Omidyar Fellowship Postdocs!
REU Summer for Undergrads!
Complex Systems Summer School!

National Institute of
Standards and Technology

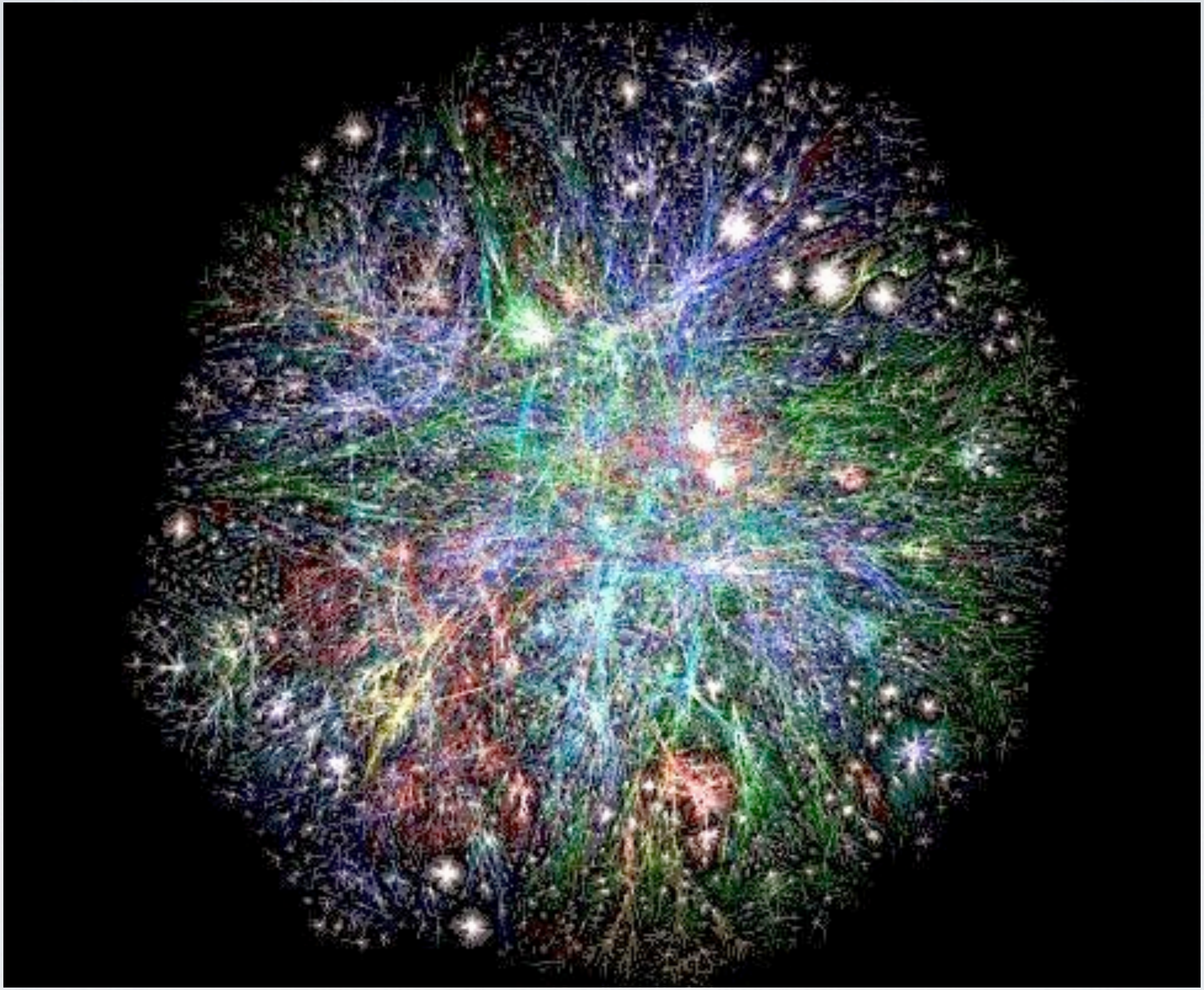
NIST
National Institute of
Standards and Technology
U.S. Department of Commerce

Los Alamos National Laboratory

# Block Models, Belief Propagation, and Phase Transitions in Community Detection

Cristopher Moore, Santa Fe Institute

joint work with
Xiaoran Yan, Yaojia Zhu, Lenka Zdeborová, Florent Krzakala,
Aurelien Decelle, Pan Zhang, Jean-Baptiste Rouquier, Tiffany Pierce,
Cosma Shalizi, Jacob Jensen, Lise Getoor, Aaron Clauset, and Mark Newman



# What is structure?

Structure is that which...

makes data different from noise: makes a network different from a random graph, or from a null model

helps us compress the data: describe the network succinctly, giving a human-readable summary of important structures

helps us generalize from data we’ve seen from data we haven’t seen: e.g. predict missing links from the links we know about

helps us understand what multiple networks have in common: e.g. structure of food webs, from the Cambrian to today

helps us coarse-grain the dynamics, reducing the number of variables: e.g. compartmentalized models in epidemiology

# The Bayesian approach

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

We can gracefully incorporate partial information: e.g. if

attributes of some nodes are known, or known with some confidence

some links are known, others not observed yet (e.g. food webs)

some links might be false positives (e.g. gene regulatory networks, protein interactions)

Use the inferred model to generalize from what we do know to what we don't: label unknown nodes, predict missing links, mark false positives

# The stochastic block model

nodes have discrete attributes: $k$ types of nodes

each node $i$ has type $t_i \in \{1,...,k\}$, with prior distribution $q_1,...,q_k$

$k \times k$ matrix $p$ of connection probabilities

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

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

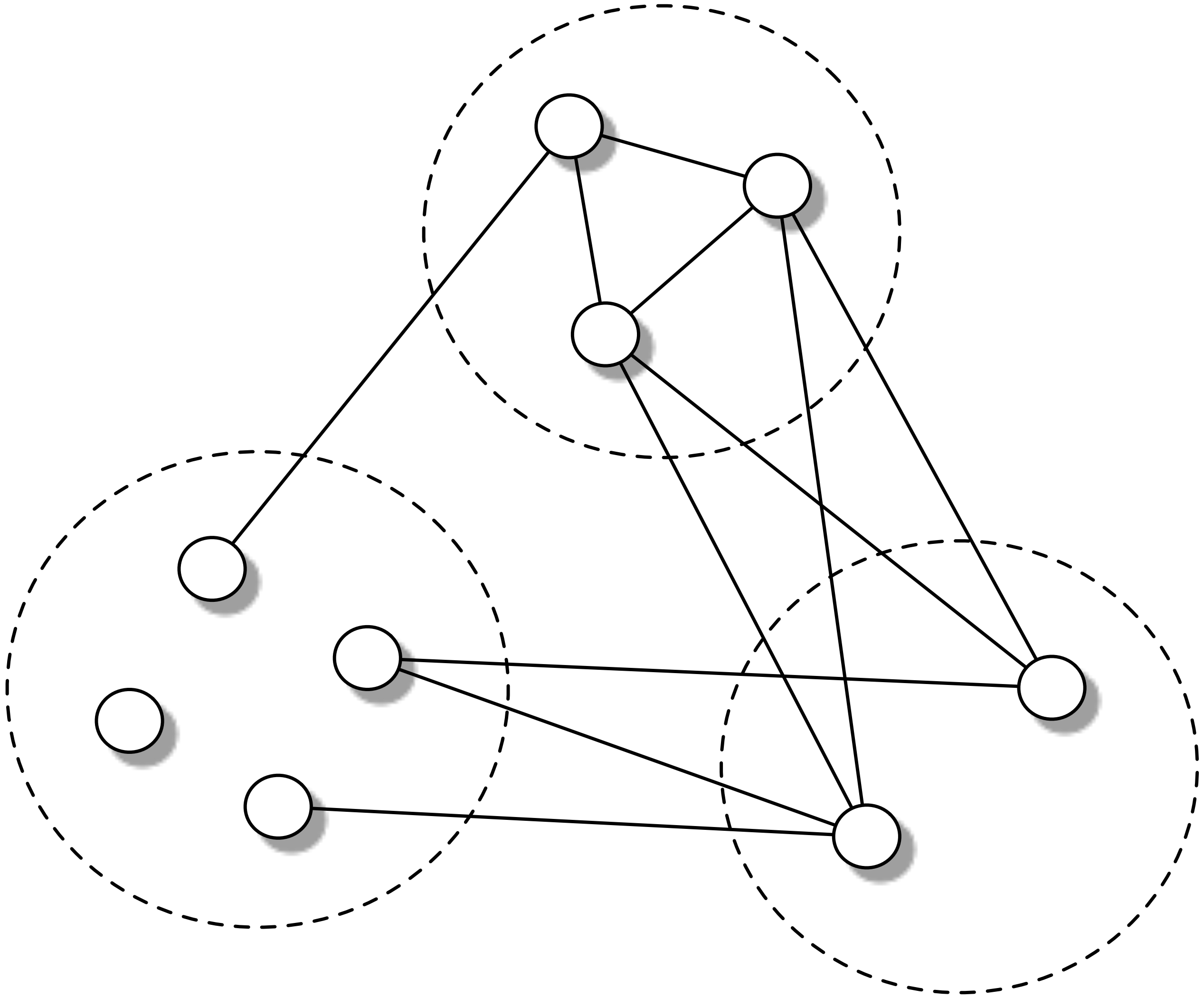
given a graph $G$, we want to simultaneously...

label the nodes, i.e., infer the type assignment $t : V \to \{1,...,k\}$

learn how types affect link probabilities, i.e., infer the matrix $p$

how do we get off the ground?

# Assortative and disassortative



functional groups, not just clumps

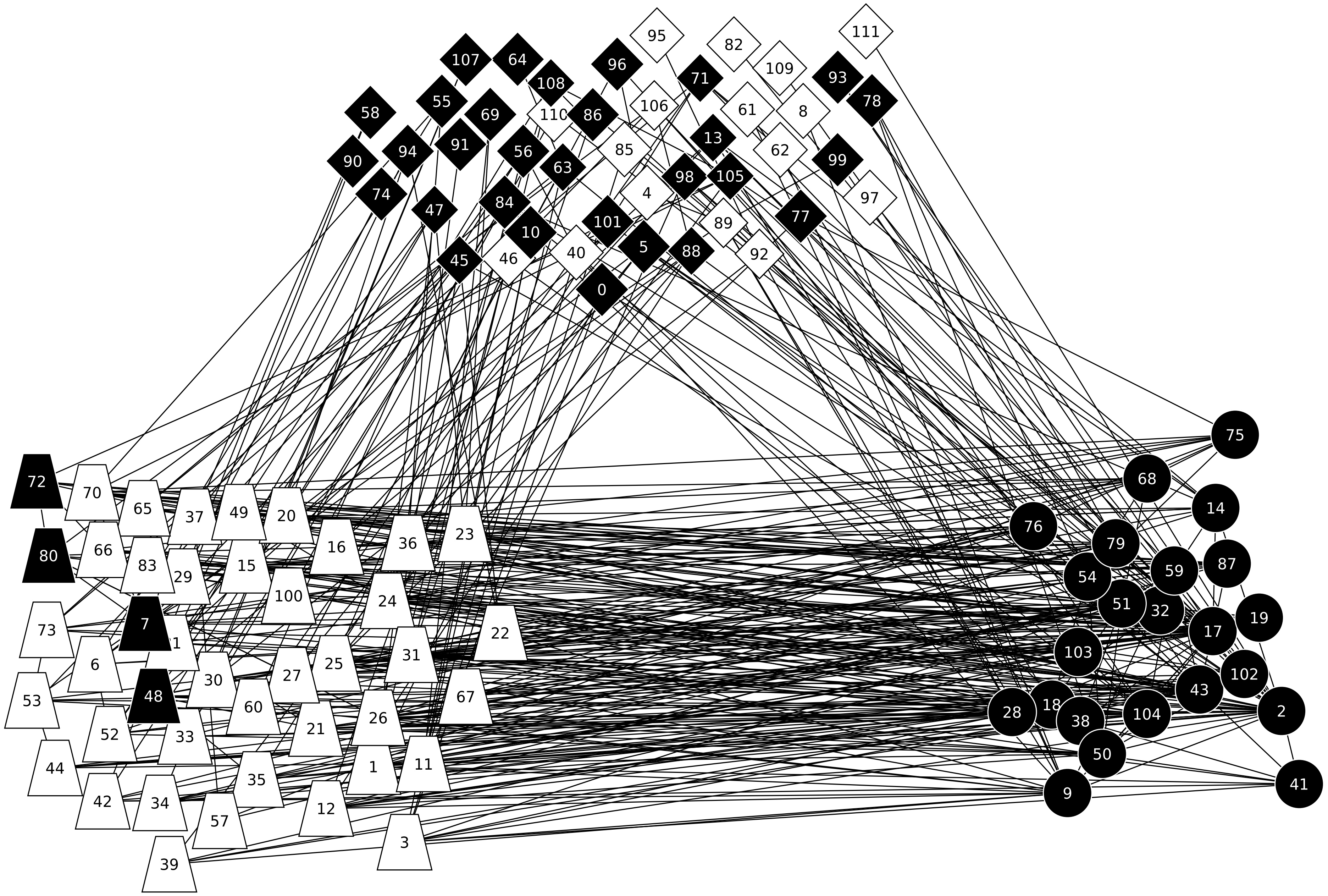
food webs: predators and prey

economics: suppliers and customers

word adjacencies: adjectives and nouns

social: leaders and followers

# Classifying words with a ground state:
# I record that I was born on a Friday



# Inferring the block model scalably

in the worst case, fitting the block model to a graph is an exponentially hard (NP-hard) optimization problem

in practice, there are now several scalable methods:

pseudolikelihood (see Liza Levina's talk)

stochastic optimization using subsampling (see Prem Gopalan's talk)

some variants have exact EM algorithms [Ball, Karrer, Newman]

spectral methods (see Mark Newman's and Elchanan Mossel's talks)

belief propagation [Decelle, Krzakala, Moore, Zdeborova]

the BP approach lets us build analogies with statistical physics, and reveals phase transitions in our ability to detect communities

# The likelihood

the probability of *G* given the types *t* and parameters θ=(*p*,*q*) is a product

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

so (after normalizing) the posterior probability of *t* given *G* is

$$P(t \mid G,\theta) = \frac{P(t \mid \theta)\, P(G \mid t,\theta)}{\sum_{t'\in\{1,\ldots,k\}^n} P(t' \mid \theta)\, P(G \mid t',\theta)}$$

$$\propto \prod_{i\in V} q_{t_i} \prod_{(i,j)\in E} p_{t_i,t_j} \prod_{(i,j)\notin E} (1 - p_{t_i,t_j})$$

# Method #1:
# Markov Chain Monte Carlo

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

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

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

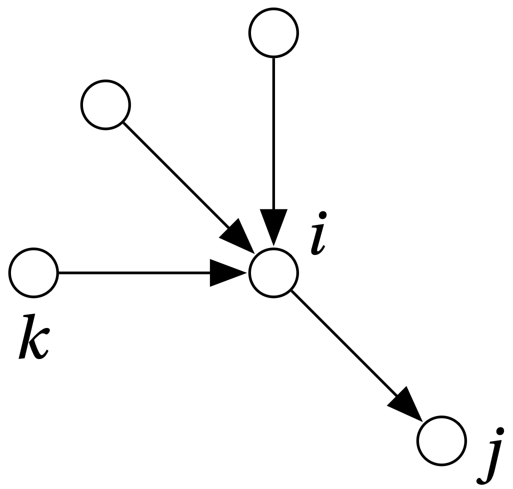
pretty good for finding ground states, but can get stuck in local optima

can speed up by introducing a temperature parameter:

- simulated annealing
- population annealing
- parallel tempering

but there's no free lunch

# Method #2:
# Belief propagation (a.k.a. the cavity method)



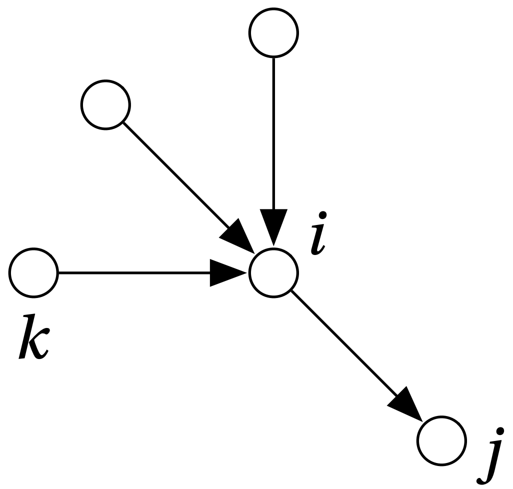
each node *i* sends a “message” to each of its neighbors j, giving *i*’s marginal distribution based on its other neighbors *k*

denote this message $\mu_r^{i \to j} = \text{estimate of } \Pr[t_i = r] \text{ if } j \text{ were absent}$

directly returns marginals, i.e. soft clustering, and two-point correlations

how do we update the messages?

# Updating the beliefs



conditional independence

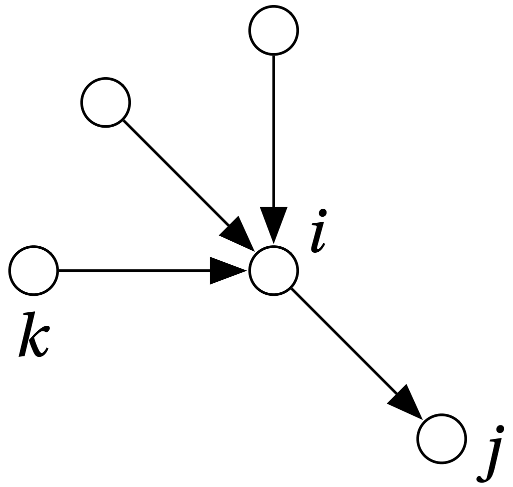
$$\mu_s^{i\to j} = \frac{1}{Z^{i\to j}}\, q_s \prod_{\substack{k\neq j\\(i,k)\in E}} \sum_r \mu_r^{k\to i} p_{rs} \times \prod_{\substack{k\neq j\\(i,k)\notin E}} \sum_r \mu_r^{k\to i}(1-p_{rs})$$

a complete graph of messages—takes O($n^2$) time to update

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

each node *k* applies an “external field” $\sum_r \mu_r^k (1-p_{rs})$ to all vertices of type *s*

# Making belief propagation scalable



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

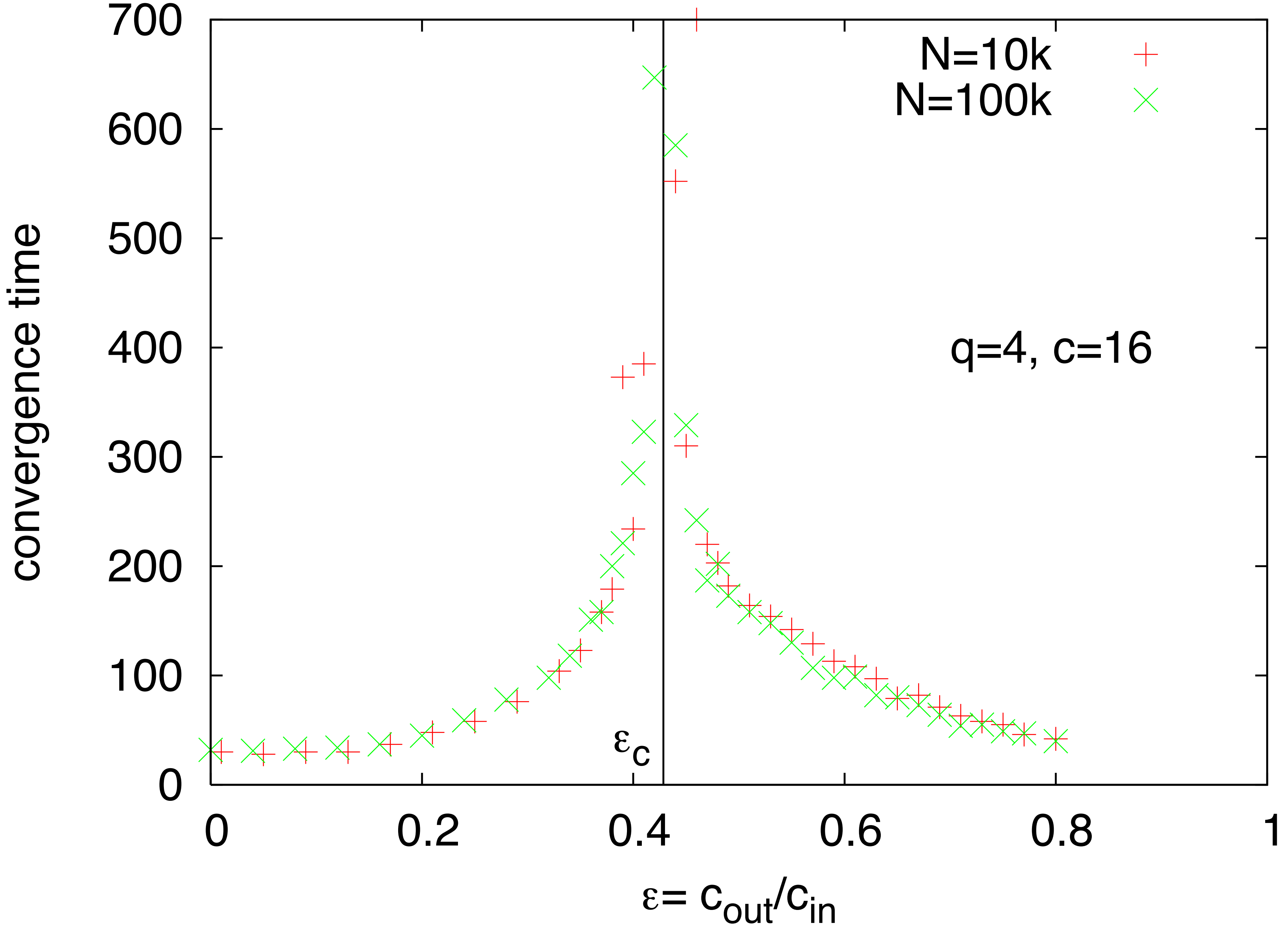
each update now takes O(*n*+*m*) time: scalable!

update until the messages reach a fixed point

converges quickly: can handle millions of nodes in minutes on a laptop

like Monte Carlo, can get stuck: try different initial messages

# BP converges in a small number of iterations on many networks: finite correlation length



[Decelle, Krzakala, Moore, Zdeborová, PRL 2011]

# A little statistical physics

the Boltzmann distribution: thermal equilibrium at temperature $T=1/\beta$

each state $t$ is a set of “spins”, or labels in our case

if a state $t$ has energy E($t$), then its probability is proportional to

$$P(t) \propto \mathrm{e}^{-\beta E(t)}$$

so (with $\beta$ =1) the “energy” of a state in the block model is

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

like an Ising or Potts model (except non-neighbors also interact, since non-edges are informative)

# Ground states vs. free energy

the most likely group assignment is a *ground state*: it maximizes

$$P(G \mid t, \theta)$$

and –log $P(G|t,\theta)$ is the *ground state energy*

one approach: find the $\theta=(p,q)$ that minimizes the ground state energy, i.e., maximize $P(G|t,\theta)$ as a function of $t$ and θ

but this overfits! good ground states even when there no real communities

for instance, random 3-regular graphs have bisections with only about 15% of the edges crossing from one side to the other

there are communities in the graph but not the underlying model

[Preview: it can be the other way around too!]

# Ground states vs. free energy

better to use the total probability of *G* given θ, summed over all $k^n$ labelings of the vertices:

$$
\begin{aligned}
P(G\,|\,\theta) &= \sum_{t\in\{1,\ldots,k\}^n} P(G,t\,|\,\theta) \\
&= \sum_{t\in\{1,\ldots,k\}^n} P(G\,|\,t,\theta)\,P(t\,|\,\theta)
\end{aligned}
$$

this is a *partition function,* and –log *P*(G|θ) is a *free energy*

total log-likelihood that *G* is generated by a block model with parameters θ

goal: find θ=*(p,q*) that minimizes the free energy, i.e., maximizes *P*(G|θ)

# Expectation-Maximization

Gradient ascent (or descent) in parameter space, maximizing total likelihood / minimizing free energy

(E step) given the current estimate of the parameters θ=*(p,q)*, estimate 1- and 2-point marginals of the Gibbs distribution (how?)

$$\mu_r^i = \Pr[t_i = r] \qquad \mu_{rs}^{ij} = \Pr[t_i = r \text{ and } t_j = s]$$

(M step) update θ=*(p,q)* to their most likely values given these marginals:

$$q_r = \frac{1}{N} \sum_i \mu_r^i \qquad p_{rs} = \frac{\sum_{(i,j)\in E} \mu_{rs}^{ij}}{q_r q_s N^2}$$

# Approximating the free energy: variational trick

$$\begin{aligned}\log P(G\,|\,\theta) &= \log \sum_t P(G\,|\,t,\theta)\\ &= \log \mathop{\mathbb{E}}_{t\sim Q} \frac{P(G\,|\,t,\theta)}{Q(t)}\\ &\geq \mathop{\mathbb{E}}_{t\sim Q} \log \frac{P(G\,|\,t,\theta)}{Q(t)}\\ &= \mathop{\mathbb{E}}_{t\sim Q} \log P(G\,|\,t,\theta) + S(Q)\end{aligned}$$

$$\begin{aligned}-\beta F = \log Z &= \log \sum_t \mathrm{e}^{-\beta E(t)}\\ &= \log \mathop{\mathbb{E}}_{t\sim Q} \frac{\mathrm{e}^{-\beta E(t)}}{Q(t)}\\ &\geq -\beta \mathop{\mathbb{E}}_{t\sim Q} E(t) + S(Q)\\ &= -\beta \langle E\rangle + S(Q)\end{aligned}$$

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 (or Kikuchi) free energy

average "energy" (log probability) depends just on 1- and 2-point marginals,

$$E(t) = \sum_{(i,j)\in E} \log p_{t_i,t_j} + \sum_{(i,j)\notin E} \log(1 - p_{t_i,t_j}) + \sum_i \log q_{t_i}$$

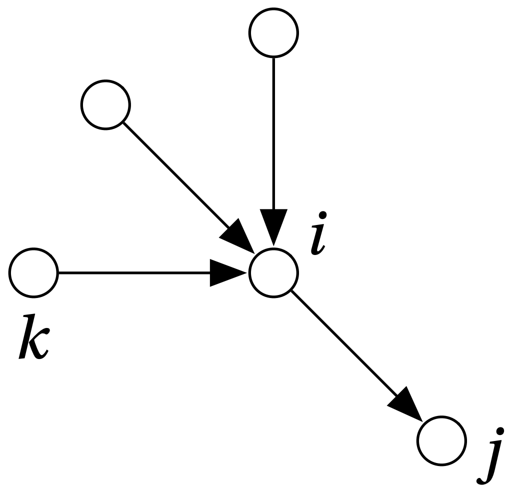
$$\langle E \rangle = \sum_{(i,j)\in E} \sum_{r,s=1}^{k} \mu_{rs}^{ij} \log p_{rs} + \sum_{(i,j)\notin E} \sum_{r,s=1}^{k} \mu_{rs}^{ij} \log(1 - p_{rs}) + \sum_{i,r} \mu_r^i \log q_r$$

but the entropy is more complicated... approximate the Gibbs distribution with a form that depends on the 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

# Belief propagation and the Bethe free energy



BP isn’t just a heuristic...

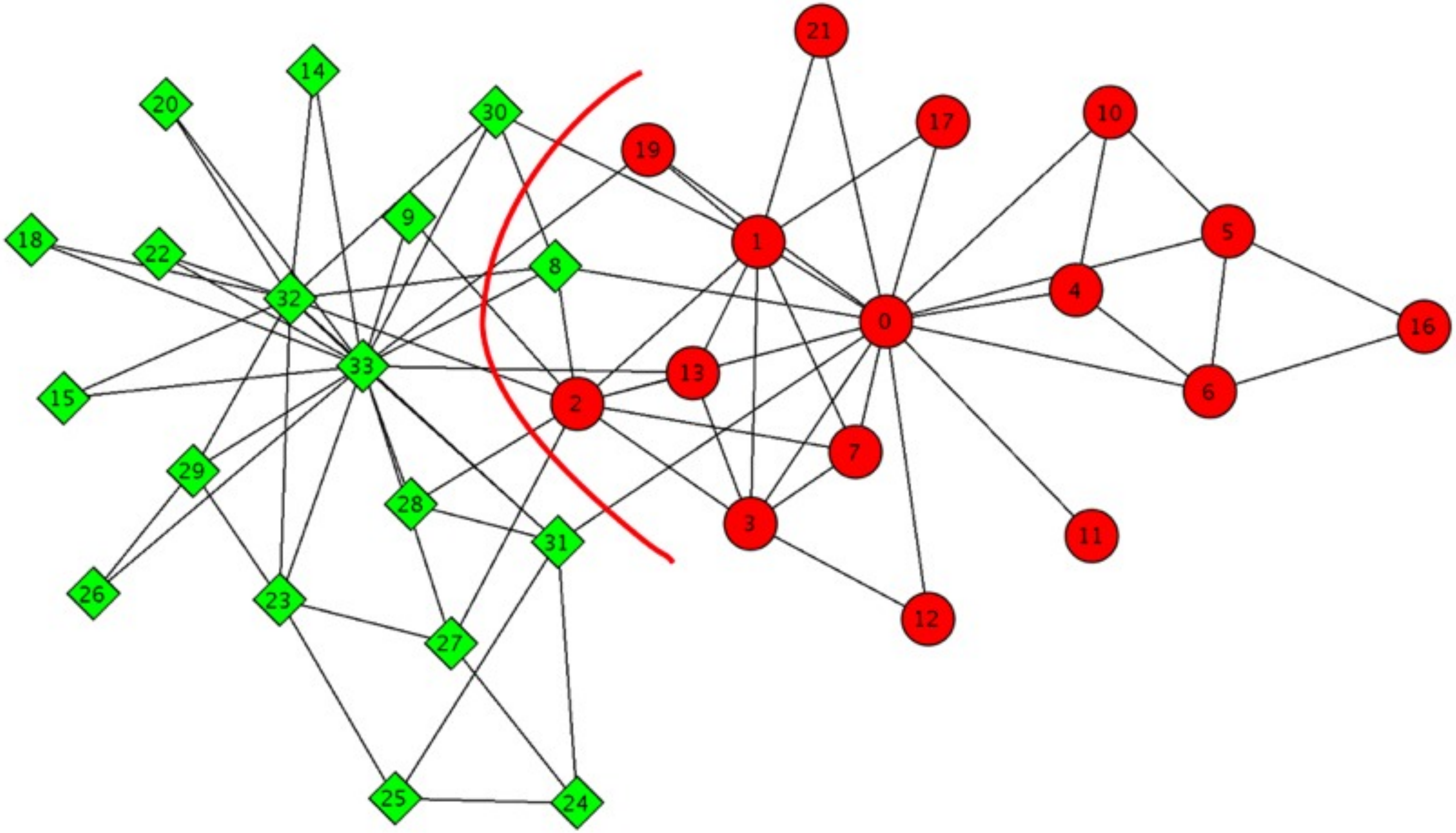
BP fixed points are exactly stationary points for the Bethe free energy [Yedidia]

for each setting of the parameters θ, can compute the Bethe free energy

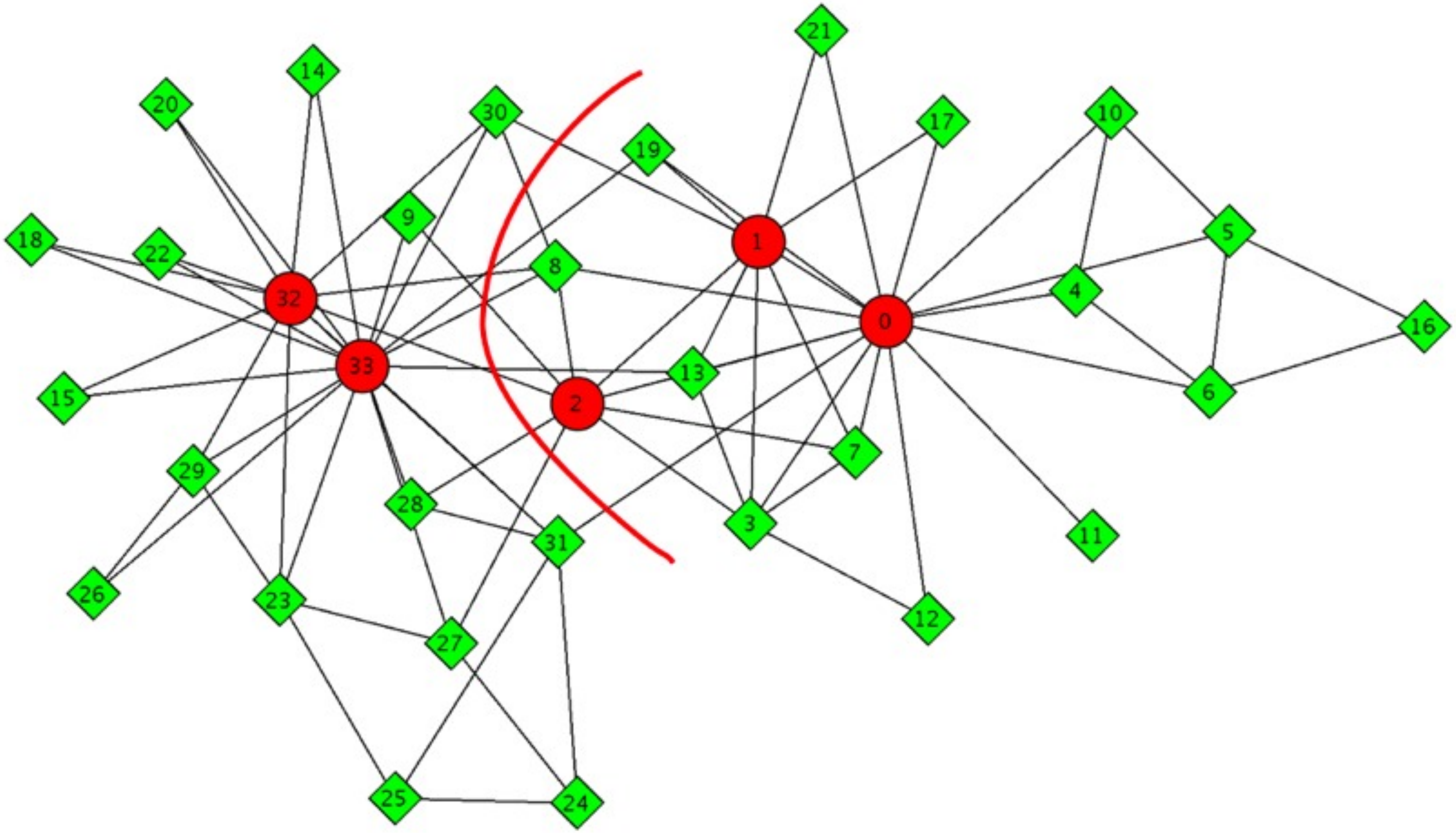
can explore free energy landscape as a function of θ

can use Bethe free energy in likelihood-ratio-based model selection (Xiaoran’s talk)

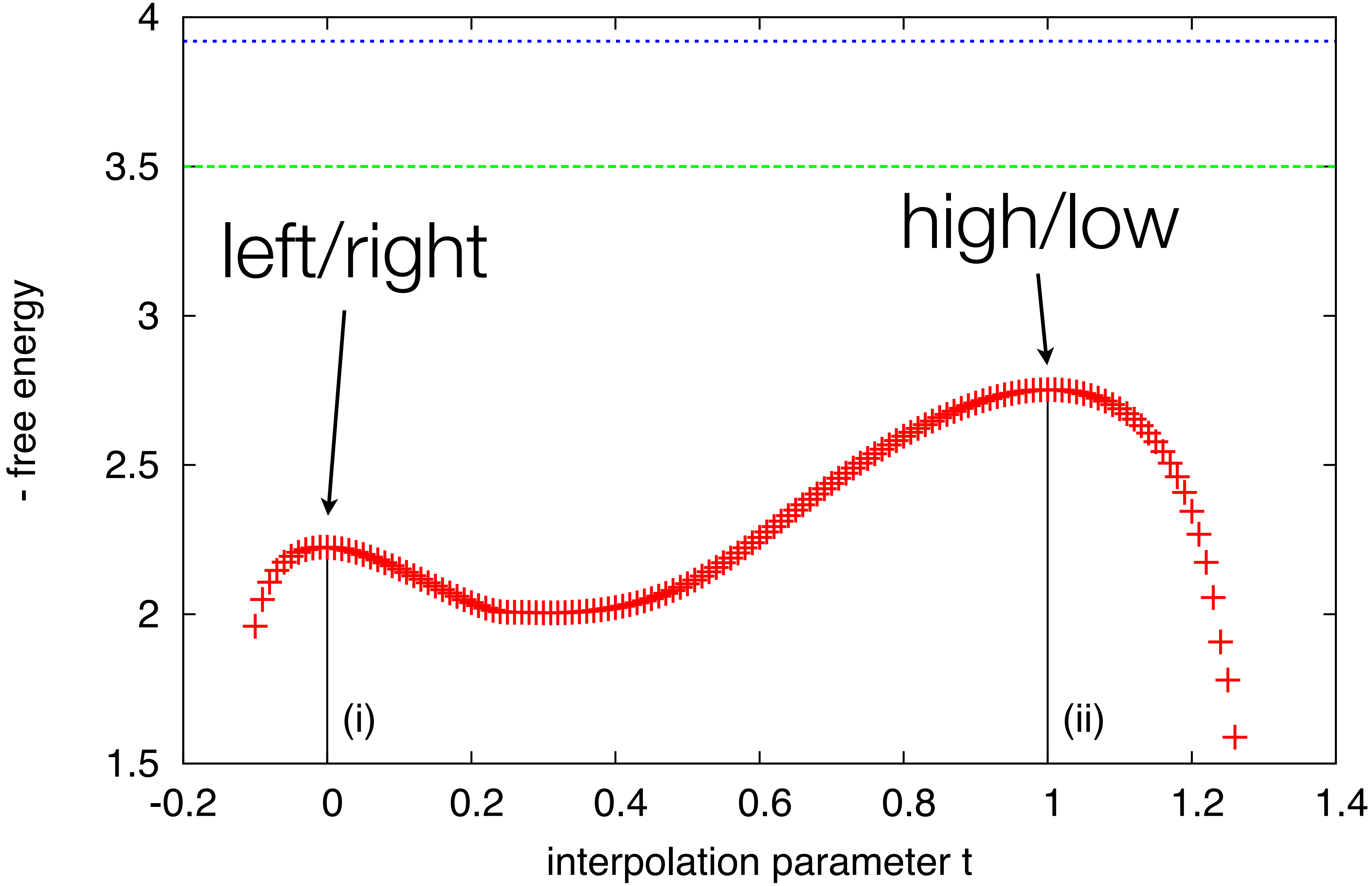
# The Karate Club: two factions



# The Karate Club: leaders vs. followers

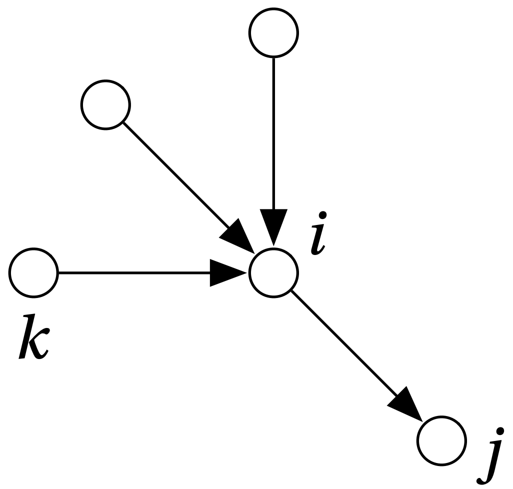


# Two local optima in free energy



degree-corrected block model focuses on factions (see Mark Newman’s talk)

# The double life of Belief Propagation



BP is a fast algorithm we can run on real networks to fit the block model...

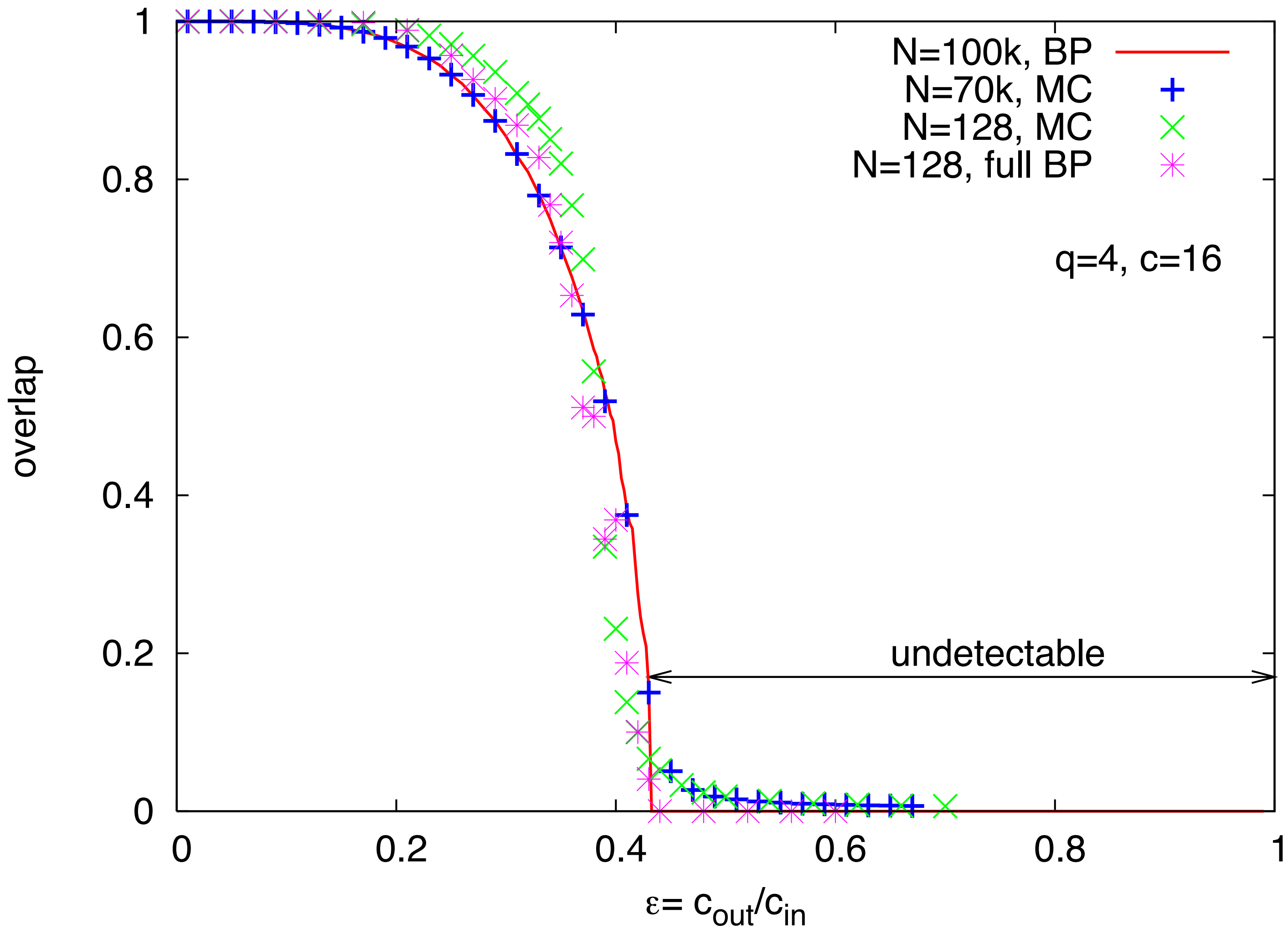
but it’s also a framework for analytic calculations

assume that the graph is generated by the block model: can we recover its parameters, and detect the communities?

given a distribution of messages and the degree distribution of your generative model, BP updates these distributions

find fixed points, their basins of attraction and their stability

# A phase transition: detectable to undetectable communities



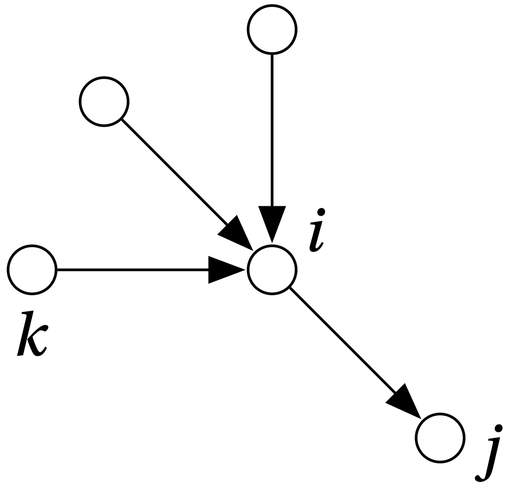
when the rows of $p_{ij}$ are different enough, BP can recover the communities

but there is a transition where it can’t — and no algorithm can!

the ensemble of graphs “knows” the communities, but a typical graph doesn’t

[Decelle, Krzakala, Moore, Zdeborová, PRL 2011]

# When uniformity is attractive



in the sparse case where $p_{rs} = c_{rs}/n$,

suppose messages are close to uniform, $\mu_r^{k\to i} = q_r + \varepsilon_r^k$

linearizing the BP equations, we find that $\varepsilon^i = T \cdot \varepsilon^k$ where $T_{rs} = q_r \left( \frac{c_{rs}}{c} - 1 \right)$

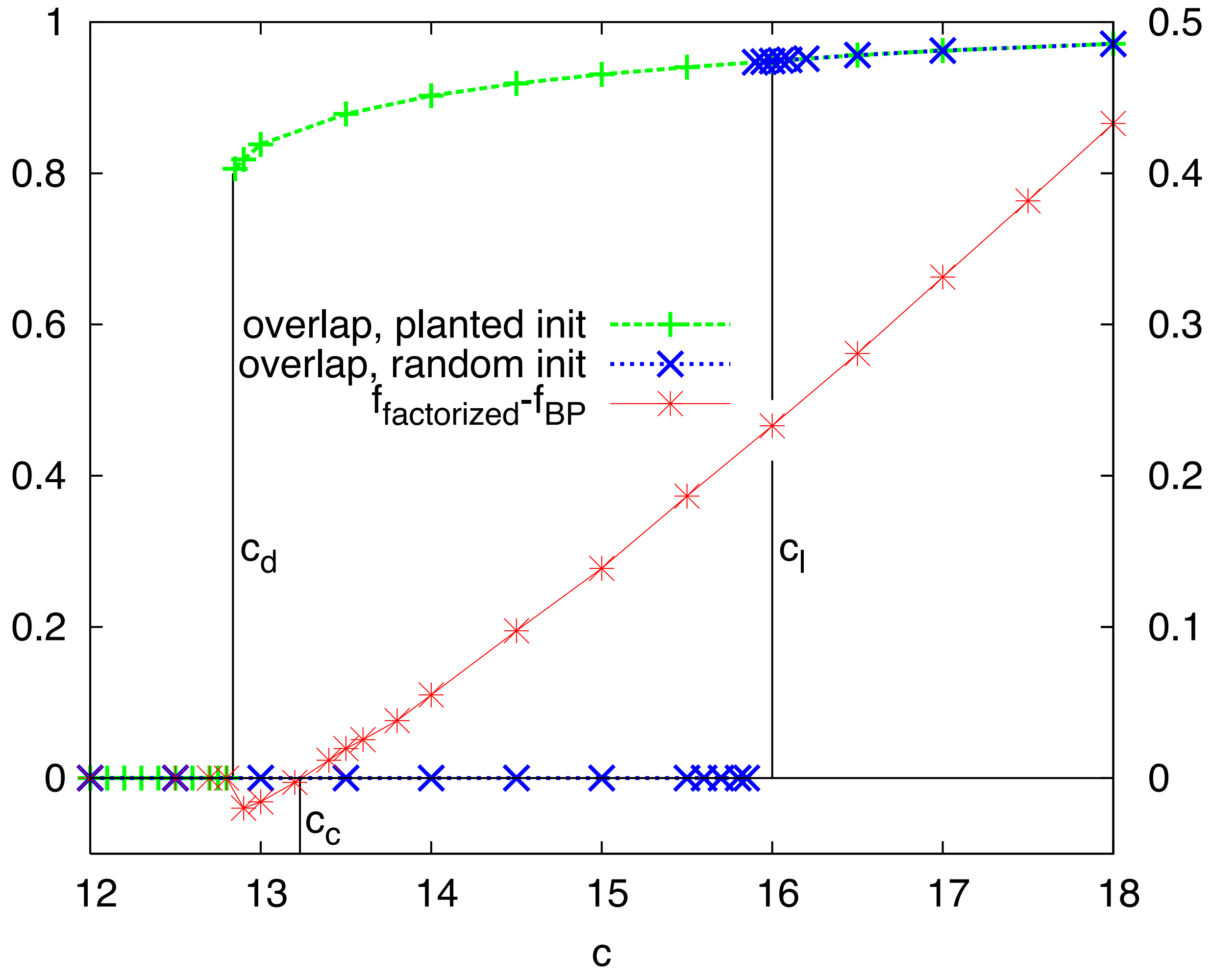
if *T*'s largest eigenvalue obeys $c\lambda^2 < 1$, then $|\varepsilon|^2$ decays exponentially when we sum over $c^\ell$ independent leaves at depth $\ell$ : Kesten-Stigum bound

in the "undetectable" regime, block model is *contiguous* to *G*(*n*,*p*): any event that holds w.h.p. in one model holds in the other [Mossel, Neeman, Sly 2012]

in the not-so-sparse case, we get the same bound from spectra of random matrices [Nadakuditi & Newman, PRL]

in the sparse case, use a different spectral operator (see Elchanan Mossel's talk)

# Detectable, but only if you know where to look

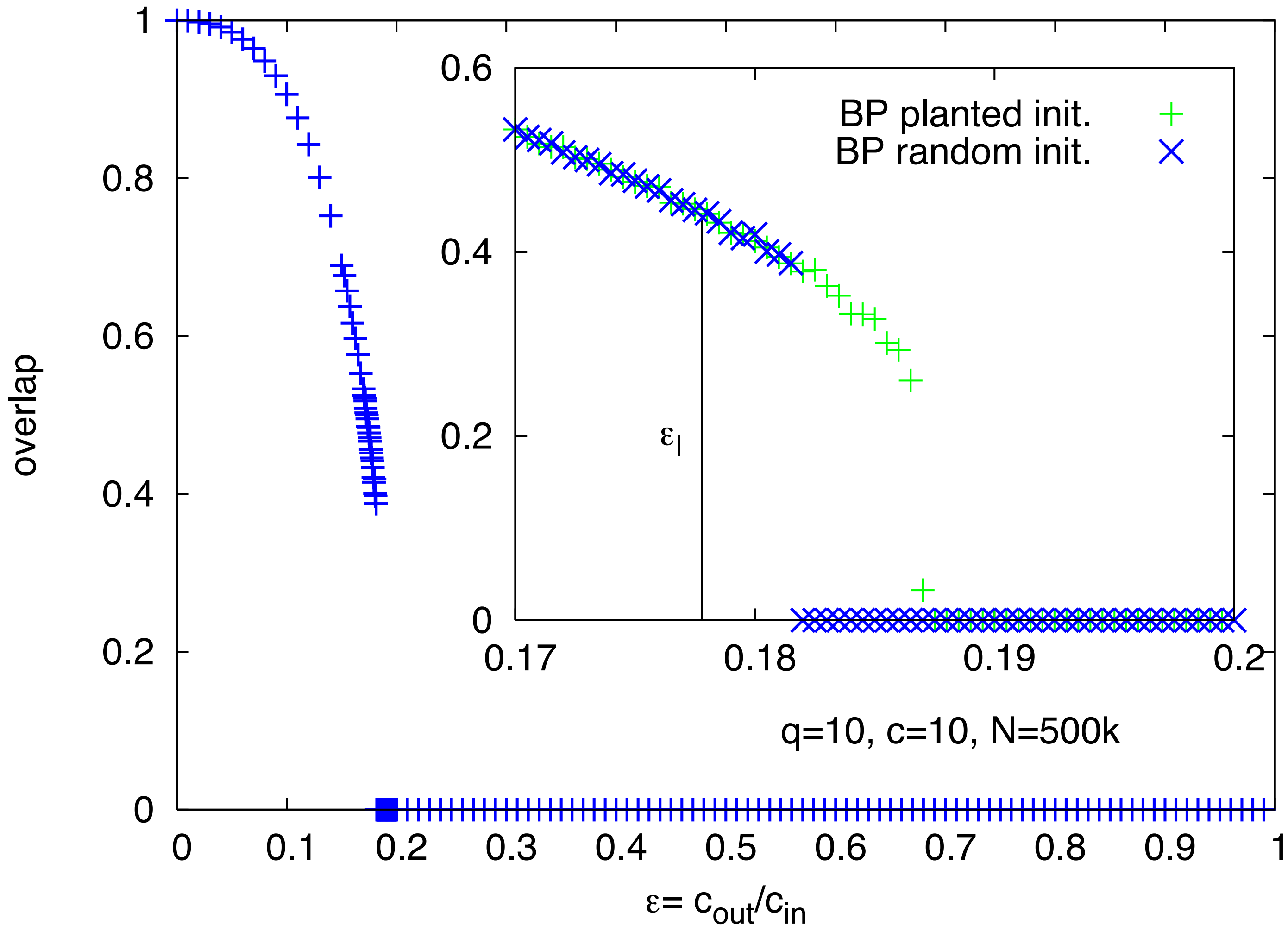


planted model: choose a 5-coloring of the vertices, build the graph around it

for c < 16 = $(5-1)^2$, random initial messages lead to an uncorrelated fixed point

for 13 < c < 16 if the initial messages are the true colors, BP finds a fixed point correlated with them: but this fixed point has a very small basin of attraction

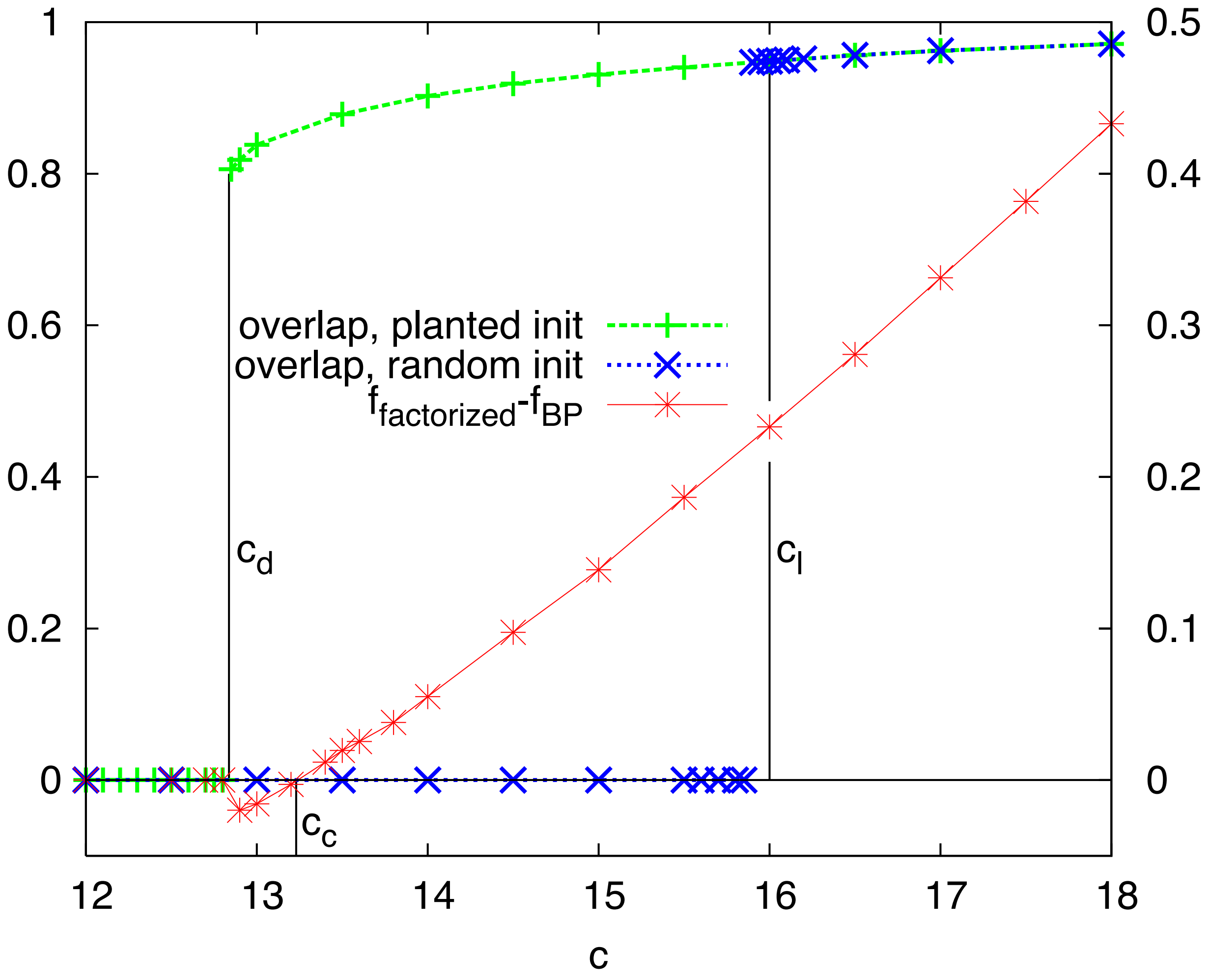
# Detectable, but only if you know where to look

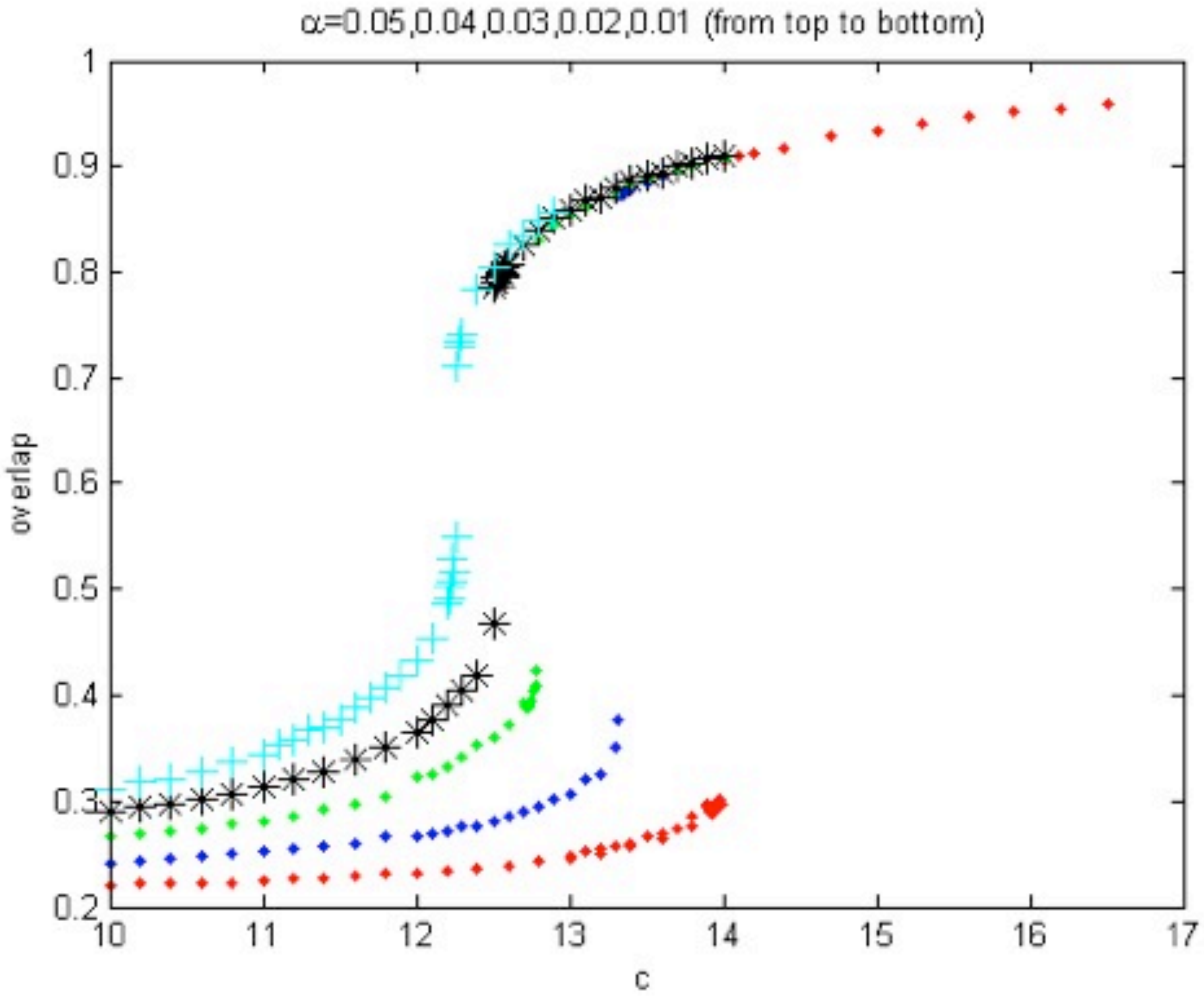


this “detectable but hard” regime also occurs in assortative community structures when the number of groups is large, though it’s narrower

10 groups, average degree 10

# Phase transitions in semisupervised learning (work in progress)





suppose we are given the correct types for *αn* random nodes. what accuracy can we achieve?

a phase transition: when *α* crosses a threshold, our knowledge percolates from the labeled nodes and their neighbors to the rest of the network, causing a discontinuous jump in the accuracy

# The story so far

statistical inference using generative models of networks lets us detect communities, classify nodes, and predict missing links

the block model allows for functional groups of nodes, not just “clumps”

Belief Propagation and expectation-maximization algorithms let us identify these groups, and learn model parameters, often in linear time: scalable!

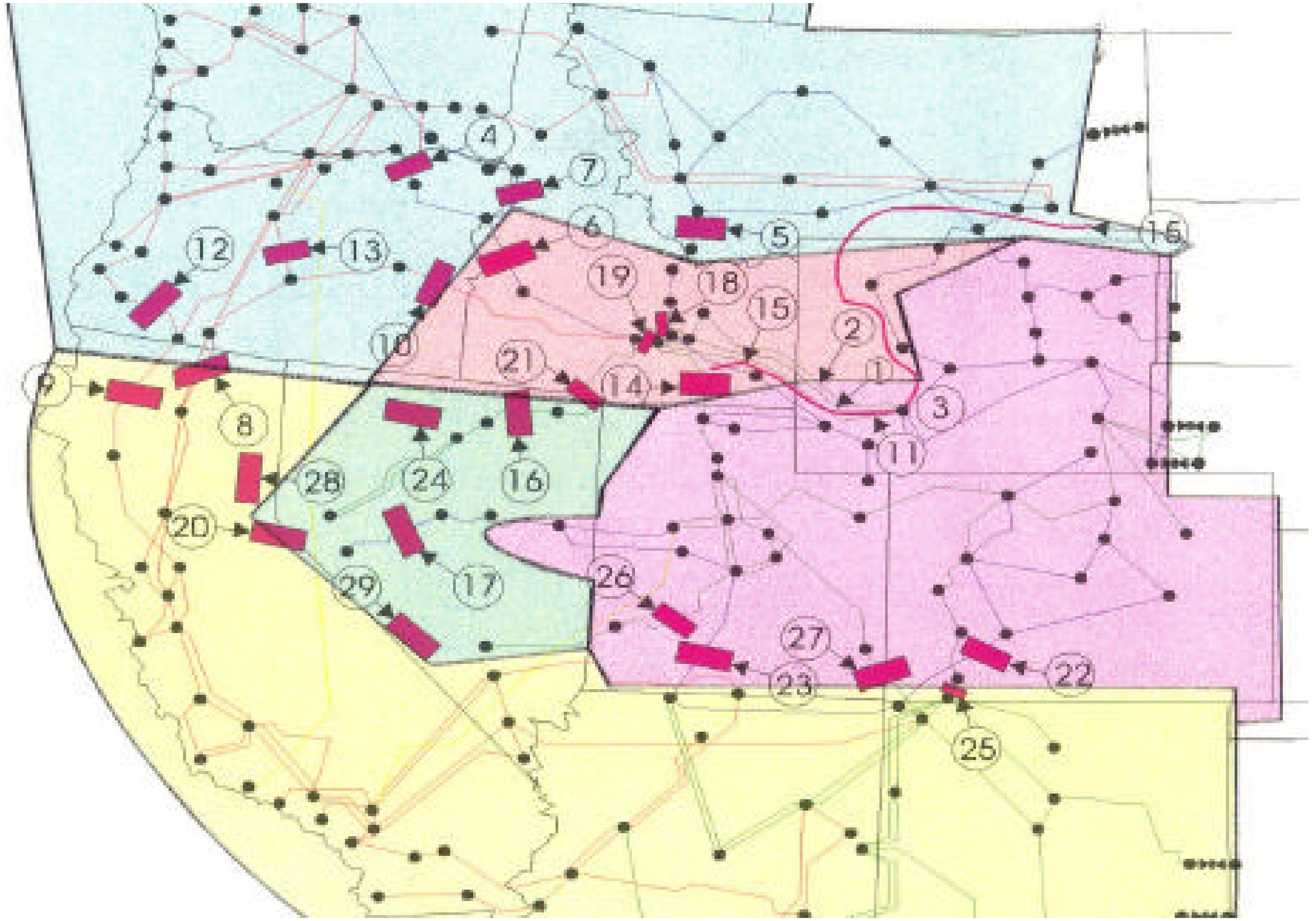
Belief Propagation also lets us analytically explore phase transitions in reconstruction of communities (assuming an underlying model)

we can elaborate these models by adding discrete or continuous attributes: degree distributions, edge types, social status, overlapping communities, hierarchy, signed edges, document content [e.g. Zhu, Yan, Getoor, Moore]

but a cautionary note...

# A real cascade of line and generator failures

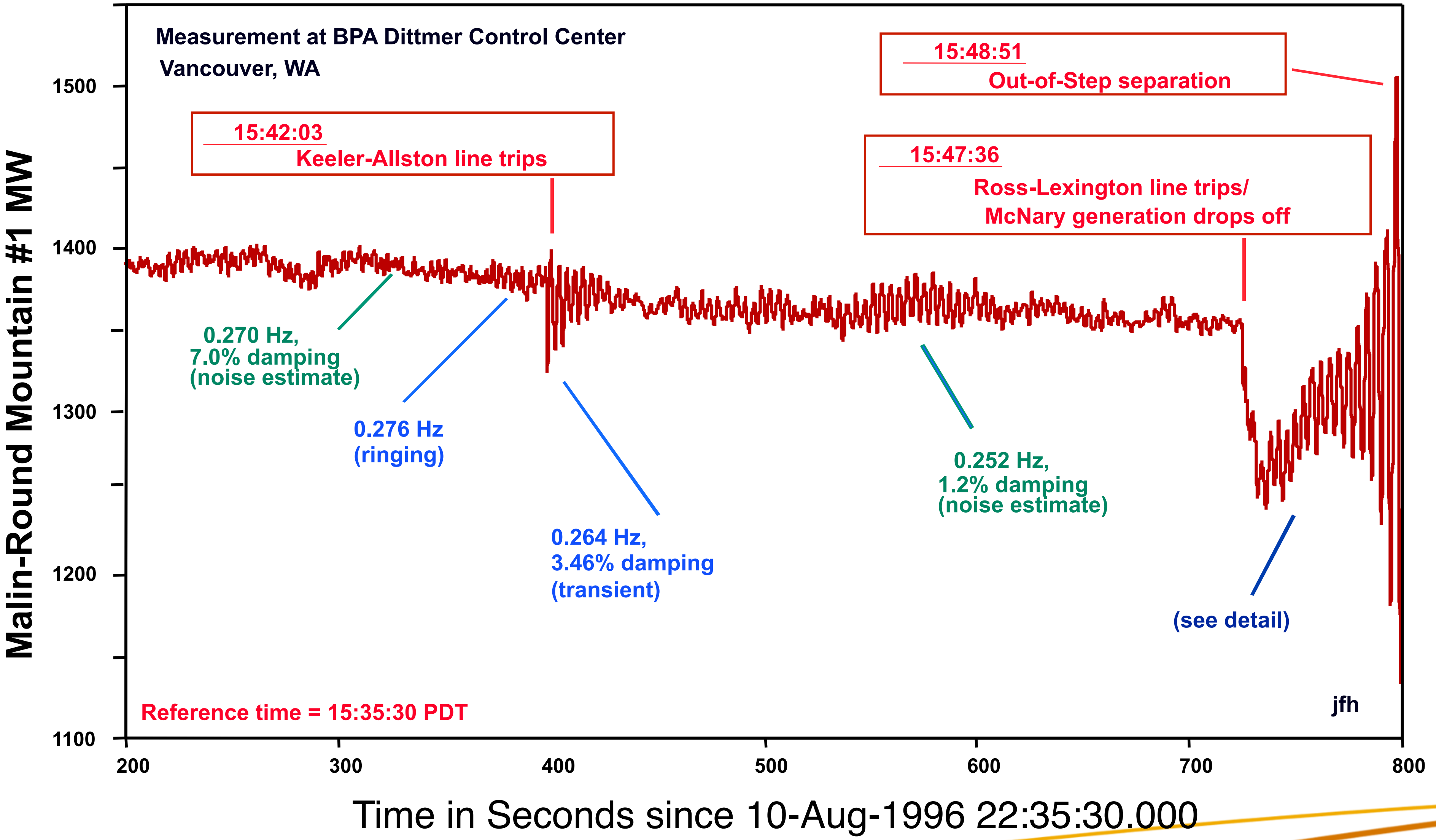
Sequence of outages in Western blackout, July 2 1996



from NERC 1996 blackout report

# Rich dynamics of coupled, nonlinear oscillators

## Sequence of Events



Pacific Northwest
NATIONAL LABORATORY

# Beyond topology

we need a new network theory that doesn't focus on topology alone

nodes and edges have rich attributes:

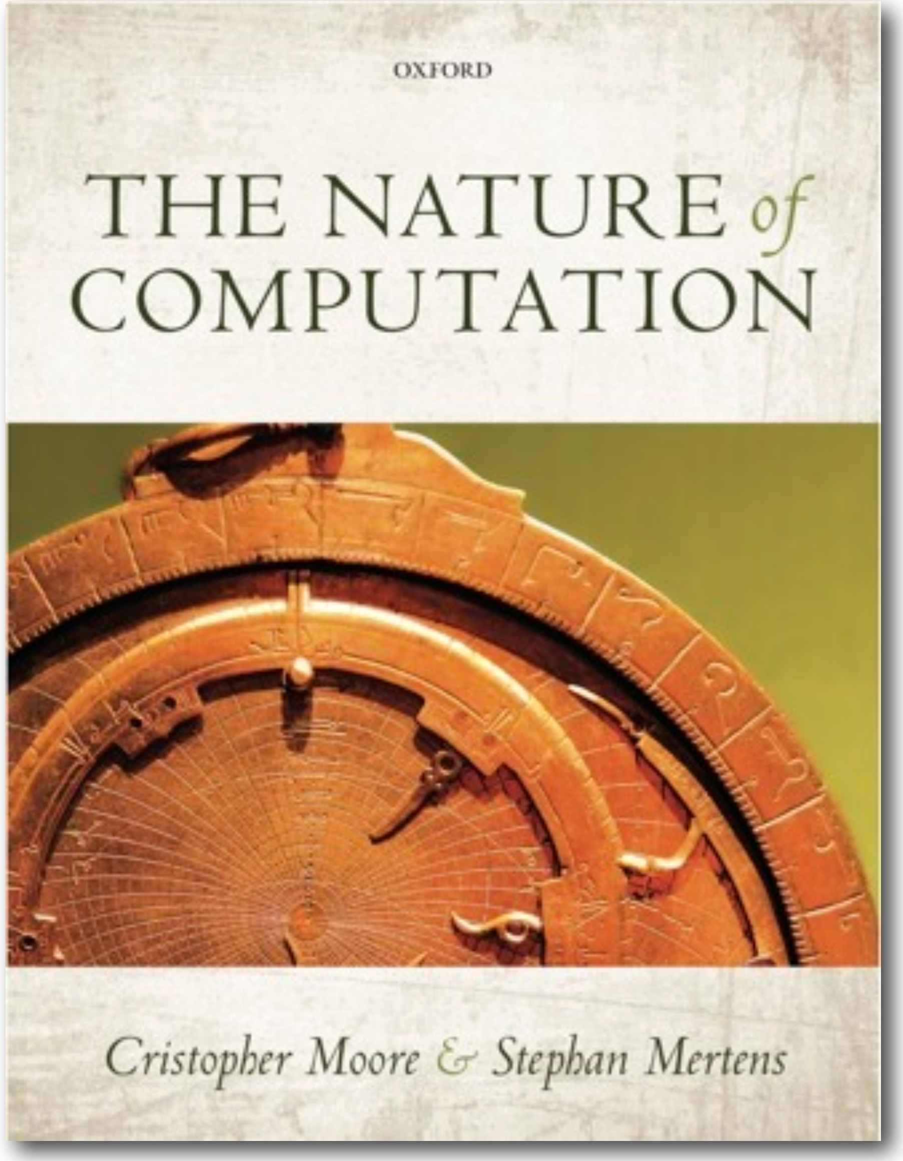
> power grid: generators have nonlinear dynamics at many time scales, transmission lines have capacities, users have fluctuating demands...
>
> cybersecurity: multiple types of links between computers (web fetches, SSH links) with timing, duration, packet size... and many links are unique
>
> food webs: species have populations, links have nutrient flows....
> dynamic response to climate change, species loss, invasive species

networks are rich, dynamic data sets, not just lists of nodes and edges

extending Bayesian inference to richer data is possible, but challenging

application dependent! do we want to label nodes? predict missing links? understand dynamics? or what?

# 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

A creative, insightful, and accessible introduction to the theory of computing, written with a keen eye toward the frontiers of the field and a vivid enthusiasm for the subject matter.

Jon Kleinberg, Cornell

A treasure trove of ideas, concepts and information on algorithms and complexity theory. Serious material presented in the most delightful manner!

Vijay Vazirani, Georgia Tech

A fantastic and unique book, a must-have guide to the theory of computation, for physicists and everyone else.

Riccardo Zecchina, Politecnico de Torino

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

Cosma Shalizi, Carnegie Mellon

# Acknowledgments

and the McDonnell Foundation, DARPA/AFOSR, NIST, and the NSF