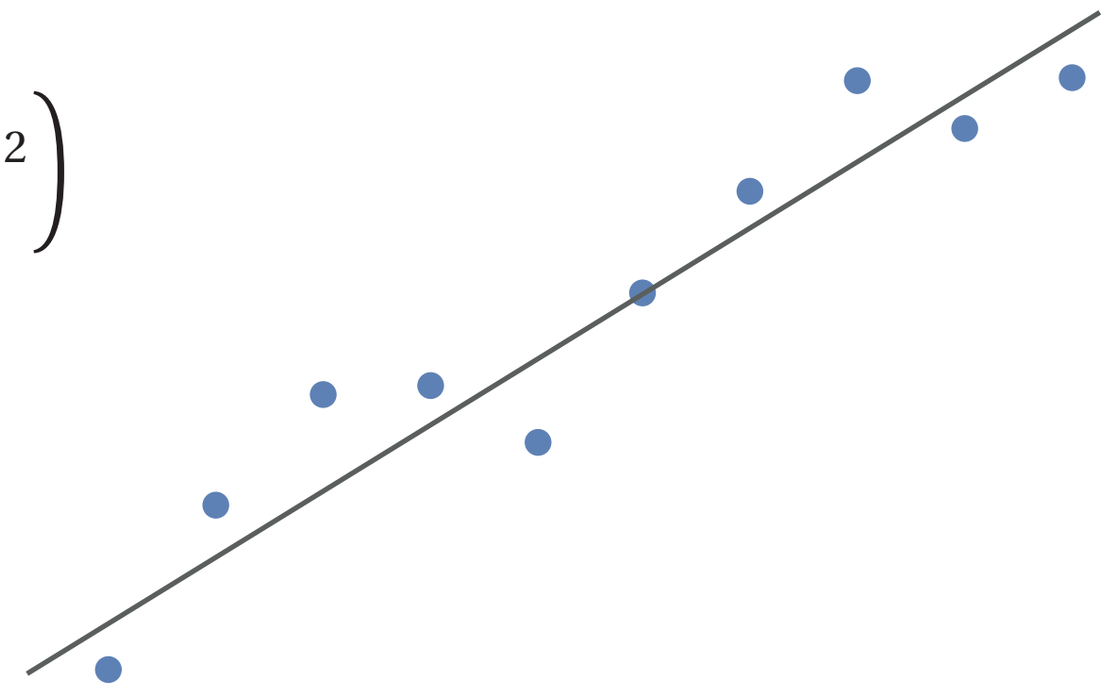
a model with noise: $y_i = a x_i + b + w$

where *w* is Gaussian, $P(w) \propto \exp\left(-\frac{1}{2\sigma} w^2\right)$

total probability of the data is

$$P(Y \mid a, b) = \prod_i P(y_i \mid a, b)$$

$$\propto \exp\left(-\frac{1}{2\sigma} \sum_i (y_i - (a x_i + b))^2\right)$$



Bayes: posterior (with flat prior) $P(a, b|Y) \propto P(Y|a, b)$

least squares = maximum likelihood estimate

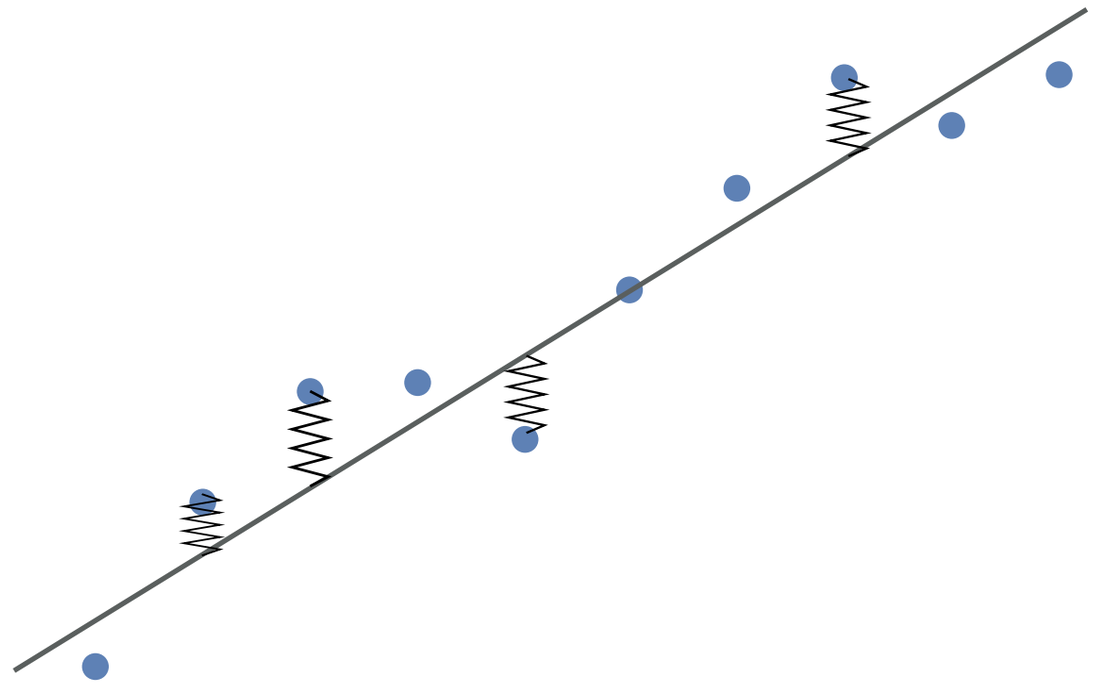
From probability to energy

define the energy of (*a*,*b*) as $E = -\log P$

$$E = \frac{1}{2\sigma} \sum_i (y_i - (a x_i + b))^2$$

springs between the model and data

$$E = \frac{1}{2} k x^2$$



maximizing *P* = minimizing *E*

maximum likelihood estimate = ground state

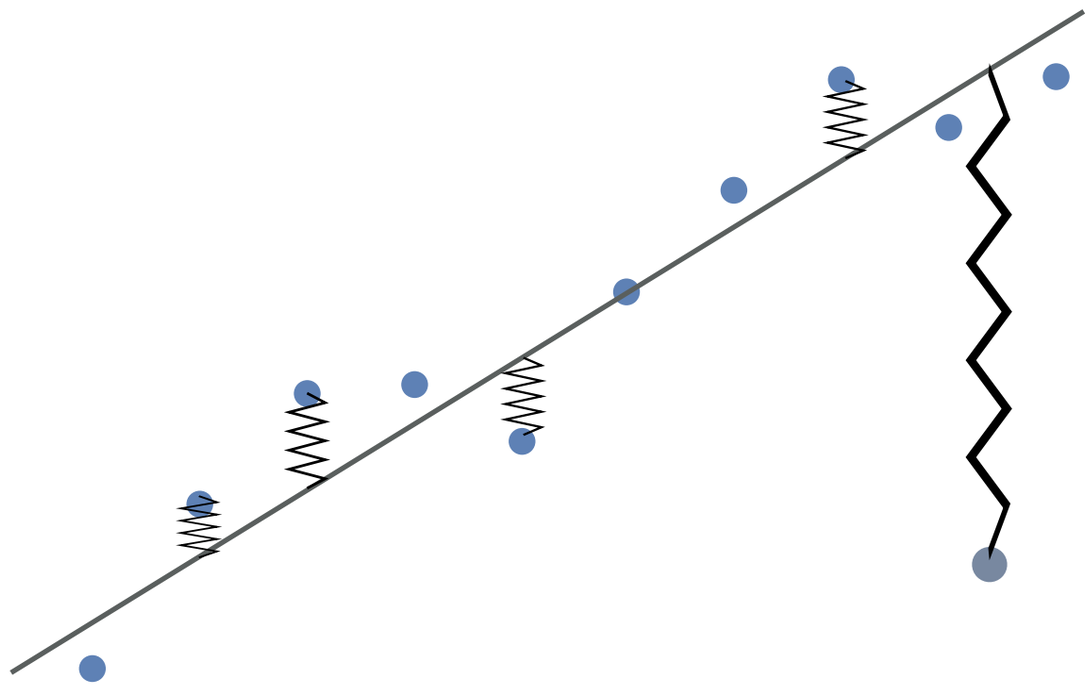
but what if the energy were different?

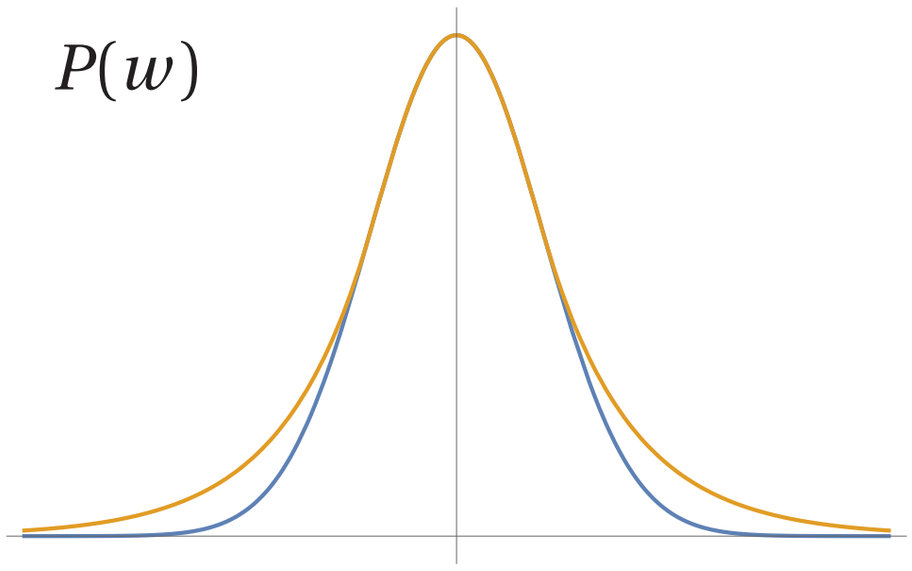
Changing the model

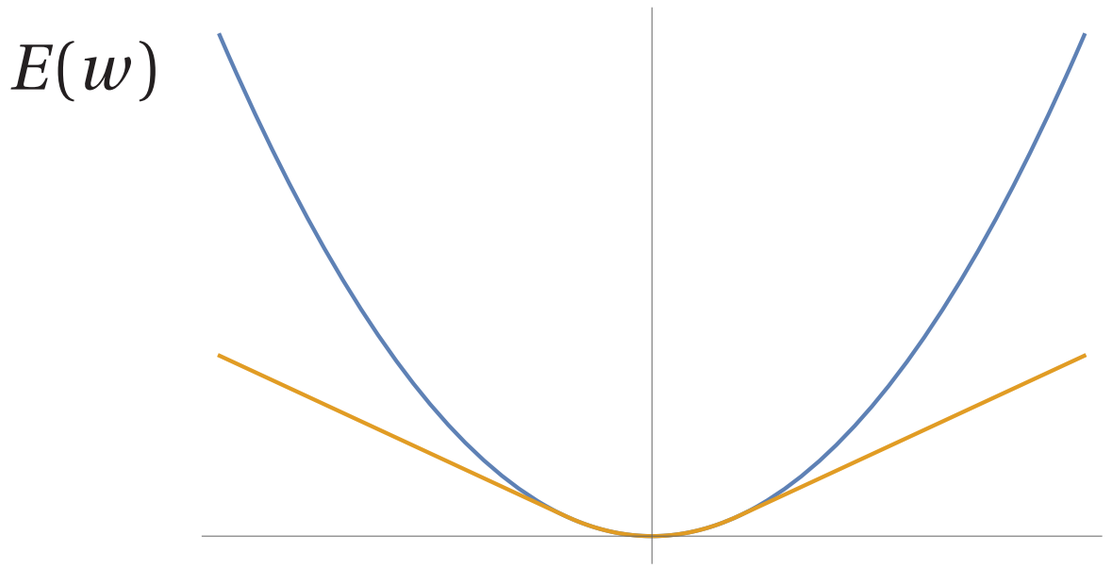
outliers skew our estimates

use a noise model with heavier tails

“gooey springs” that exert less force at large distances







Uncertainty, equilibrium, and the energy landscape

[Bayes] don't just give an estimate!
what's the posterior distribution?

[Boltzmann] at thermal equilibrium,

$$P(s) \propto \mathrm{e}^{-E(s)/T}$$

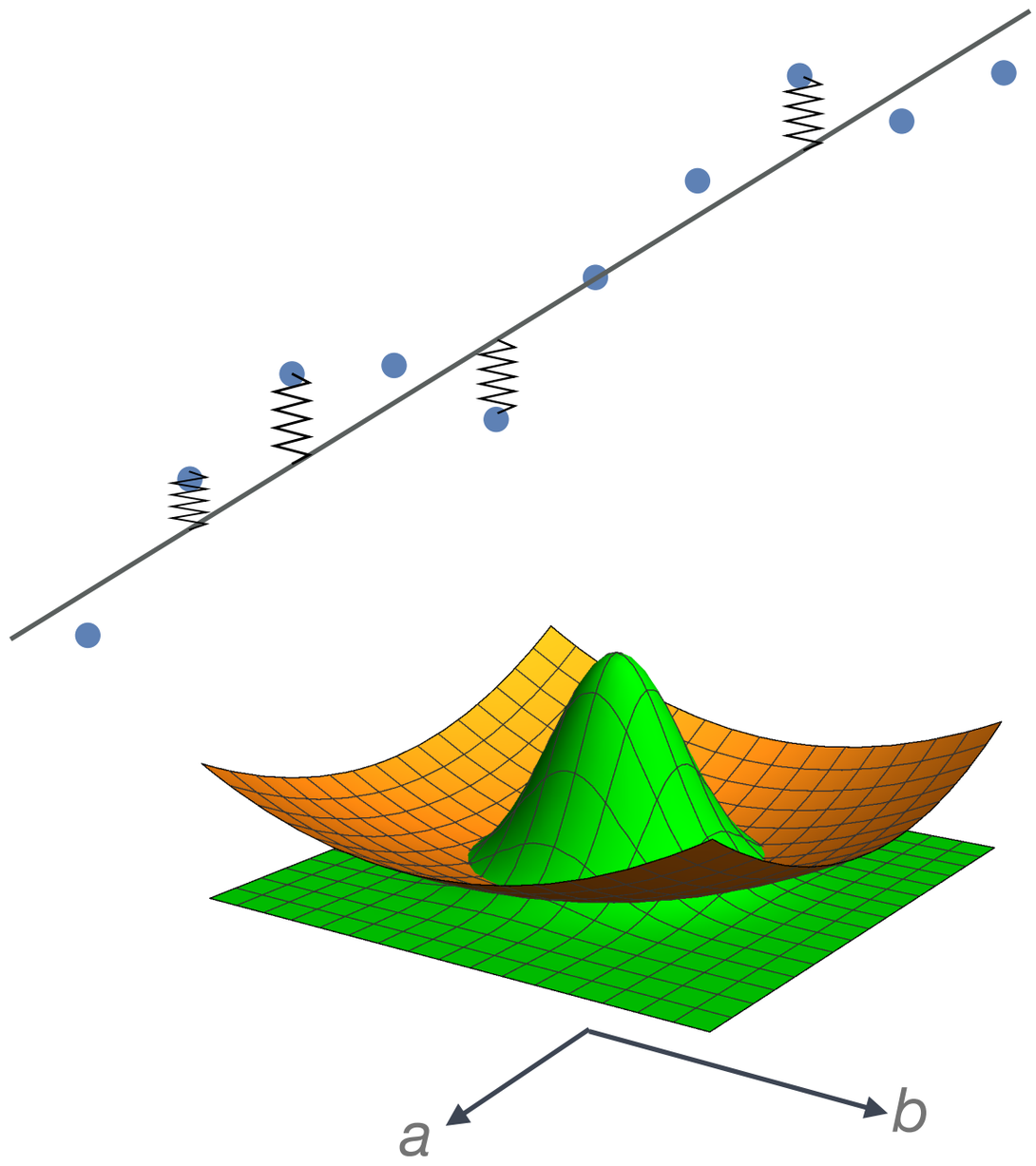
low T: concentrated on ground states

high T: uniform

thermal noise: $T=\sigma$ (or looser springs)

$E(a,b)$ defined by model and data

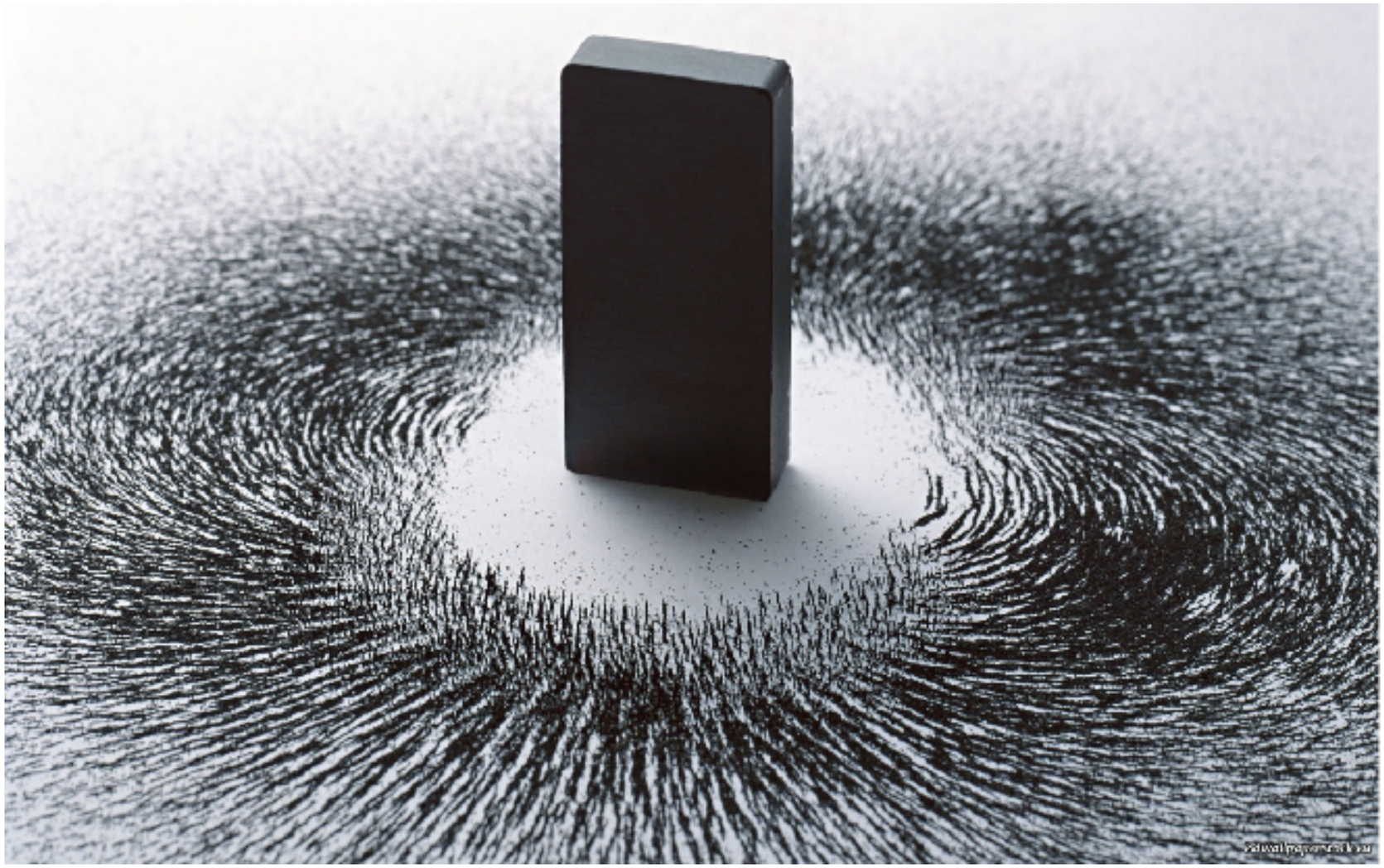
posterior distribution = equilibrium



The Ising model of magnetism

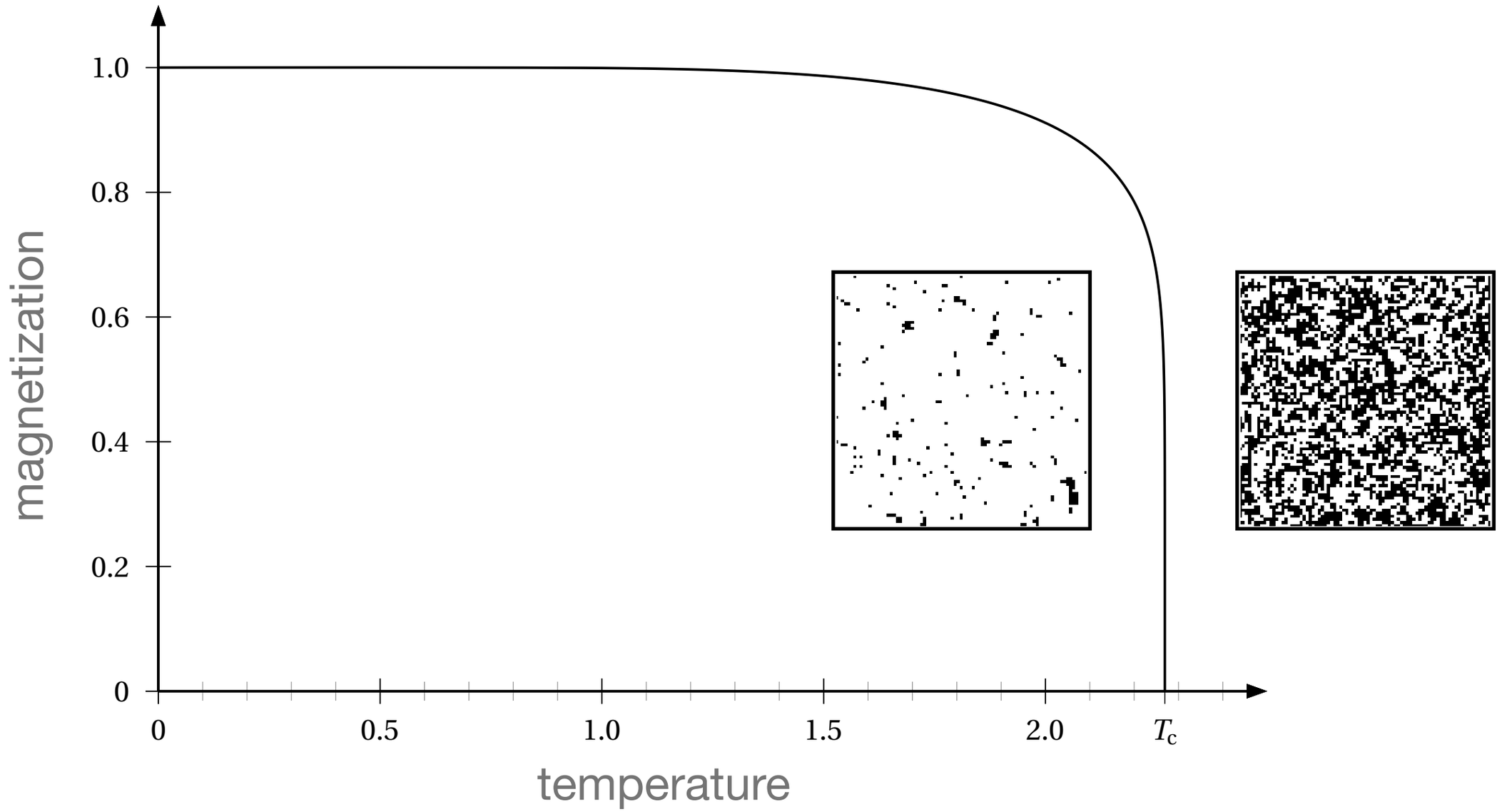
the atoms of a block of iron interact with their neighbors ↑↑↑↓↓↑↑↓↑↑↑↑

when these interactions are strong enough, or the temperature is low enough, they line up and form a magnetic field



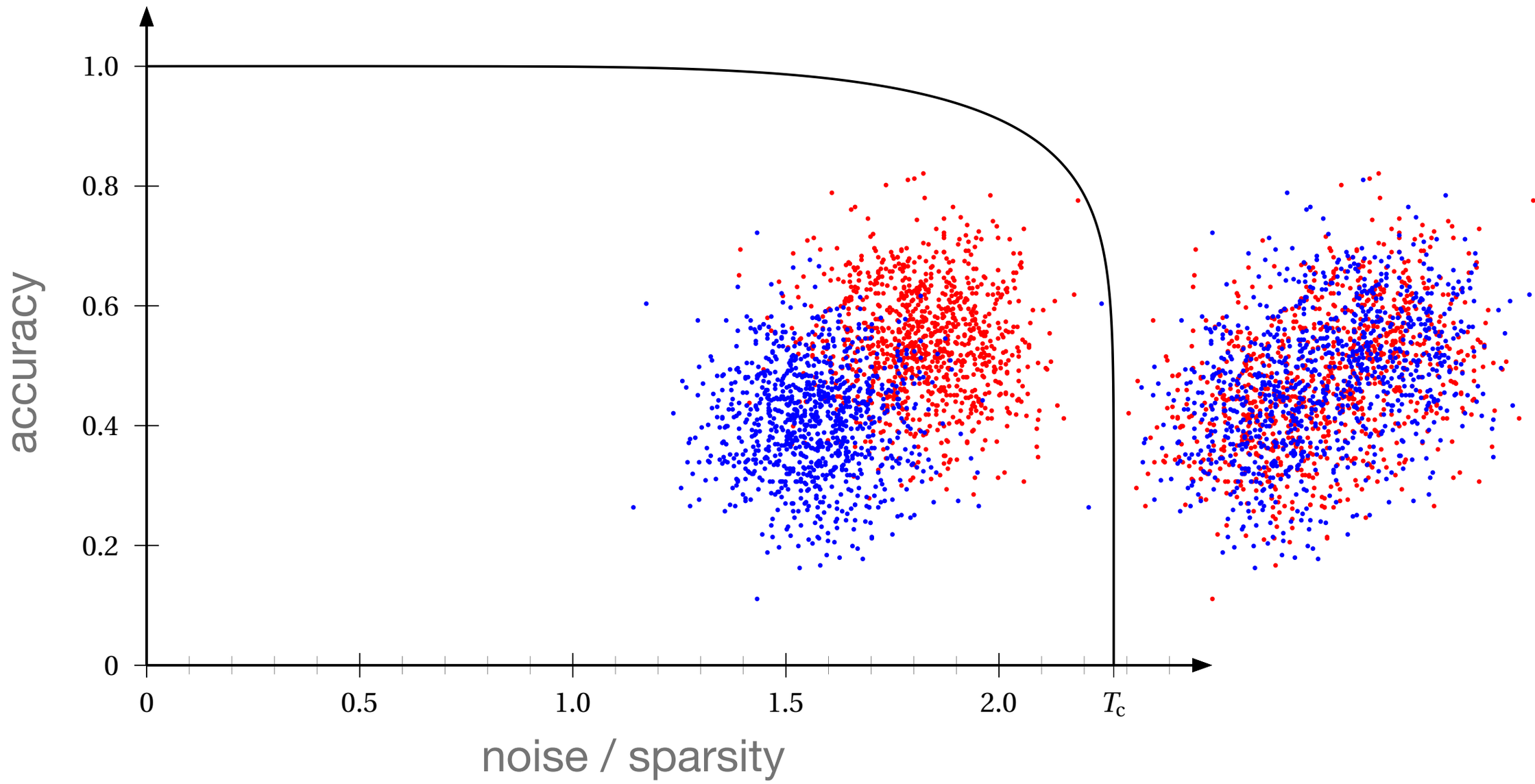
The Ising model of magnetism

at a critical temperature, the iron suddenly loses its magnetic field: atoms become uncorrelated, no long-range information



Fitting models to data

when data is too noisy or too sparse, the posterior distribution of a model becomes uncorrelated with the ground truth



Bumpy landscapes

least squares has a landscape with one optimum, and the Ising model has two

but a “spin glass” with energy $E = -\sum_{(i,j)} J_{ij} s_i s_j$ can have exponentially many

suppose the interactions J_{ij} depend on the data and the model

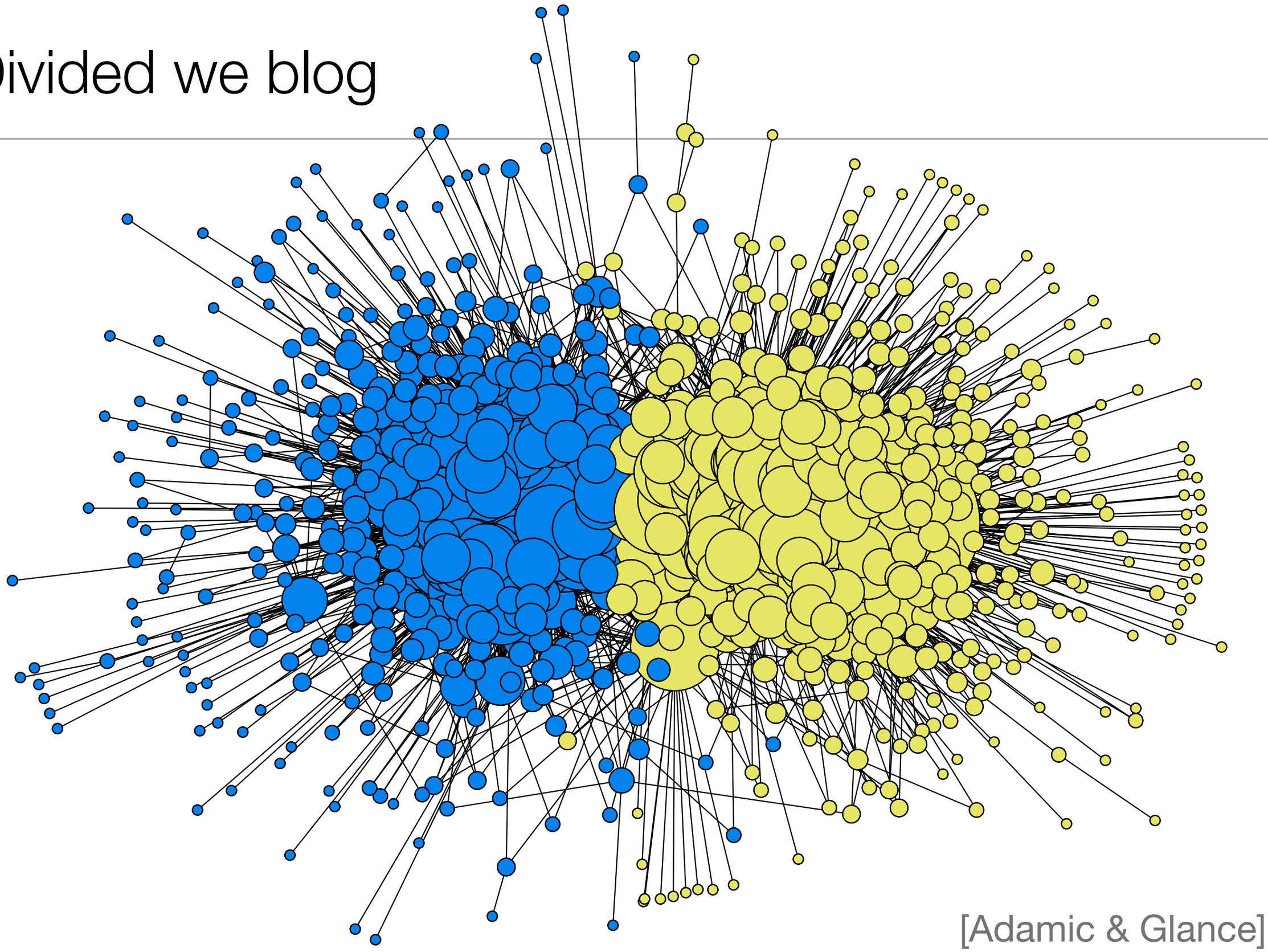
which local optimum is the true one?

can we find it efficiently? can we find it at all, given the posterior distribution?

let’s look at a classic problem in social networks…

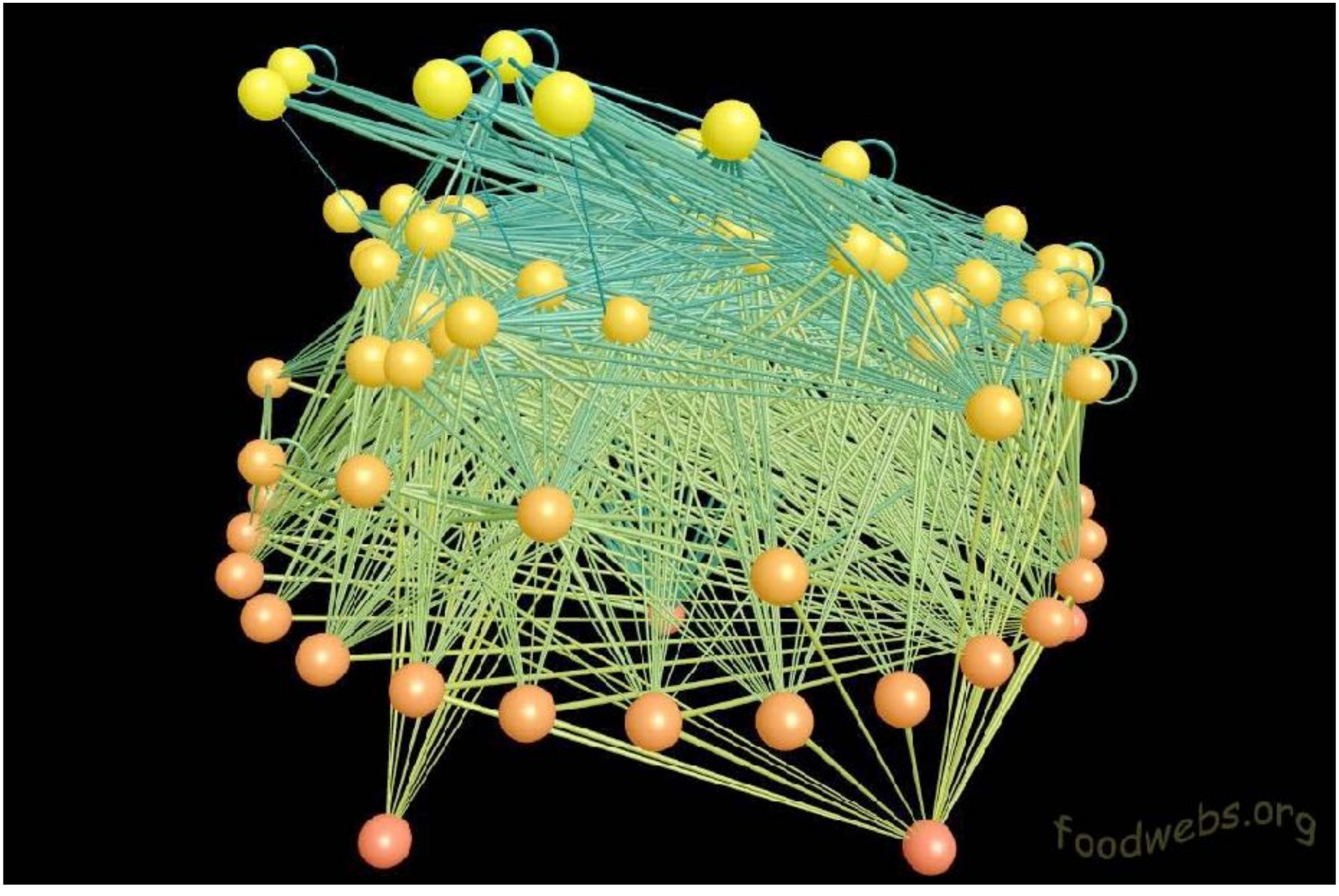


Divided we blog

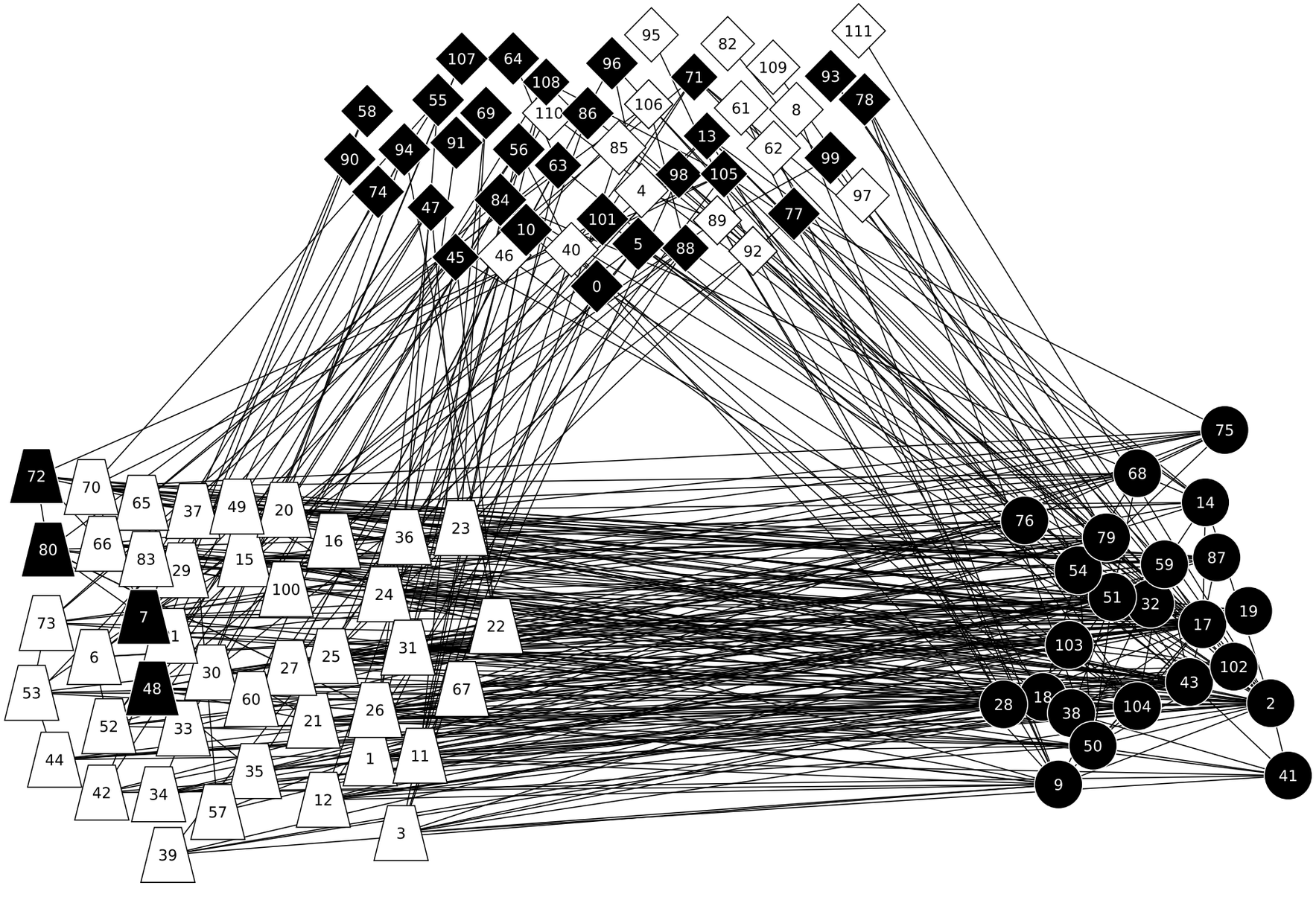


[Adamic & Glance]

Who eats whom



I record that I was born on a Friday



The stochastic block model

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

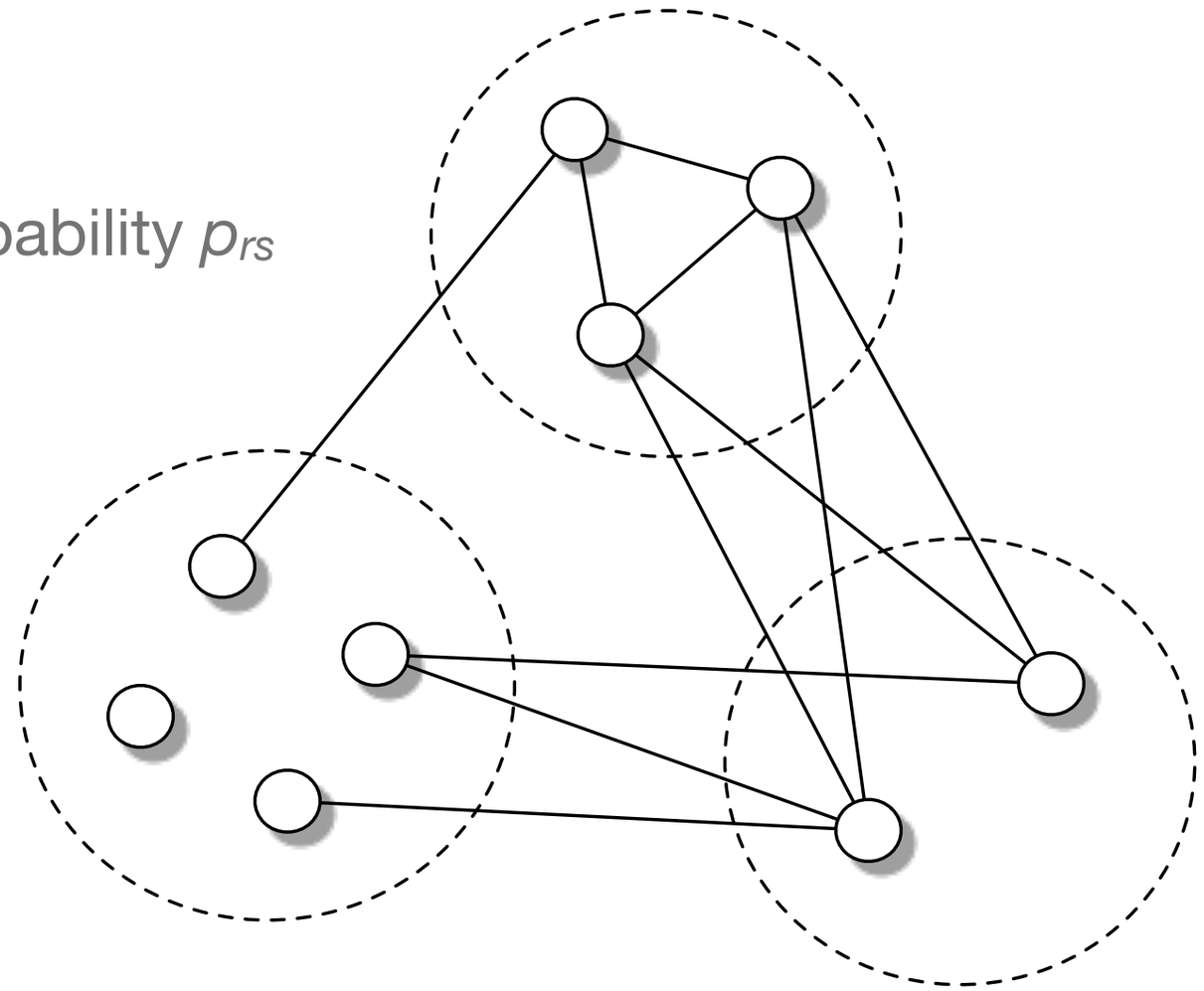
$k \times k$ matrix p of connection probabilities

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

sparse: $p = O(1/n)$

popular special case:

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



ferromagnetic (assortative, homophilic) if $c_{\text{in}} > c_{\text{out}}$

Likelihood and energy

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

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

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

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

like Ising model, but with weak antiferromagnetic interactions on non-edges

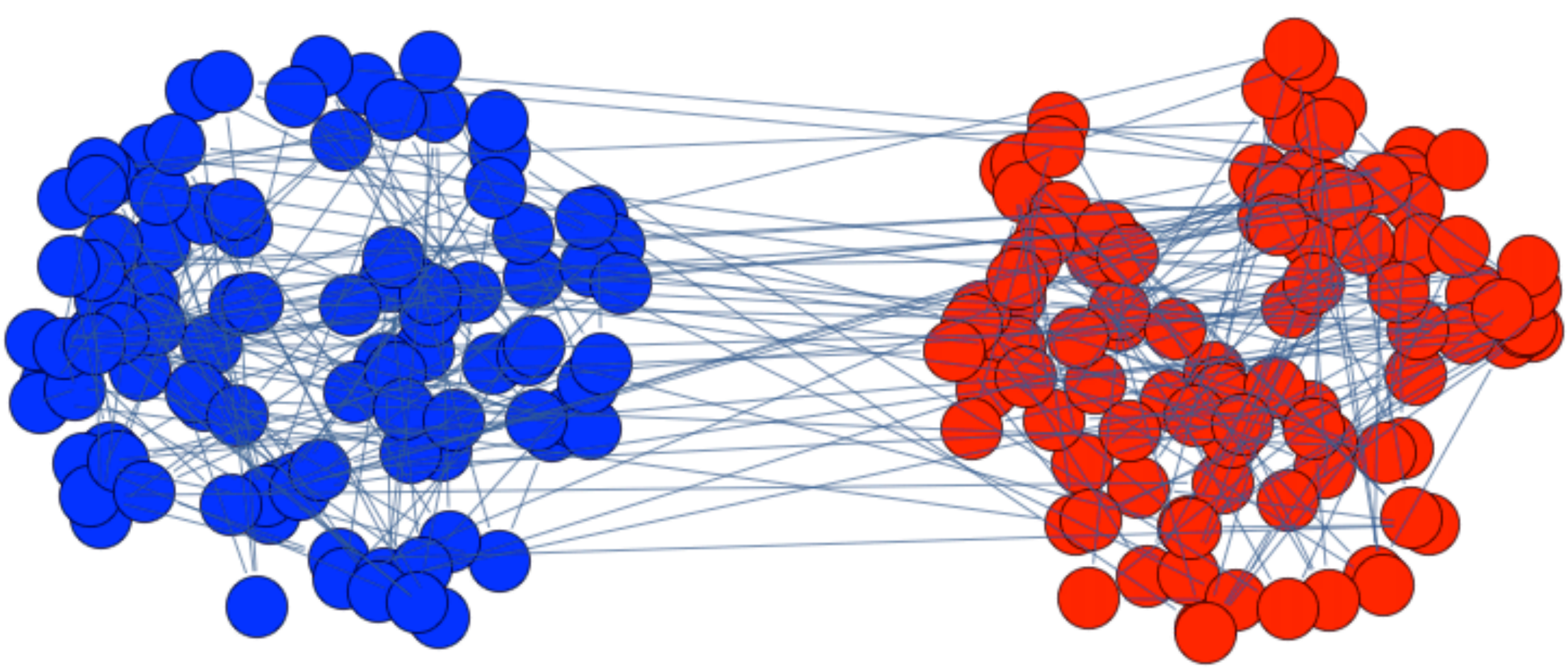
what can we learn from the “physics” of the block model?

Ground states vs. the landscape

the most likely labeling (MLE, MAP) is the ground state

even random 3-regular graphs have labelings with only 11% edges crossing [Zdeborová & Boettcher] — many of them, which don't agree!

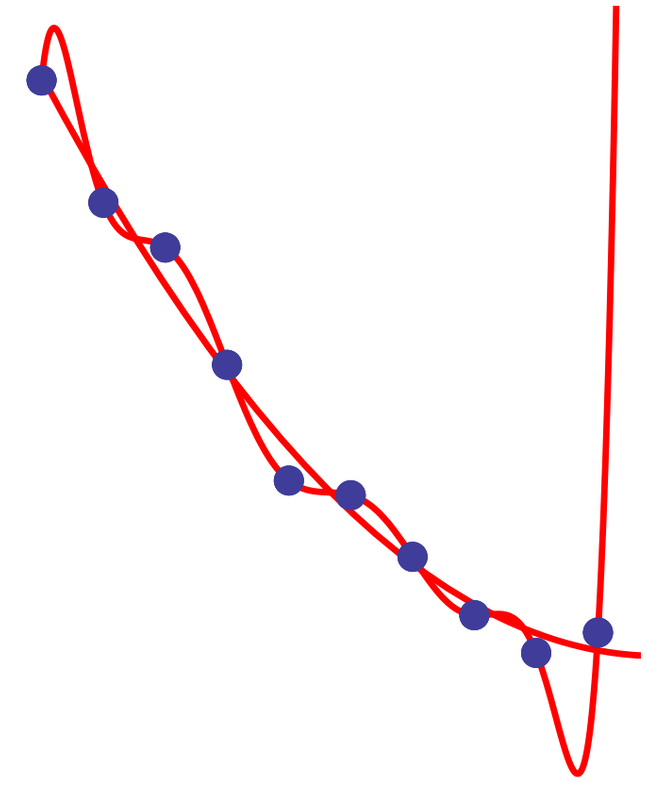
we need to understand the entire landscape, not just the optimum



Overfitting

we, and our algorithms, are prone to false positives

fitting the data with fancy models is tempting…



but often we're really fitting the noise, not the underlying process

we want to understand the coin, not the coin flips

Statistical significance and the energy landscape

explore the landscape of models, not just the best one

if there is real structure in the data, there is a robust optimum

but the landscape can be “glassy”: many local optima with nothing in common

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

instead, sample from the entire landscape, and look for consensus

Information in the block model: the effect of a link

k equal groups, $p = \frac{1}{n}\begin{pmatrix} c_{\text{in}} & \cdots & c_{\text{out}} \\ \vdots & \ddots & \\ c_{\text{out}} & & c_{\text{in}} \end{pmatrix}$: average degree $c = \frac{c_{\text{in}} + (k-1)c_{\text{out}}}{k}$

if there is a link $i \to j$, the probability distribution of t_j is related to that of t_i by a transition matrix

$$\frac{1}{kc}\begin{pmatrix} c_{\text{in}} & \cdots & c_{\text{out}} \\ \vdots & \ddots & \\ c_{\text{out}} & & c_{\text{in}} \end{pmatrix} = \lambda \mathbb{1} + (1-\lambda)\begin{pmatrix} 1/k & \cdots & 1/k \\ \vdots & \ddots & \\ 1/k & & 1/k \end{pmatrix}$$

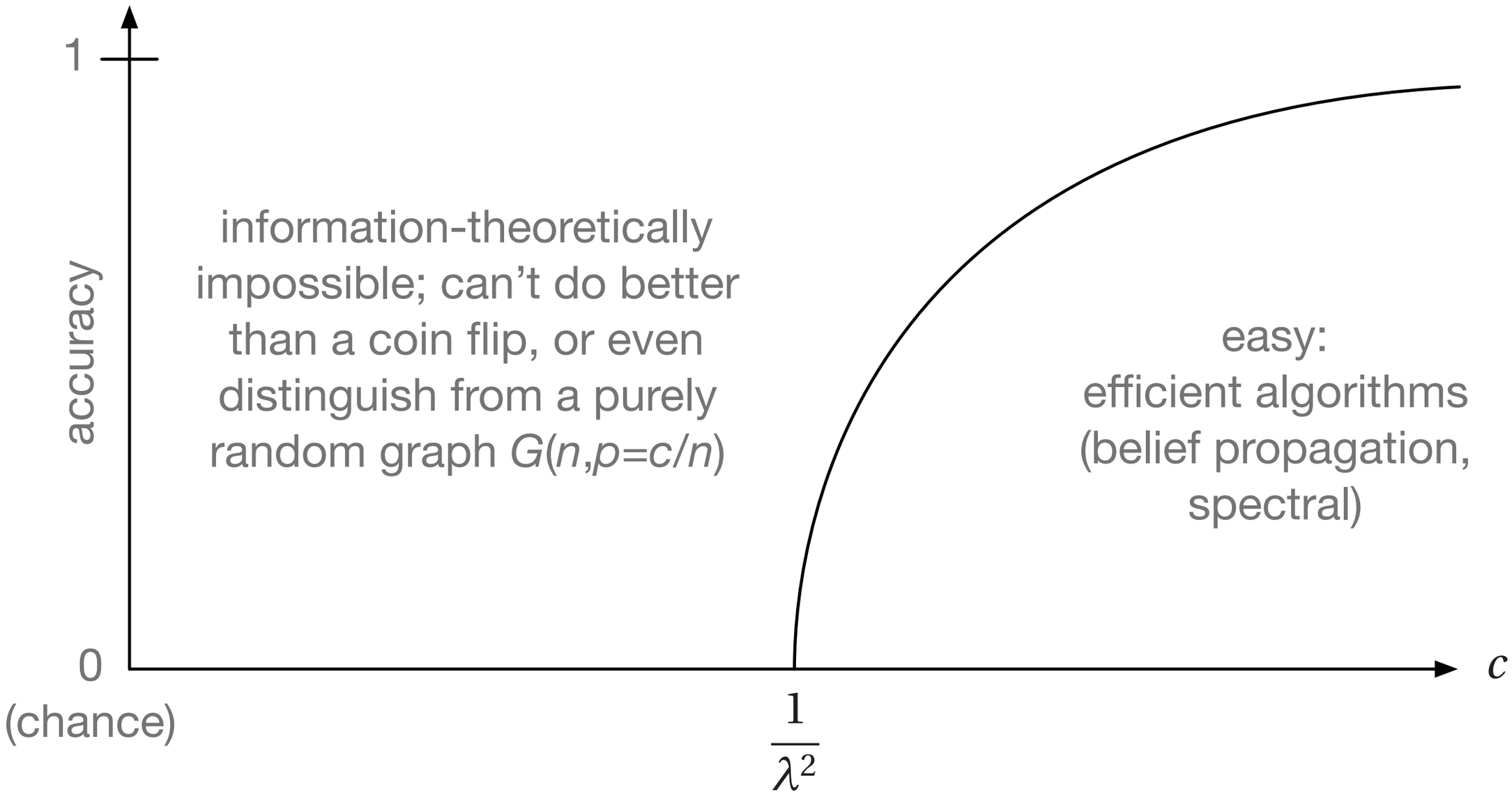
where $\lambda = \frac{c_{\text{in}} - c_{\text{out}}}{kc}$

with probability λ, copy from *i* to *j;* with probability $1-\lambda$, set *j*'s type randomly

if λ is fixed, community detection gets easier as c increases…

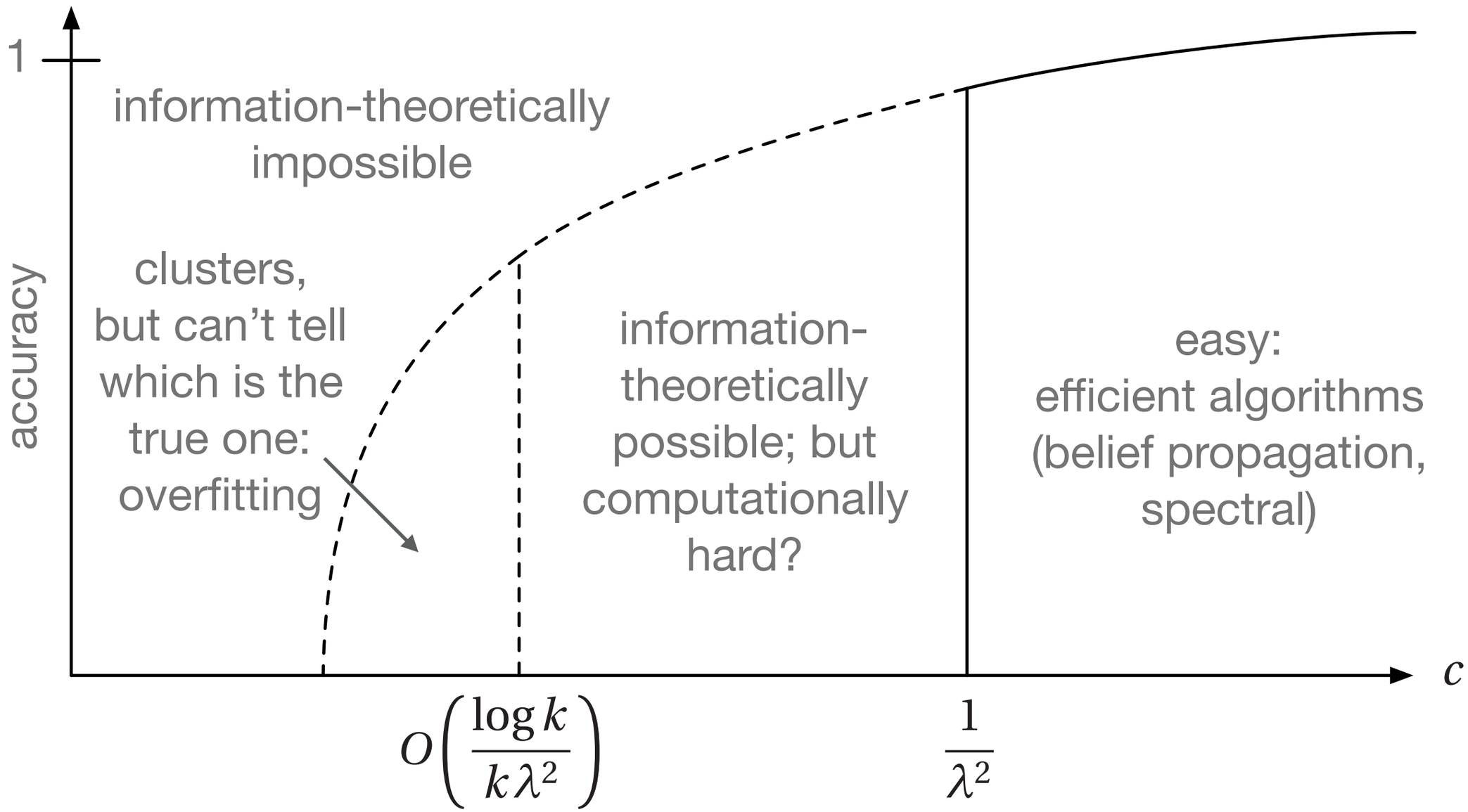
Detectability thresholds

For two groups of equal size [DKMZ, MNS, M, KMMNSSZ, BLM]:



Detectability thresholds

For k≥4 groups [DKMZ, KMMNSSZ, BLM, BMNN, AS]:



Markov Chain Monte Carlo

want to sample from the Gibbs/Boltzmann/posterior distribution $P(t|G)$

computing $P(t|G)$ is hard, but it's proportional to $P(G|t)$, a product of local terms

can compute ratios between $P(t|G)$ and $P(t' \mid G)$ if t and t' differ at one node

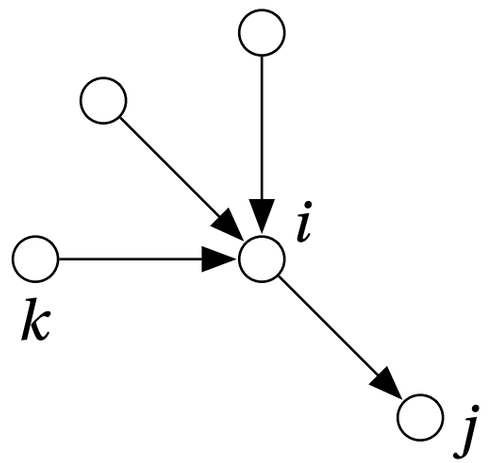
heat-bath dynamics: choose a random node v, fix labels of all other nodes, update v's label according to its marginal distribution

can also use population annealing, parallel tempering, etc.

but to compute marginals we need many independent samples...

...and if we want free energies, we need many temperatures

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



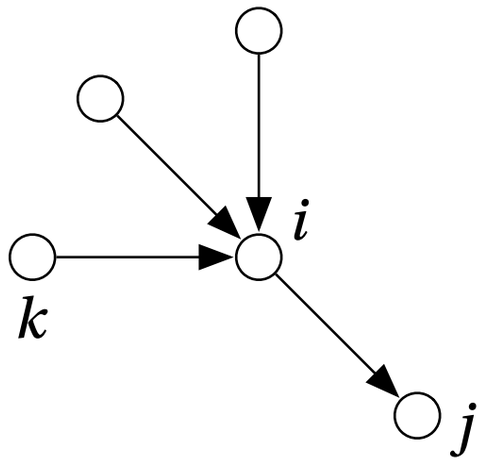
each node *i* sends a “message” to each of its neighbors *j*, giving *i*’s marginal distribution based on its other neighbors *k*

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



conditional independence

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

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

each update takes O(*n*+*m*) time, for constant *k*

A phase transition: detectable to undetectable communities

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

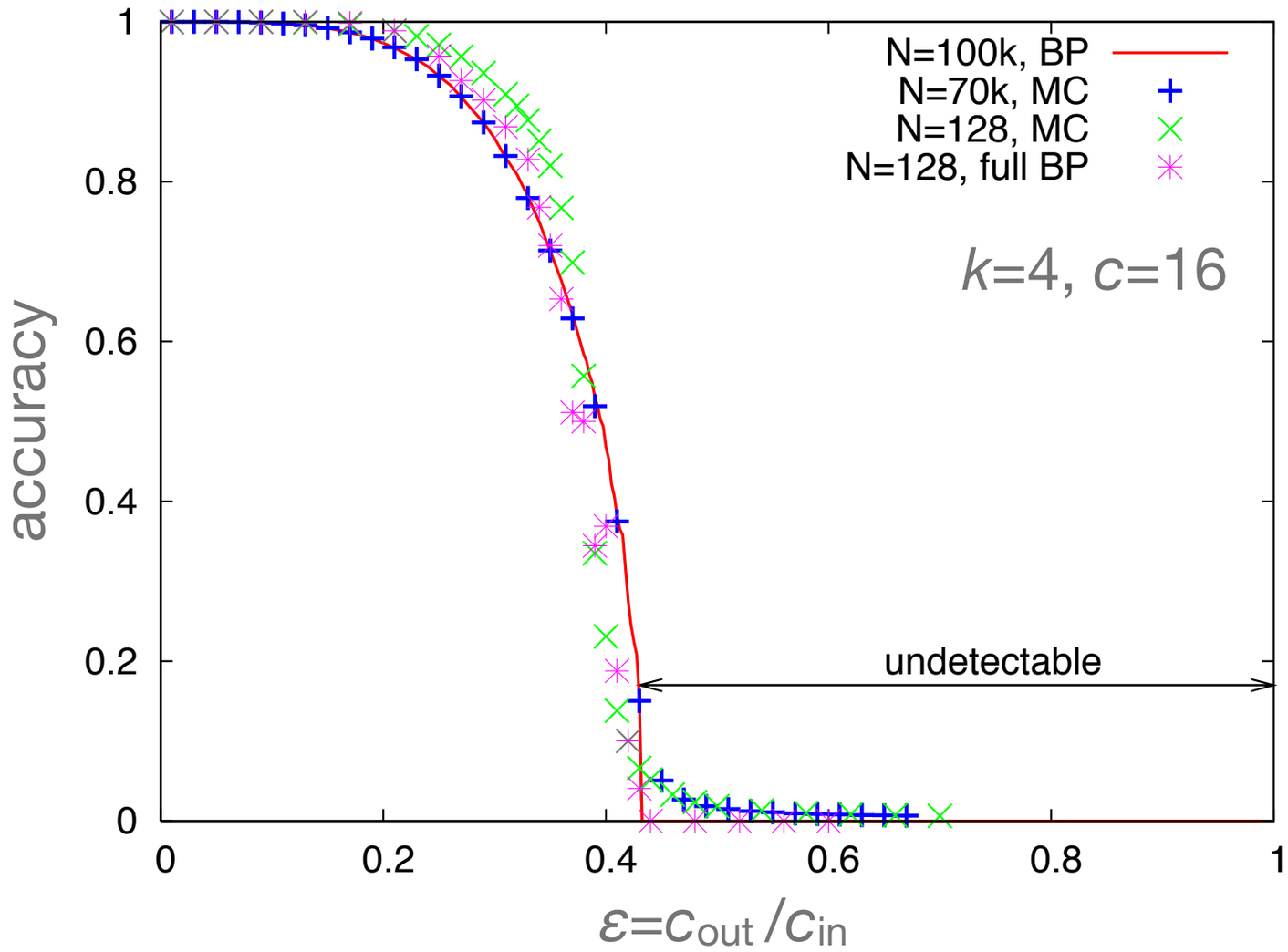
there is a regime where it can't, and no algorithm can!

for 2 groups, the threshold is at

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

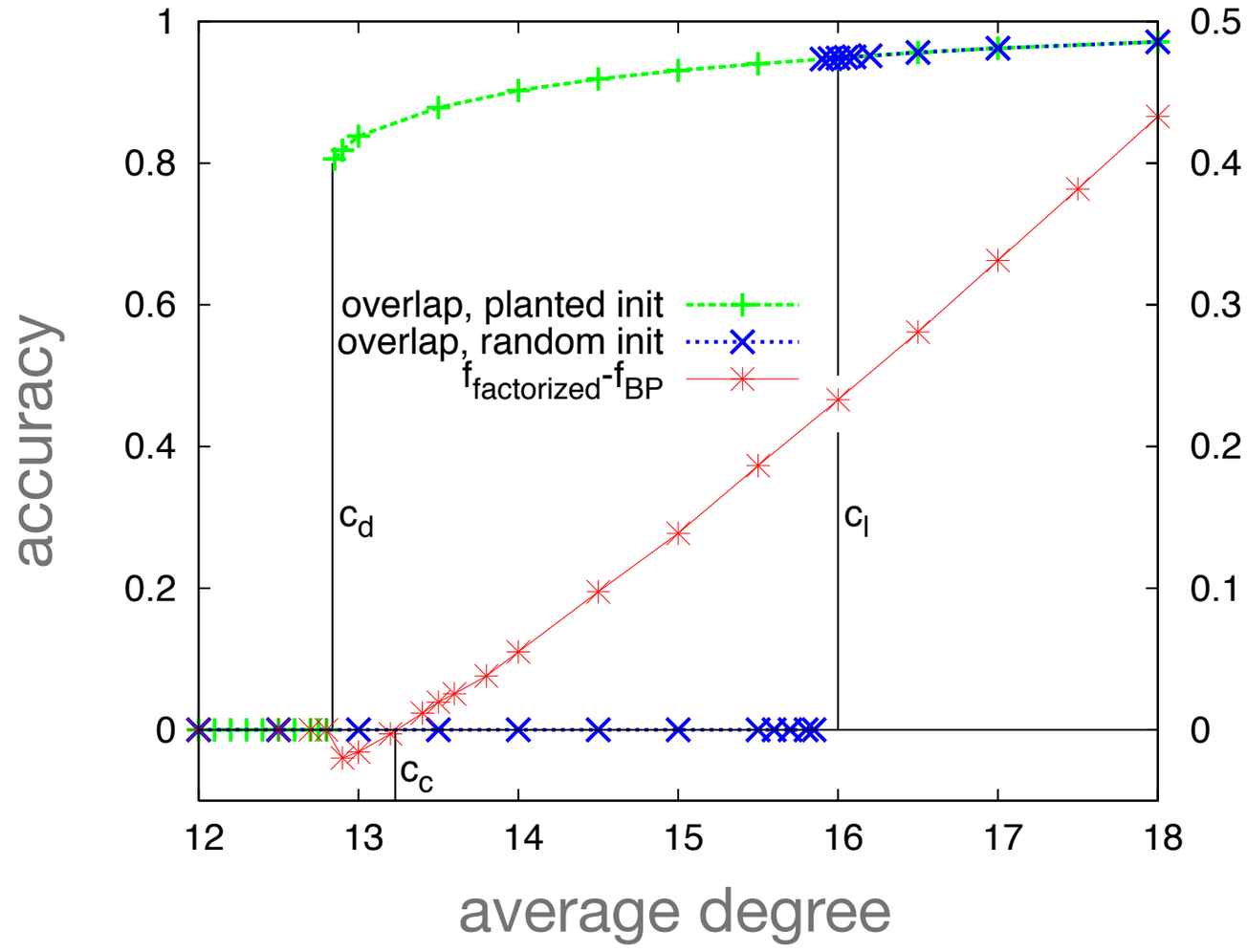
there is a fixed point where all nodes have uniform marginals...

at the transition, it becomes stable



conjectured by [Decelle, Krzakala, Moore, Zdeborová, '11]
proved by [Mossel, Neeman, Sly, `13; Massoulié '13]

Another regime: detectable but hard

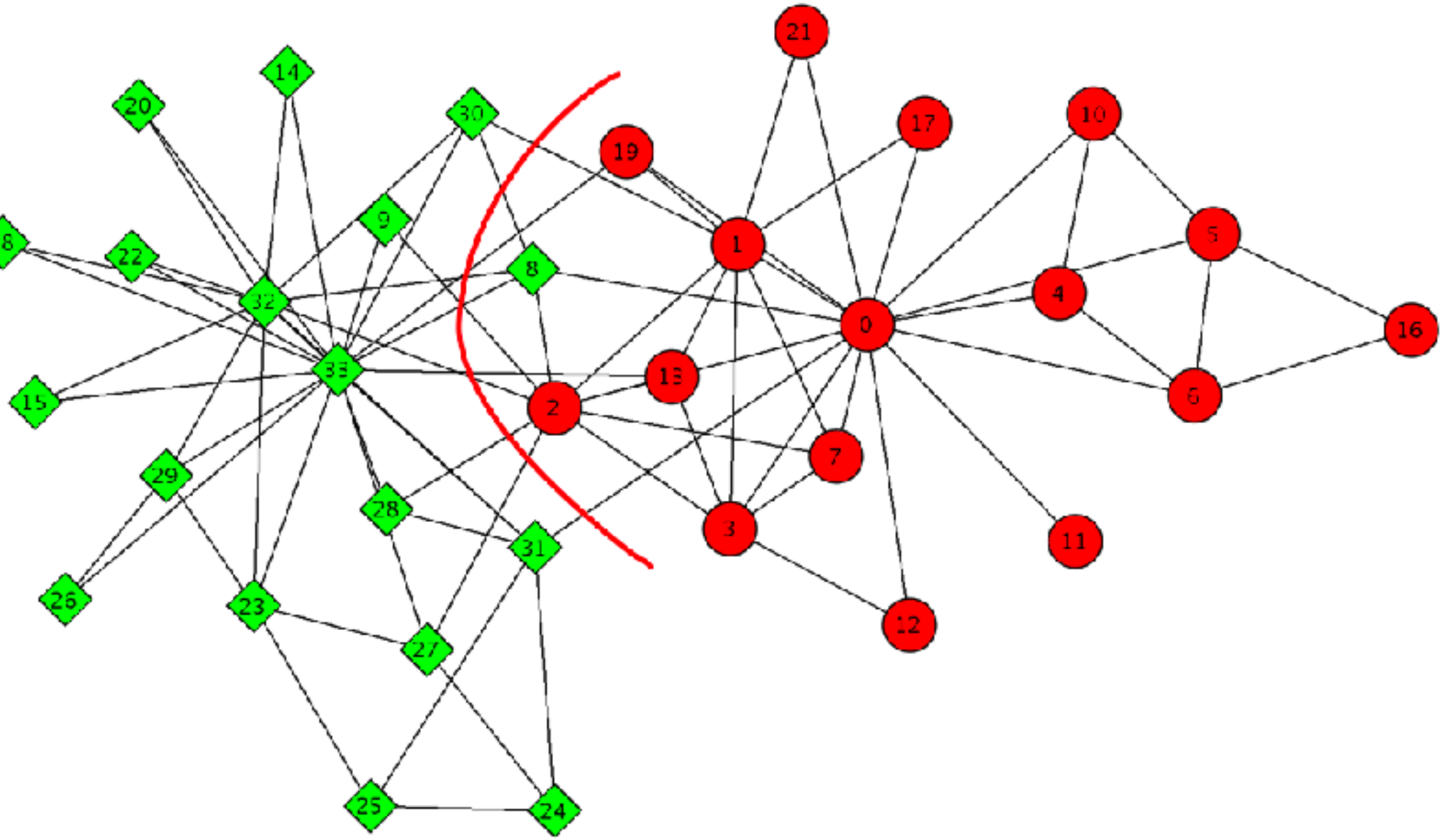


BP has two fixed points, but the accurate one has a small basin of attraction

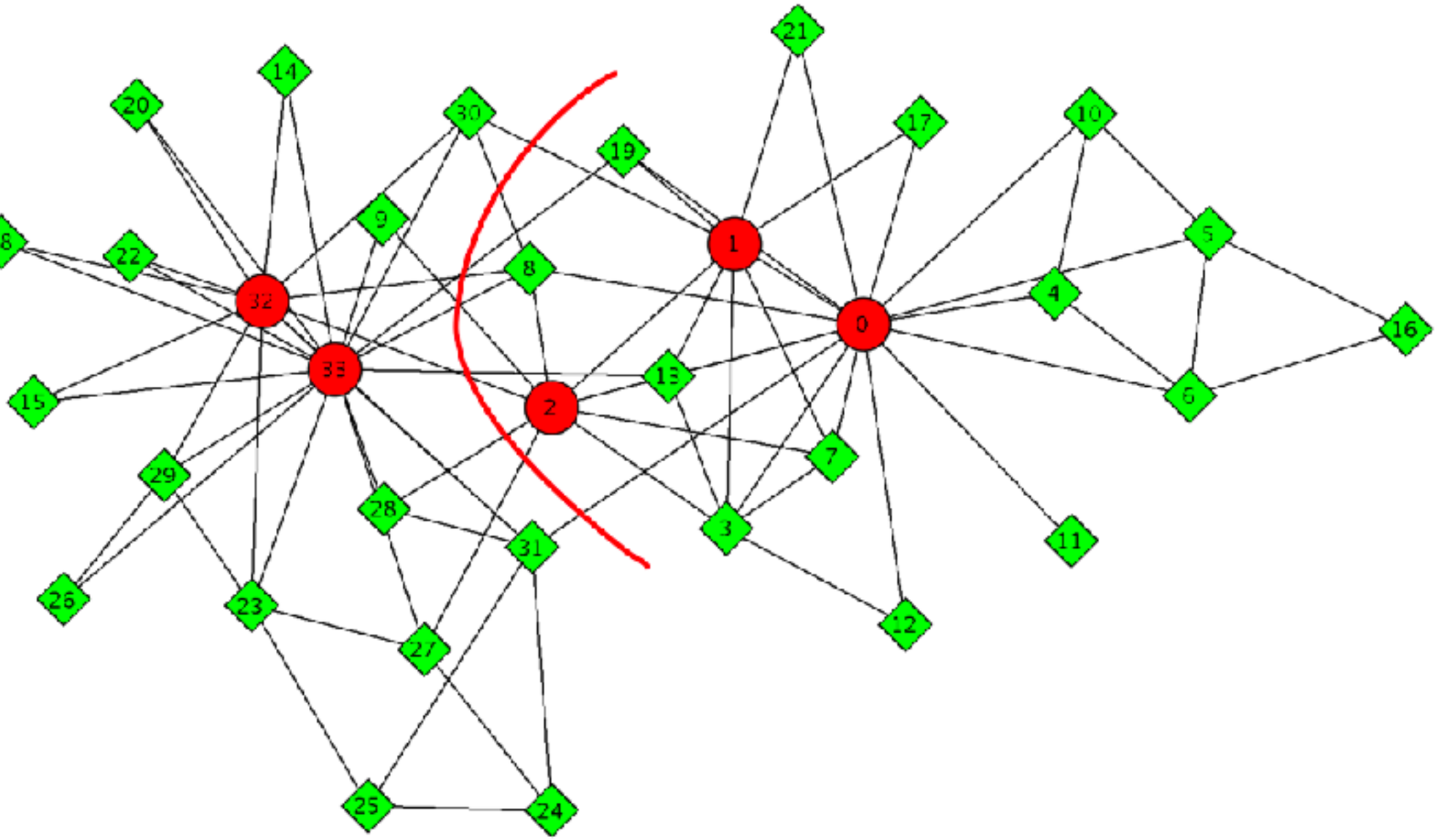
a free energy barrier between “paramagnetic” and “ferromagnetic” phases

detection is information-theoretically possible [Banks, Moore, Neeman, Netrapalli, COLT `16; Abbe and Sandon] but we believe it’s computationally hard

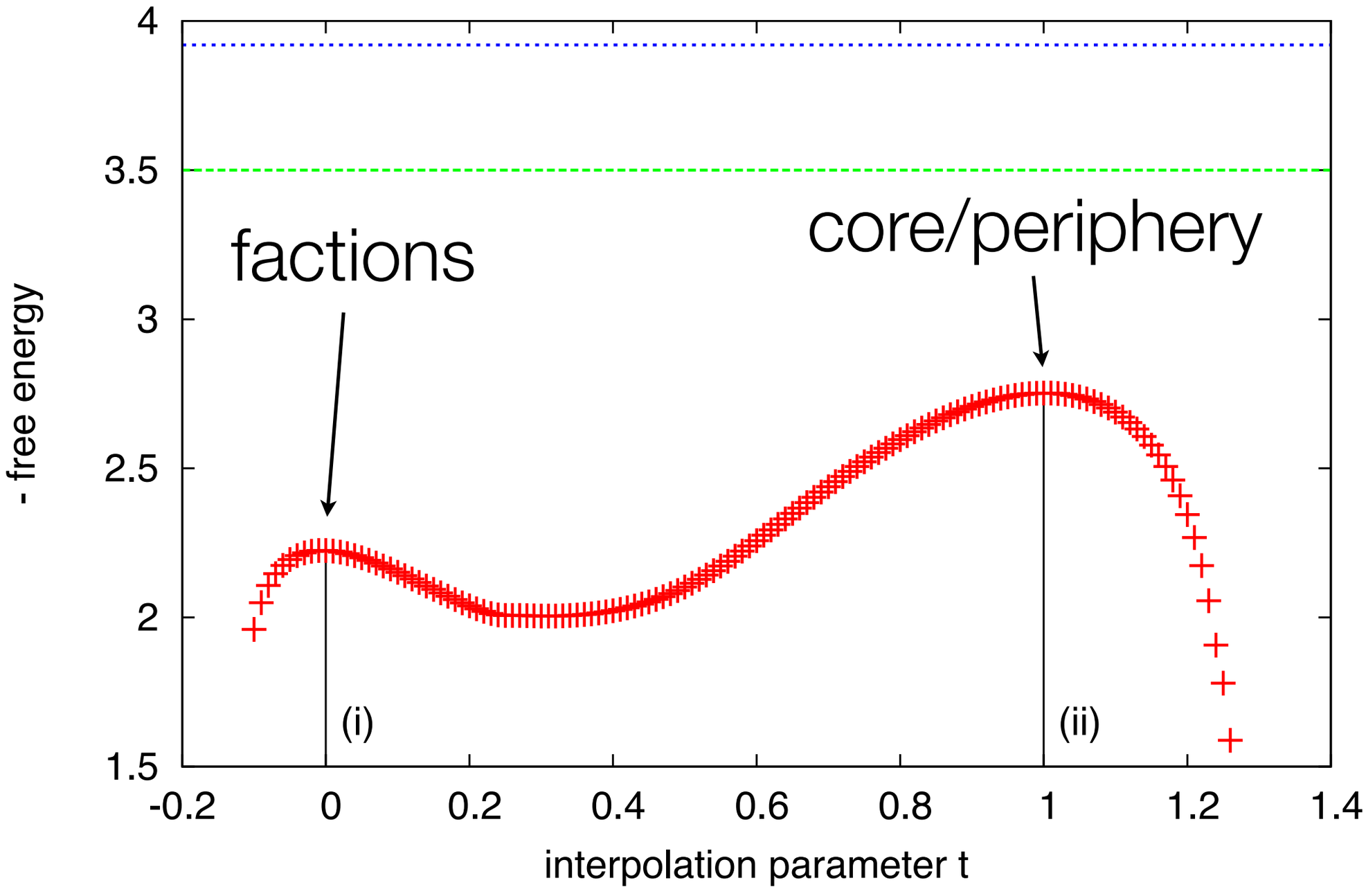
Zachary's Karate Club: Two factions



Zachary's Karate Club: Core-periphery



Two local optima in free energy



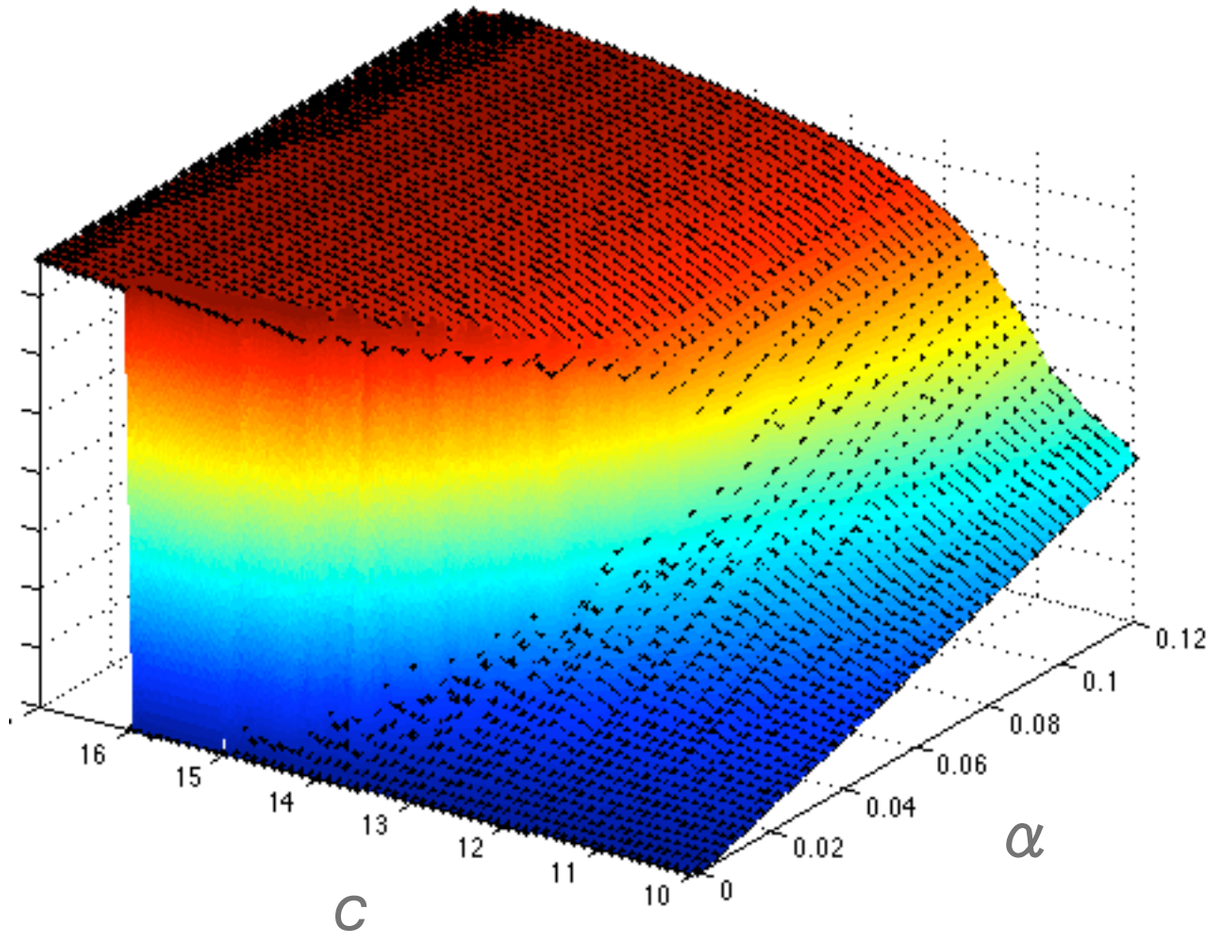
Phase transitions with metadata: what if we know some labels?

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

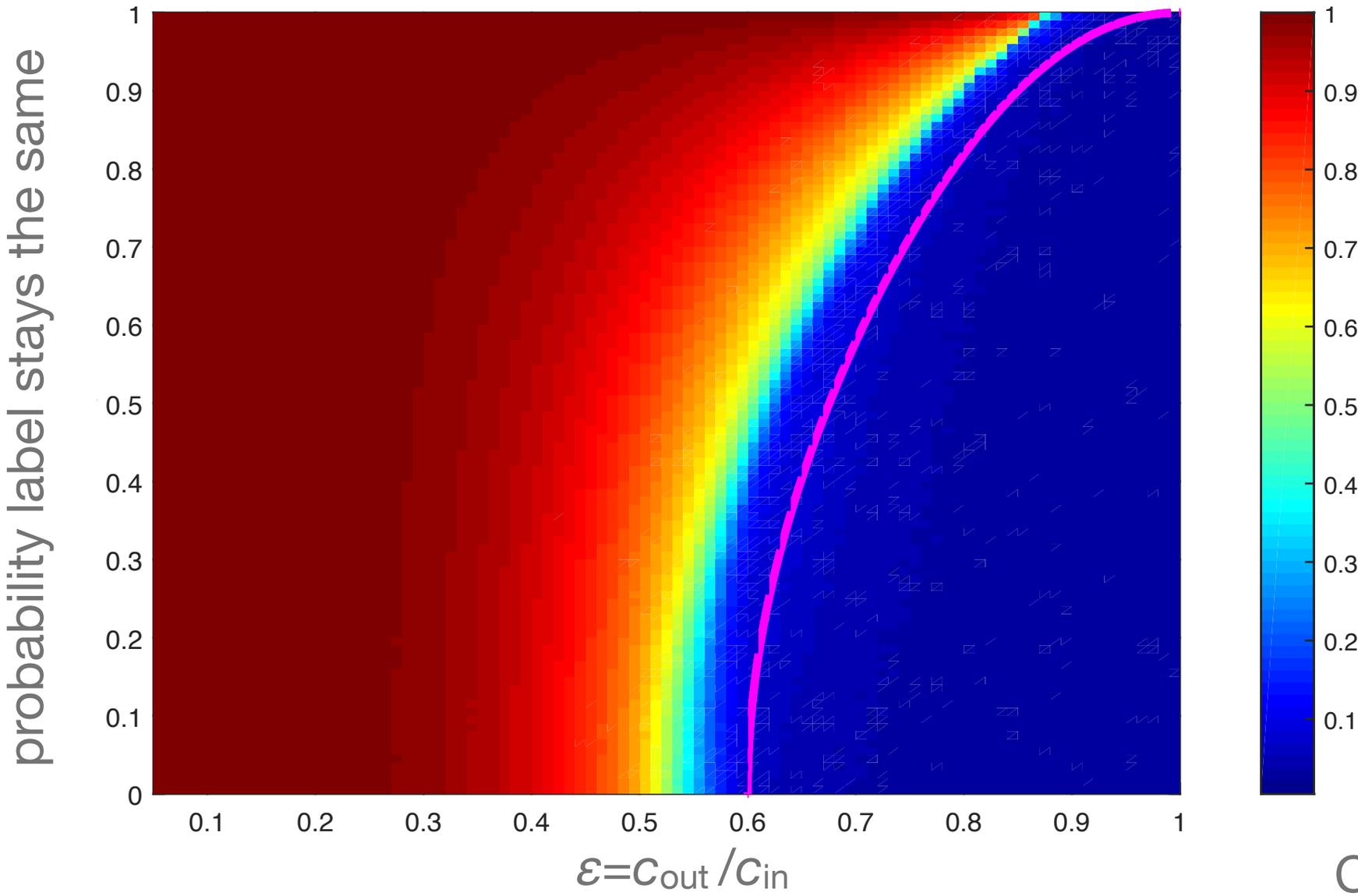


[Zhang, Moore, Zdeborová ’14]

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

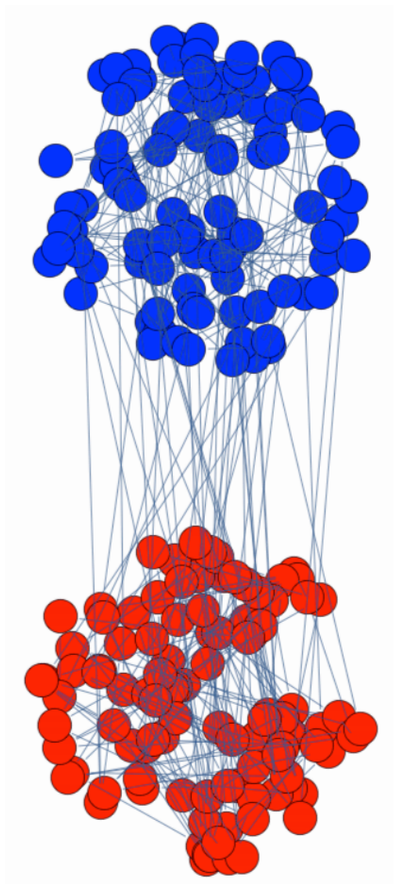


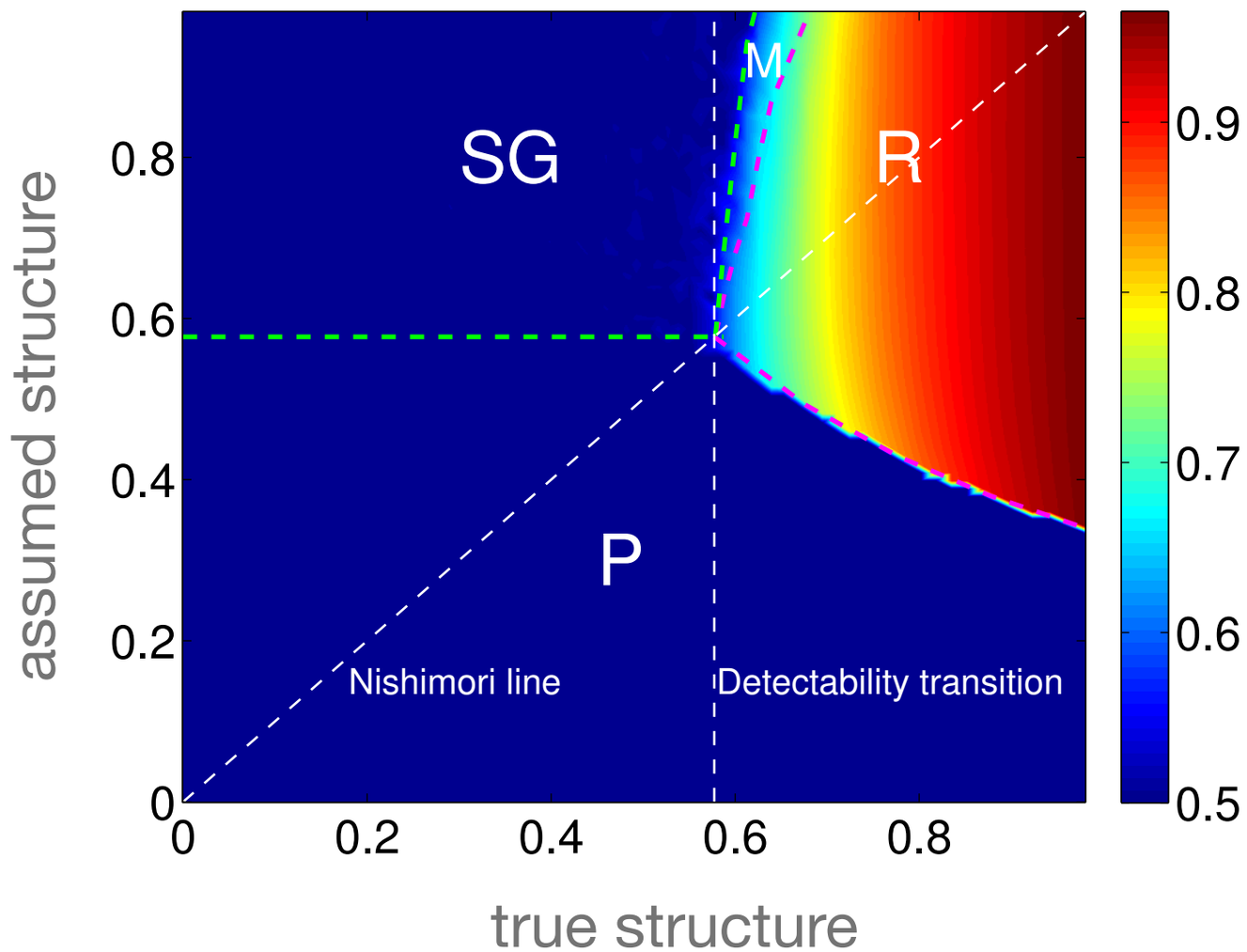
[Ghasemian, Zhang, Clauset, Moore, Peel]

What if we don’t know how strong the structure is?

lower temperature = greedier algorithm = assume stronger structure

if we get too greedy, we enter a “spin glass” where BP fails to converge





[Zhang+Moore]

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

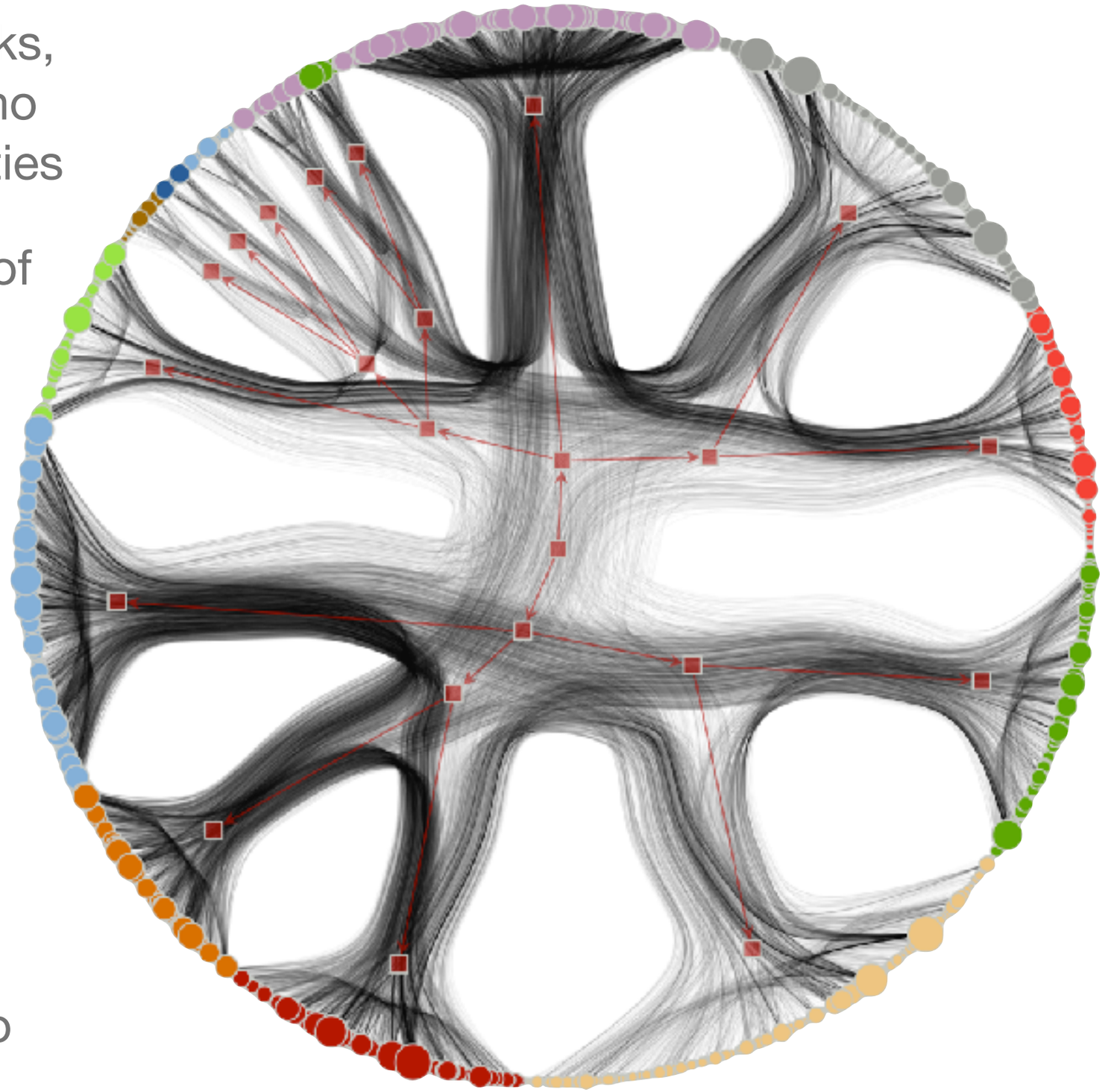
Bayes+physics: use free energy to decide if structure is statistically significant

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

Extensions to richer data, e.g. text+links

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		

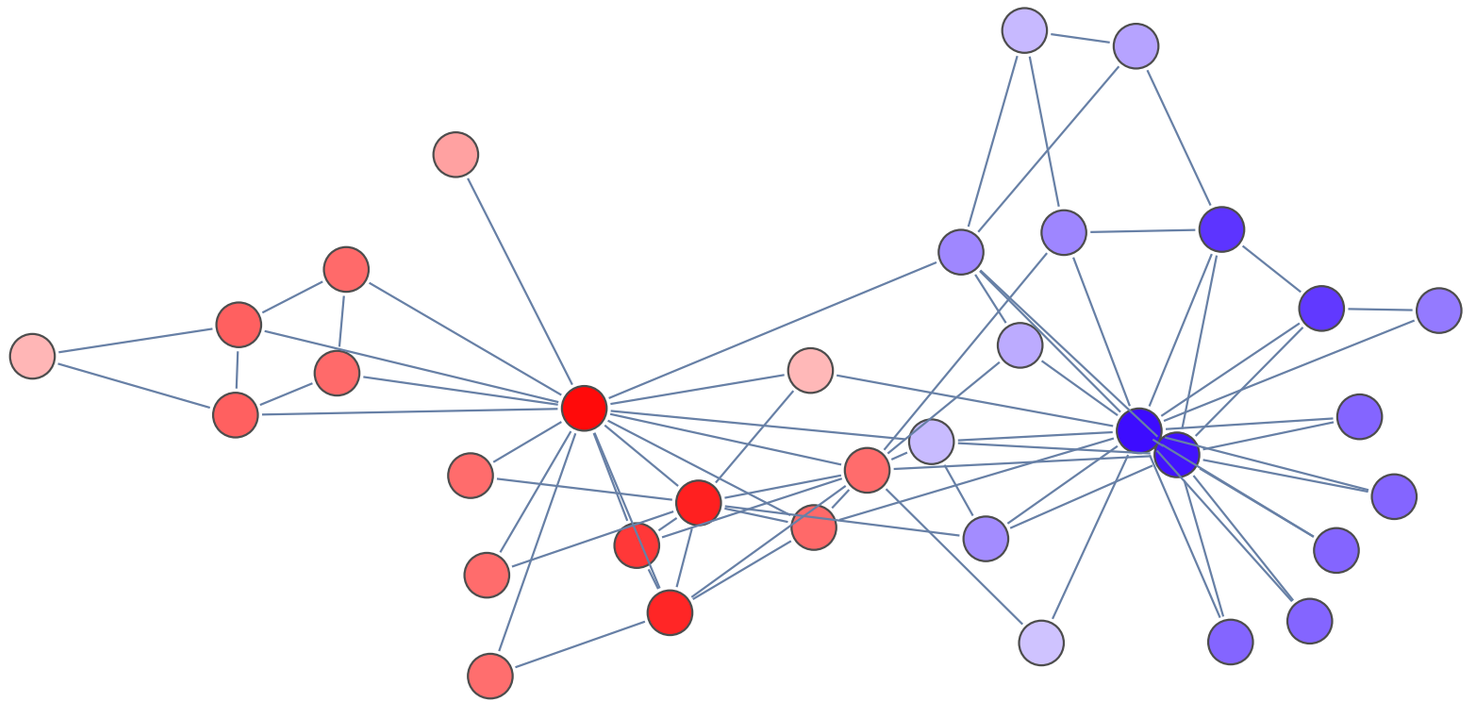
using both text and links does better than either one alone

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

Spectral clustering

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

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

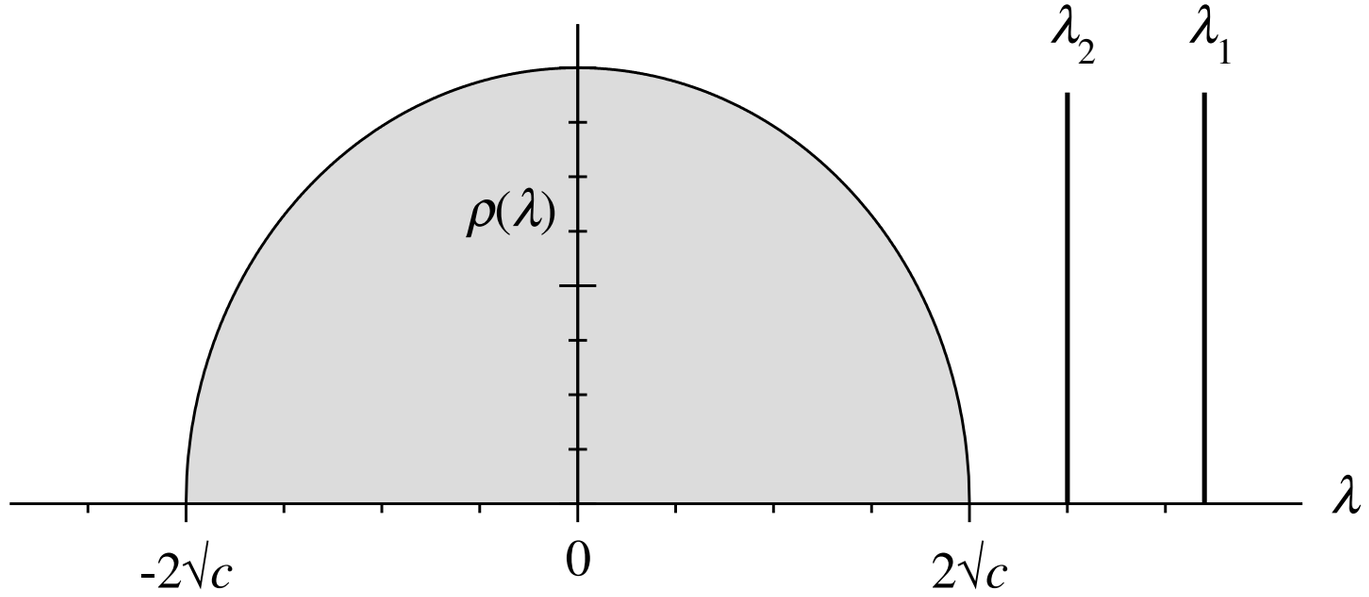


if there are *k* groups, look at the first *k* eigenvectors, and use your favorite clustering algorithm in R^k

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...

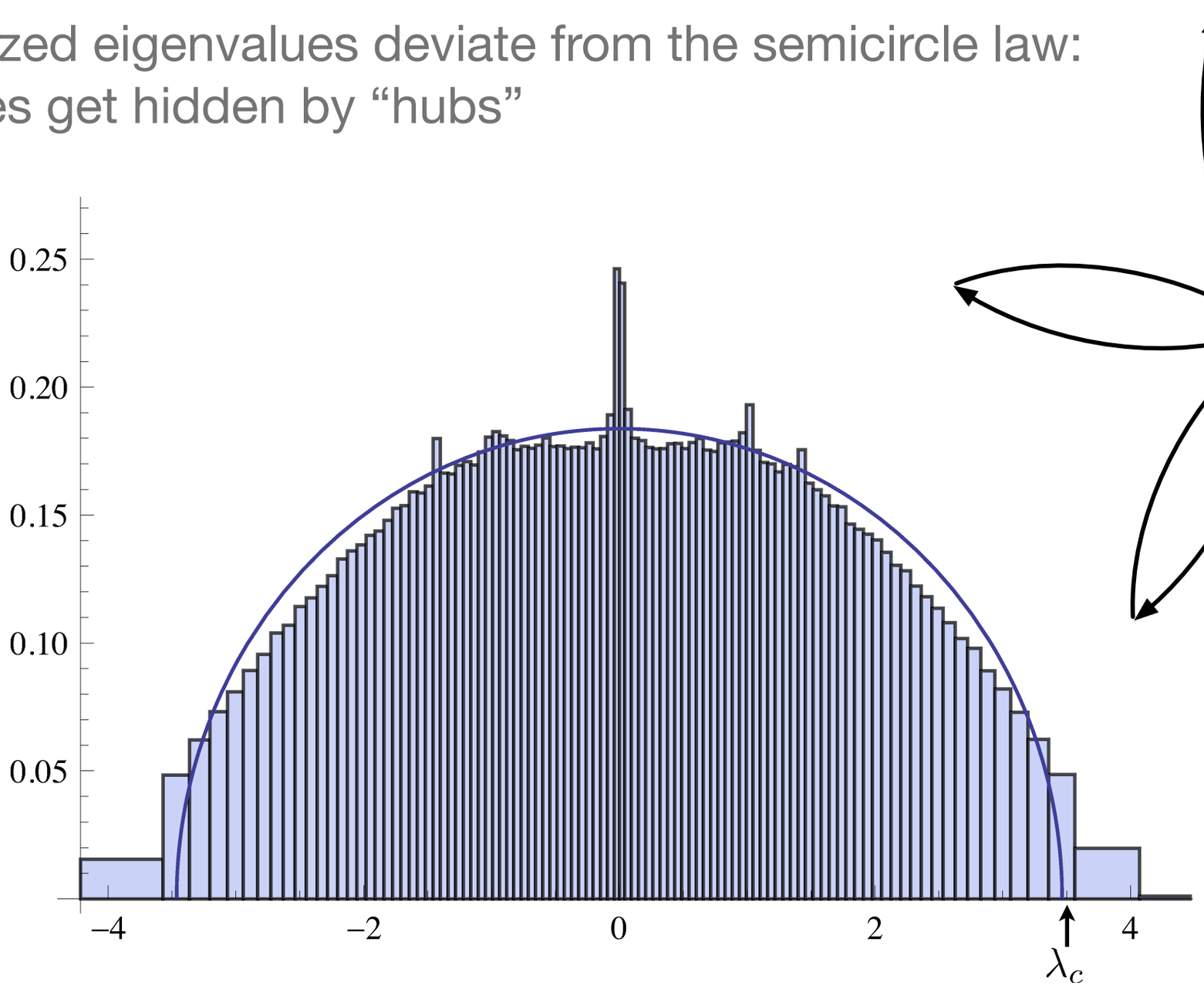
crosses at the detectability transition... if the graph is dense enough

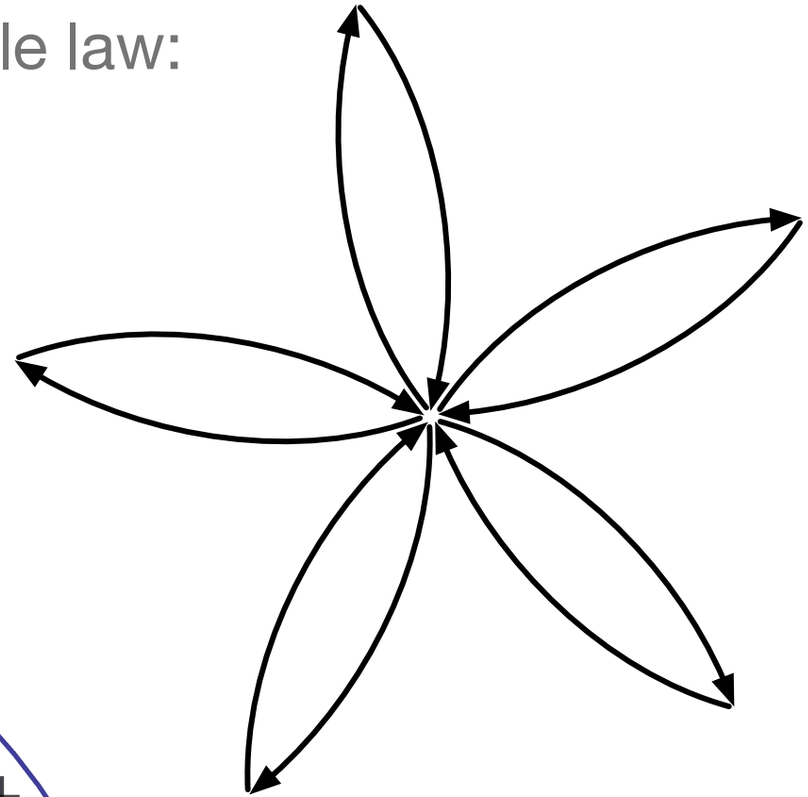
[Nadakuditi and Newman, `12]

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:
communities get hidden by “hubs”

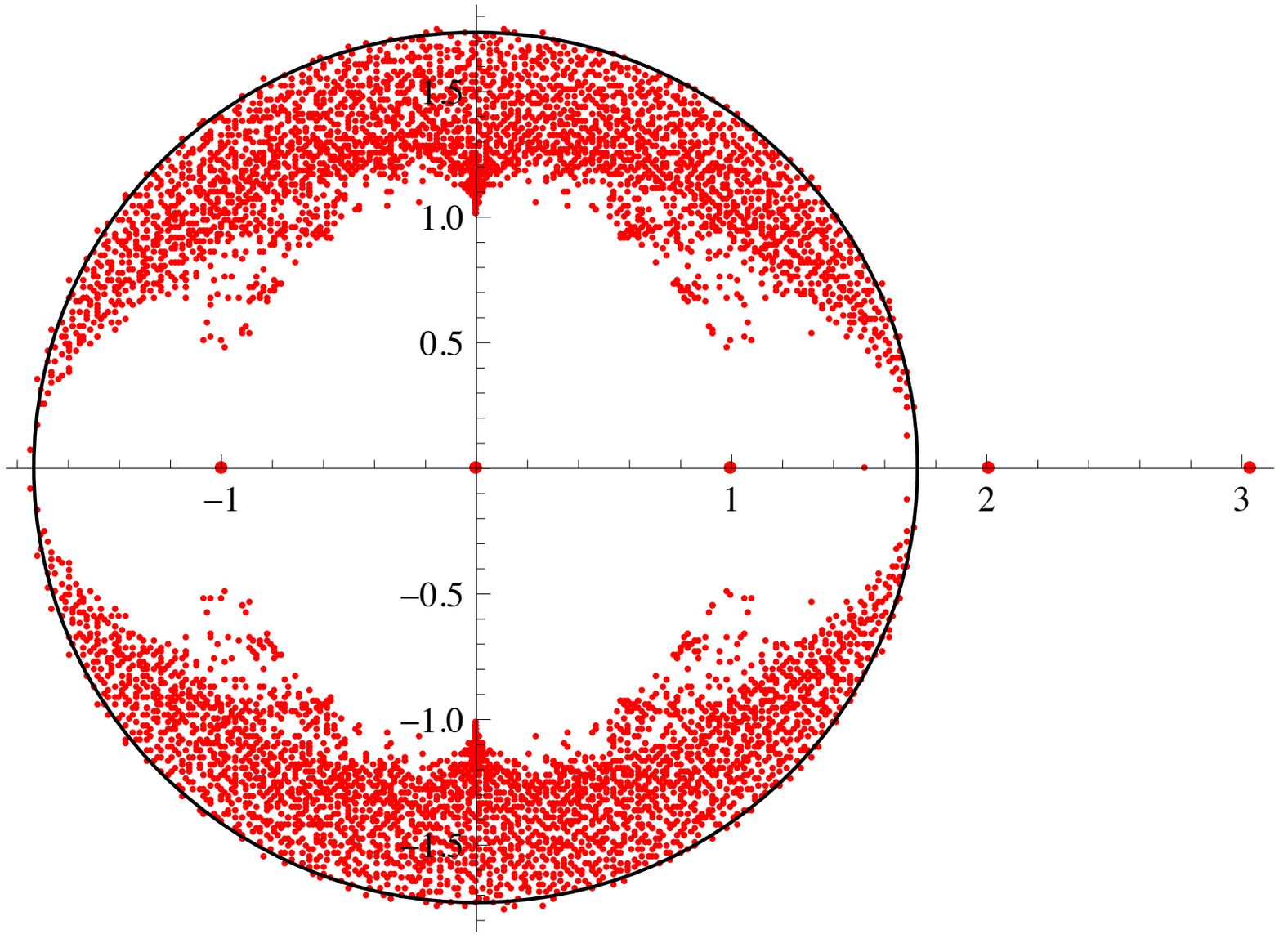


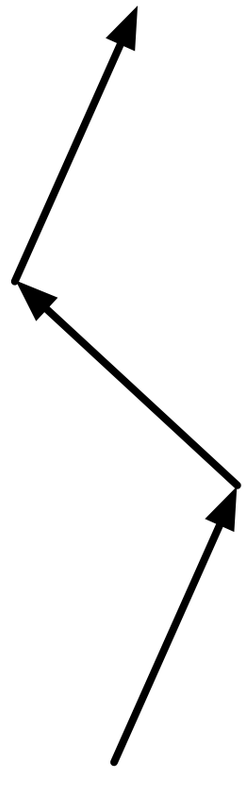


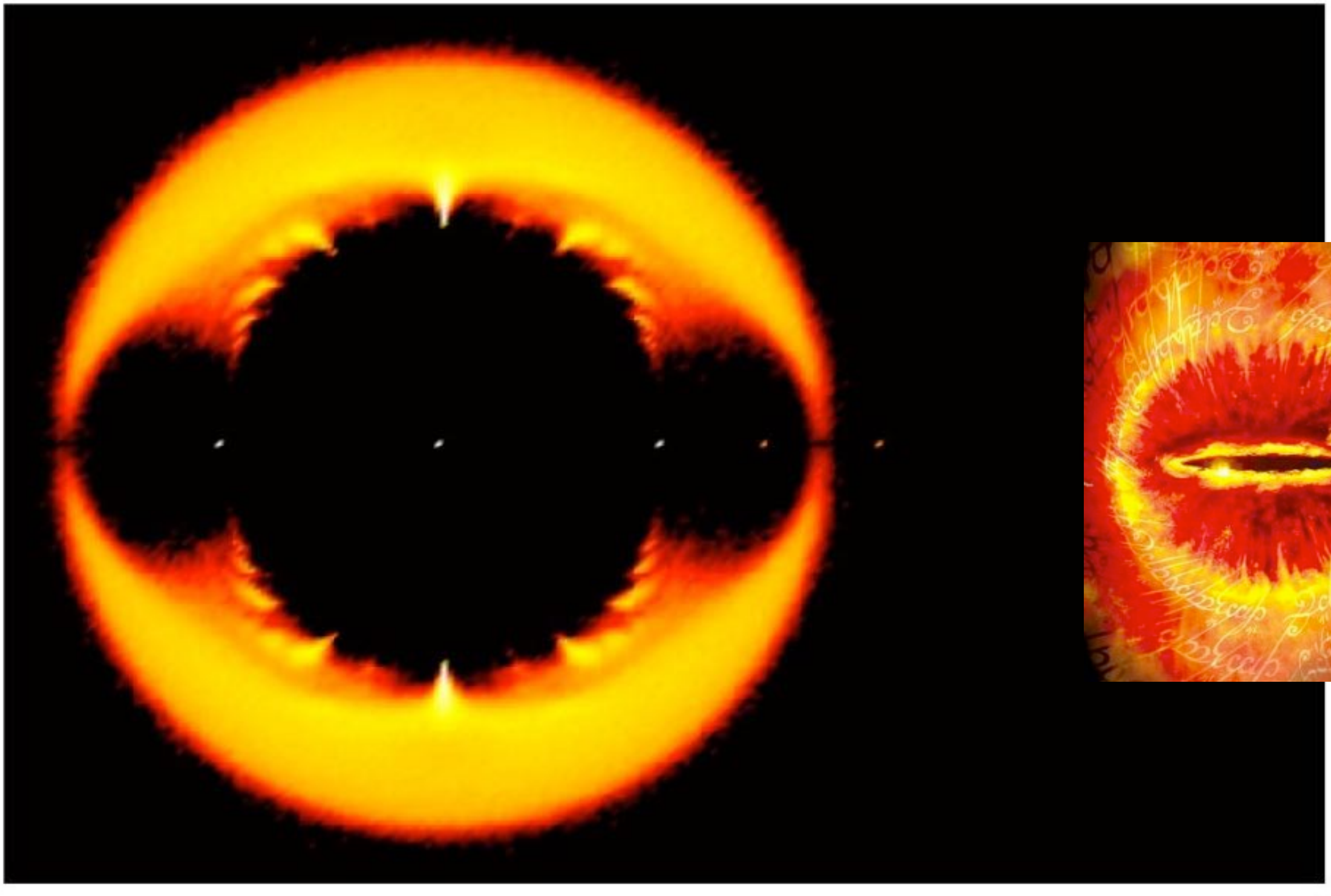
The non-backtracking operator

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

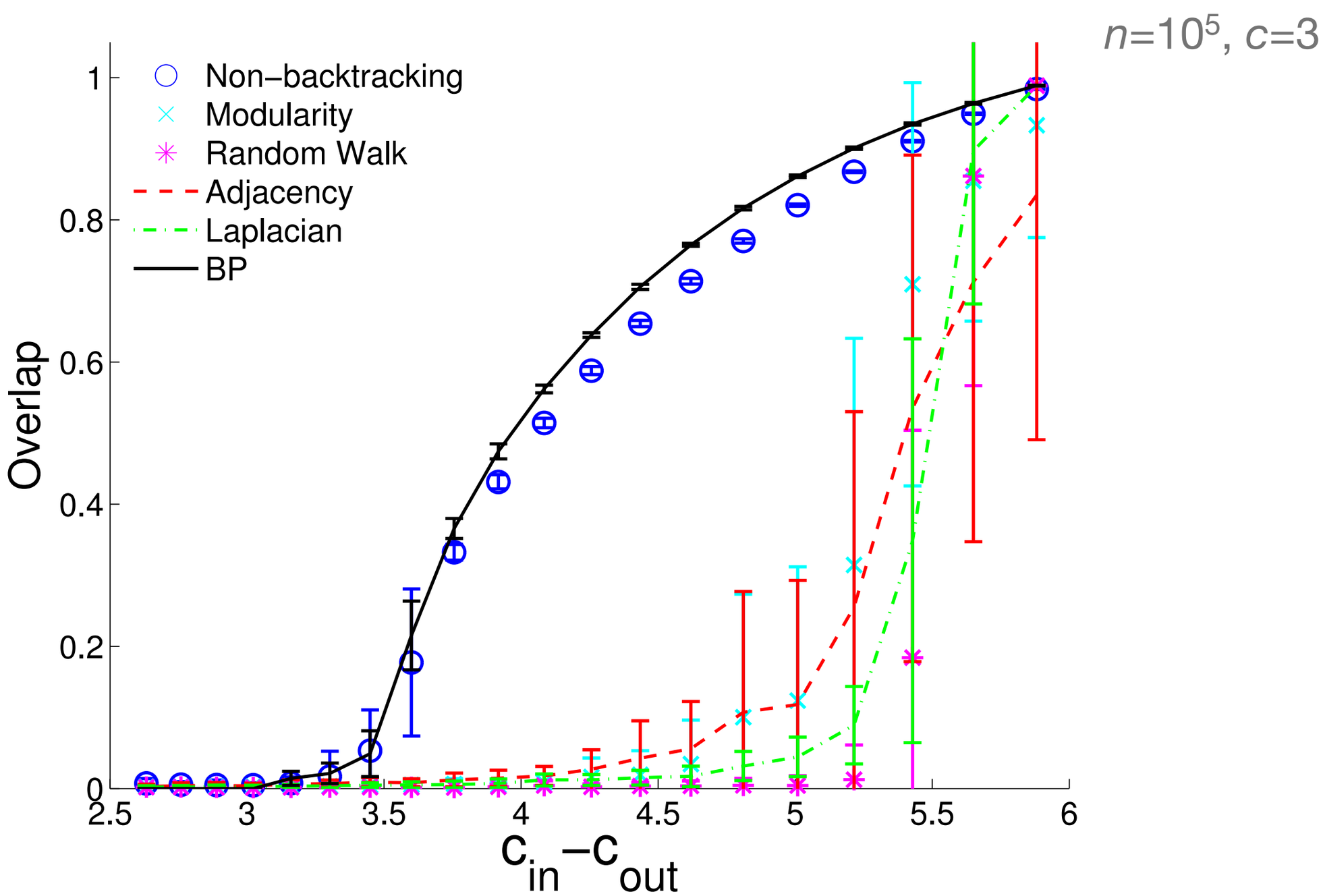
bulk of *B*'s spectrum is confined to a disk of radius $\sqrt{c}$, even in the sparse case







Comparing with standard spectral methods

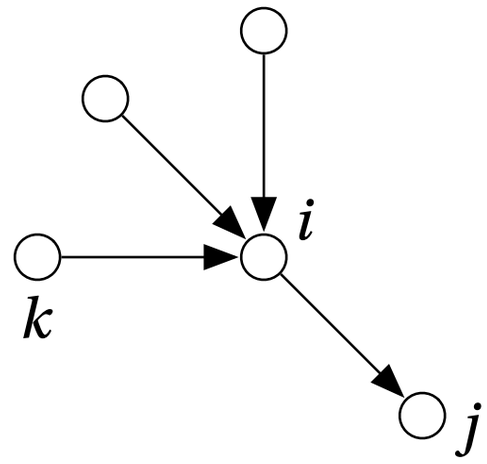


You may ask yourself,
Well, how did we get here?

---

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, *FOCS* 2016]

Clustering high-dimensional data

m points in *n*-dimensional space, where *m*=O(*n*)

k clusters with Gaussian noise

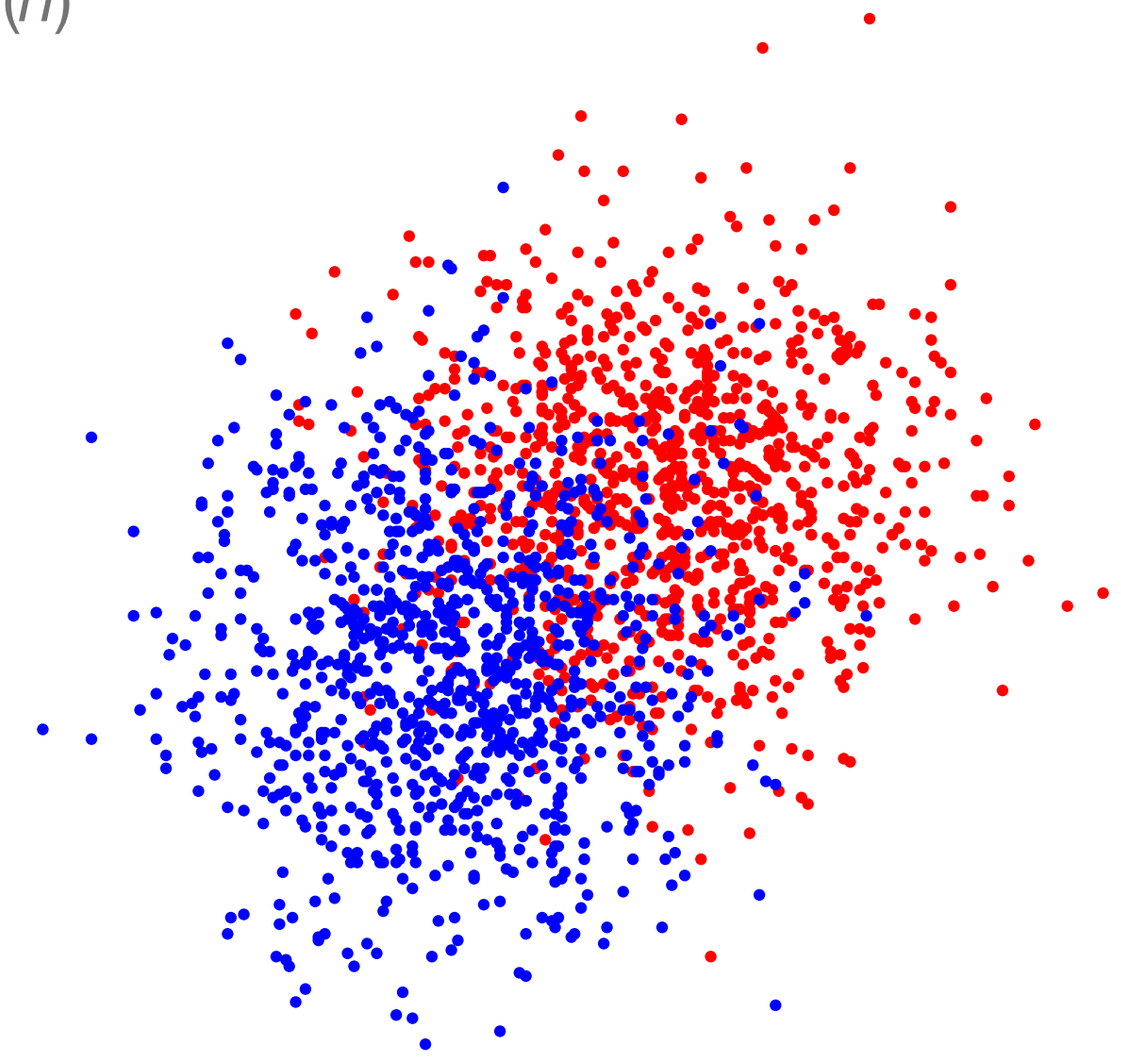
when can we...

find the cluster centers?

label the points better than chance?

tell that there are clusters, i.e., distinguish from a null model with one big cluster?

phase transitions as a function of noise vs. cluster distances, and *m*/*n*



PCA (Principal Component Analysis)

find the direction along which the points have the largest variance

first eigenvector of the matrix

$$\frac{1}{m}\sum_{i=1}^{m} x_i \otimes x_i$$

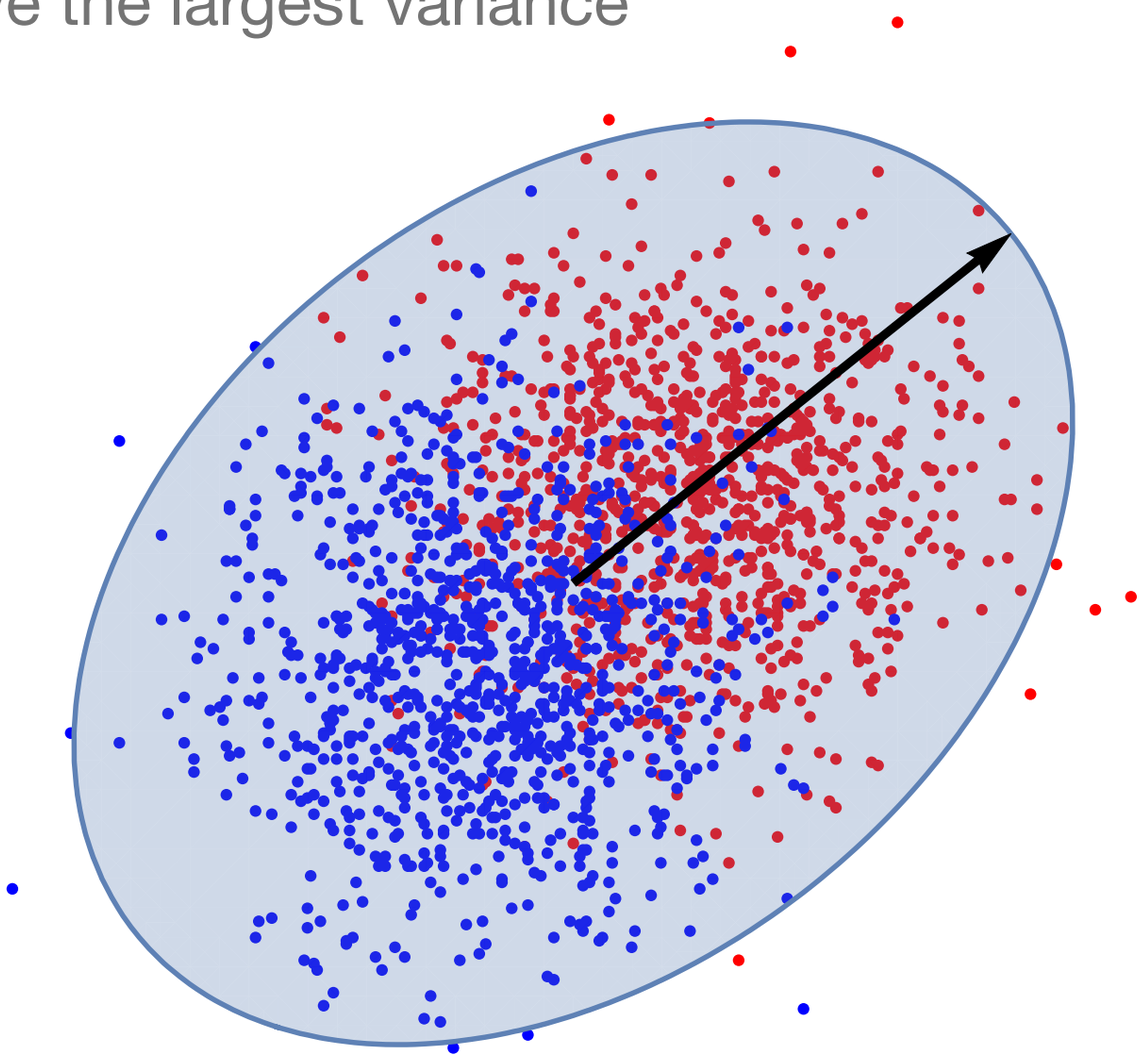
this is a *Wishart random matrix*

$$\frac{1}{m}\sum_{i=1}^{m} u_i \otimes u_i$$

plus a rank-1 perturbation

$$(v+\bar{u})\otimes(v+\bar{u})$$

when does PCA work? and how accurately?



A phase transition

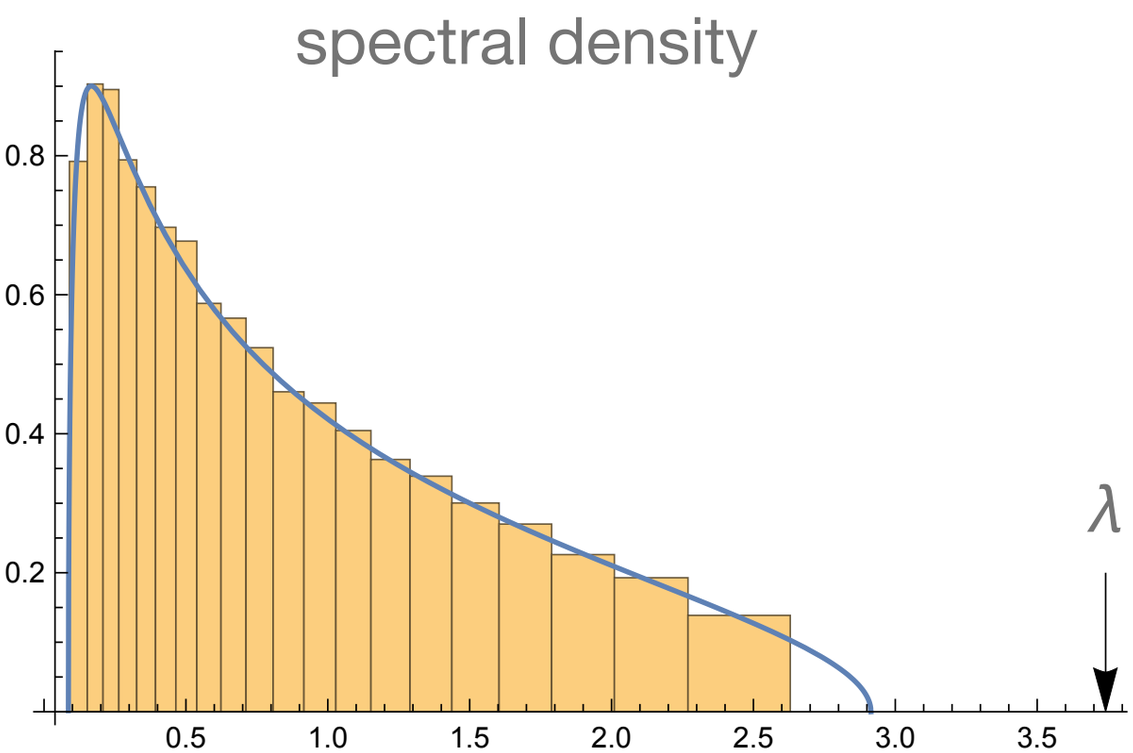
when does this perturbation rise above the “bulk” of random eigenvectors?

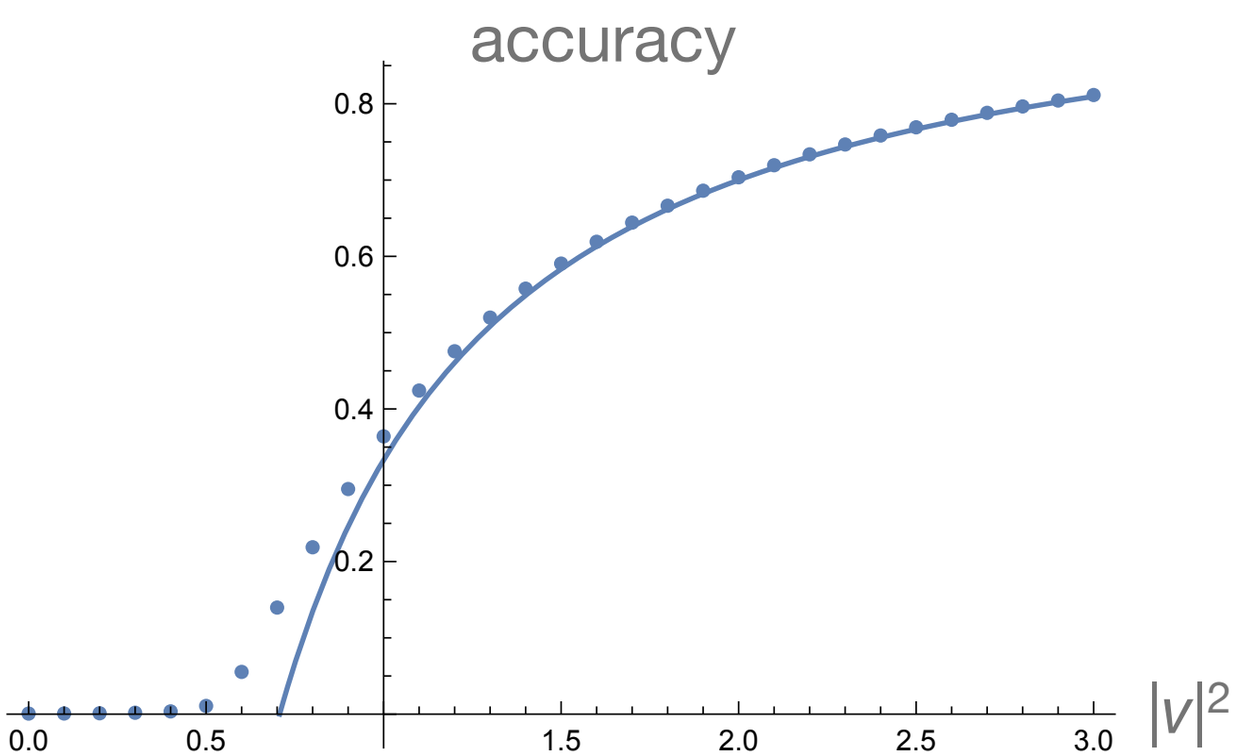
when it does, how accurately does the leading eigenvector point to v?

a phase transition at $m/n = 1/|v|^4$

when k is large, a gap between information and computation:

PCA is not optimal [Lesieur, De Bacco, Banks, Krzakala, Moore, Zdeborová]





Morals: physics meets machine learning

many problems involving sparse, noisy data have **phase transitions** beyond which no algorithm can find underlying structure

ideas from physics can help us find **optimal algorithms** that succeed all the way up to these transitions

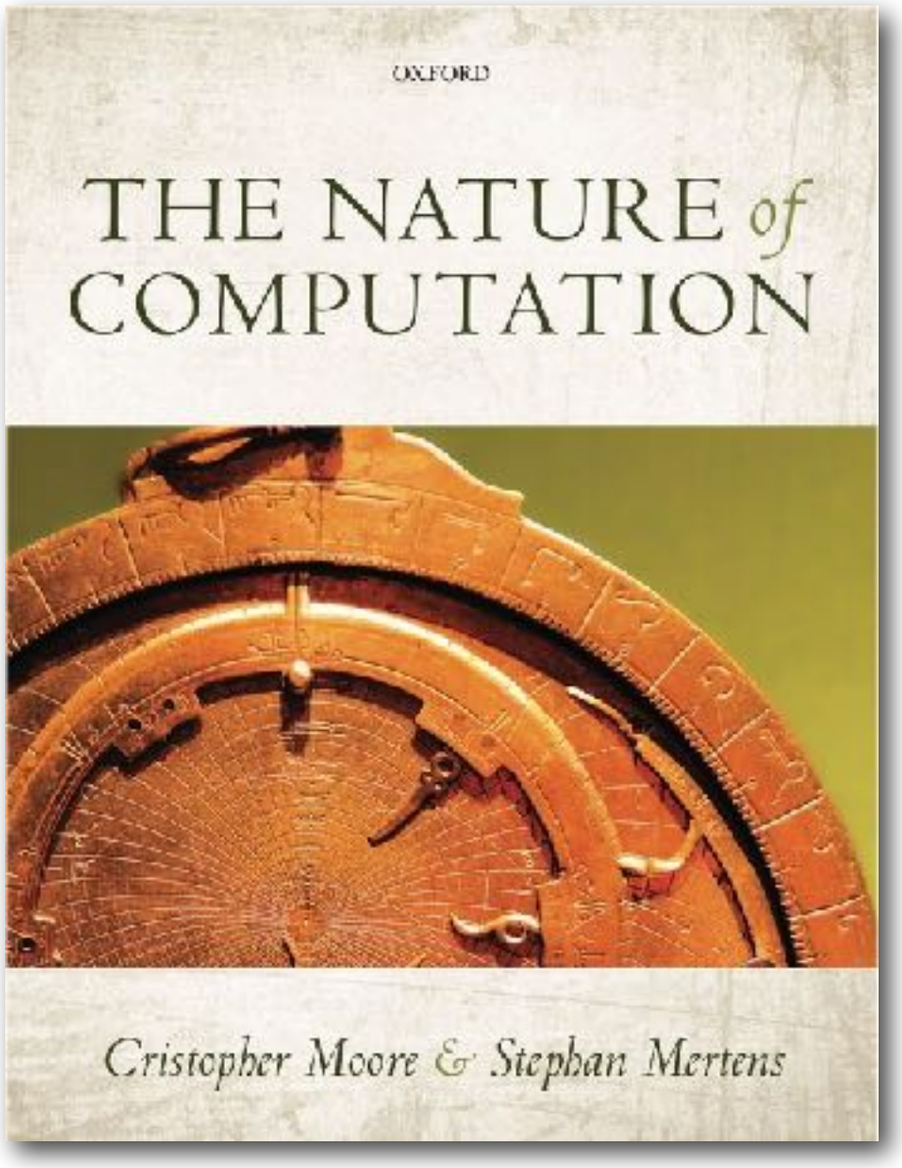
much of this work can be made mathematically rigorous

mathematical elegance pays off, even with real data: simple algorithms are faster, and we can understand their strengths and weaknesses

“as simple as possible, but no simpler”

Everything is a steam engine

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