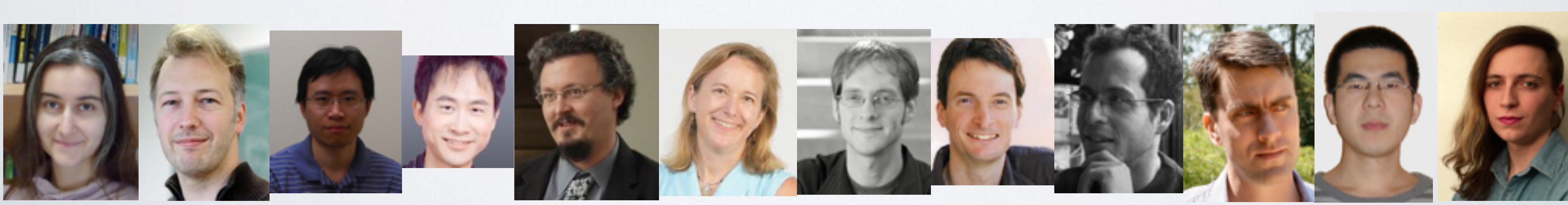
Phase Transitions in Community Detection and Clustering

Cristopher Moore, Santa Fe Institute

joint work over the years with Aurelien Decelle, Lenka Zdeborová, Florent Krzakala, Xiaoran Yan, Yaojia Zhu, Cosma Shalizi, Lise Getoor, Aaron Clauset, Mark Newman, Elchanan Mossel, Allan Sly, Pan Zhang, and Jess Banks



How can we find patterns in data?

How can we find patterns in data?
Fundamental limits

How can we find patterns in data?
Fundamental limits
Phase transitions

How can we find patterns in data?
Fundamental limits
Phase transitions
Optimal algorithms

How can we find patterns in data?

Fundamental limits

Phase transitions

Optimal algorithms

Statistical inference $\Leftrightarrow$ statistical physics

The analogy between physics and inference

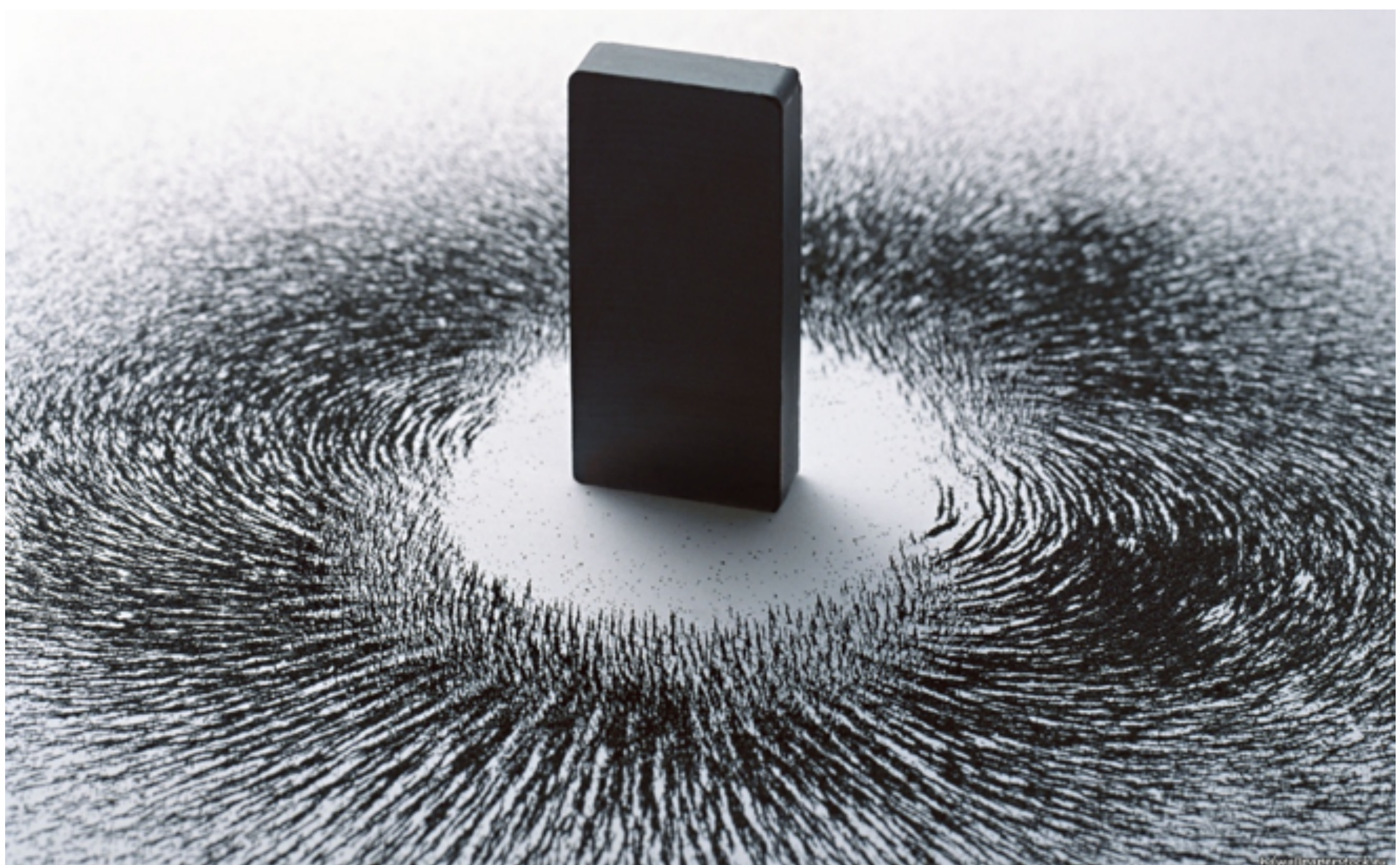
The analogy between physics and inference

the atoms of a block of iron interact with their neighbors $\uparrow\uparrow\uparrow\downarrow\downarrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\uparrow$

The analogy between physics and inference

the atoms of a block of iron interact with their neighbors $\uparrow\uparrow\uparrow\downarrow\downarrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\uparrow$

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



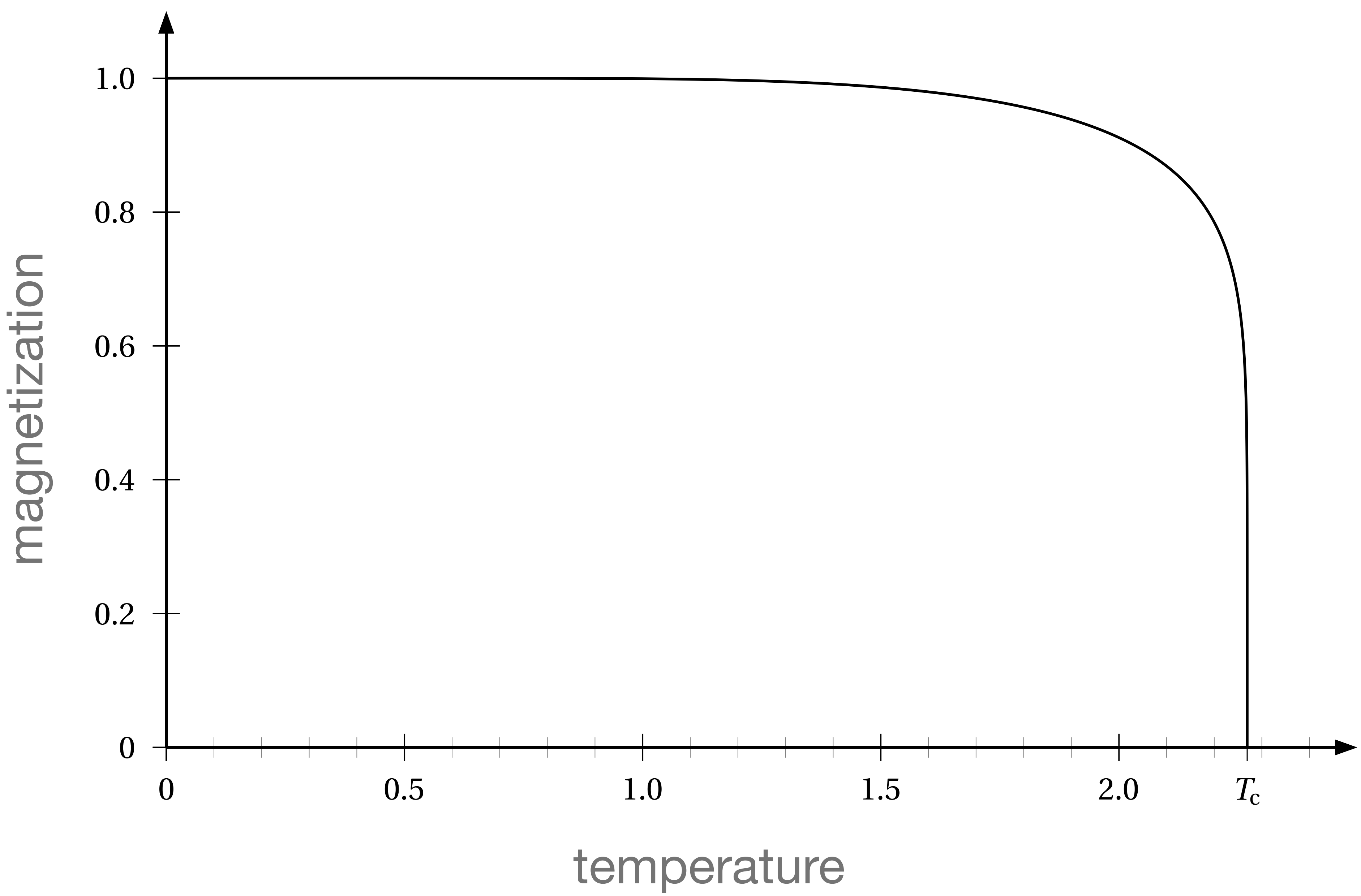
The analogy between physics and inference

The analogy between physics and inference

but when the iron is heated beyond a critical temperature, it suddenly loses its ability to hold a magnetic field: the atoms become uncorrelated

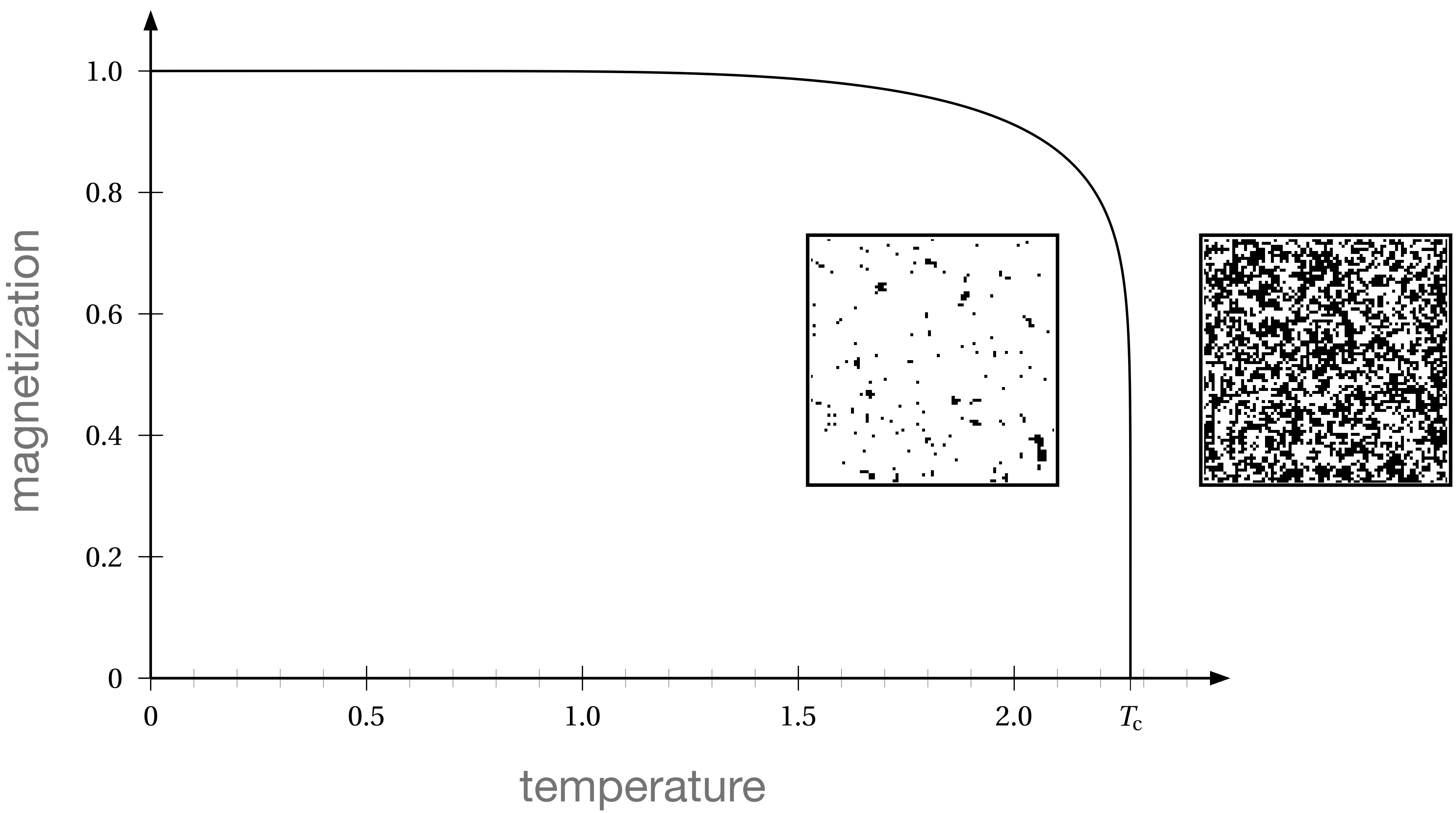
The analogy between physics and inference

but when the iron is heated beyond a critical temperature, it suddenly loses its ability to hold a magnetic field: the atoms become uncorrelated



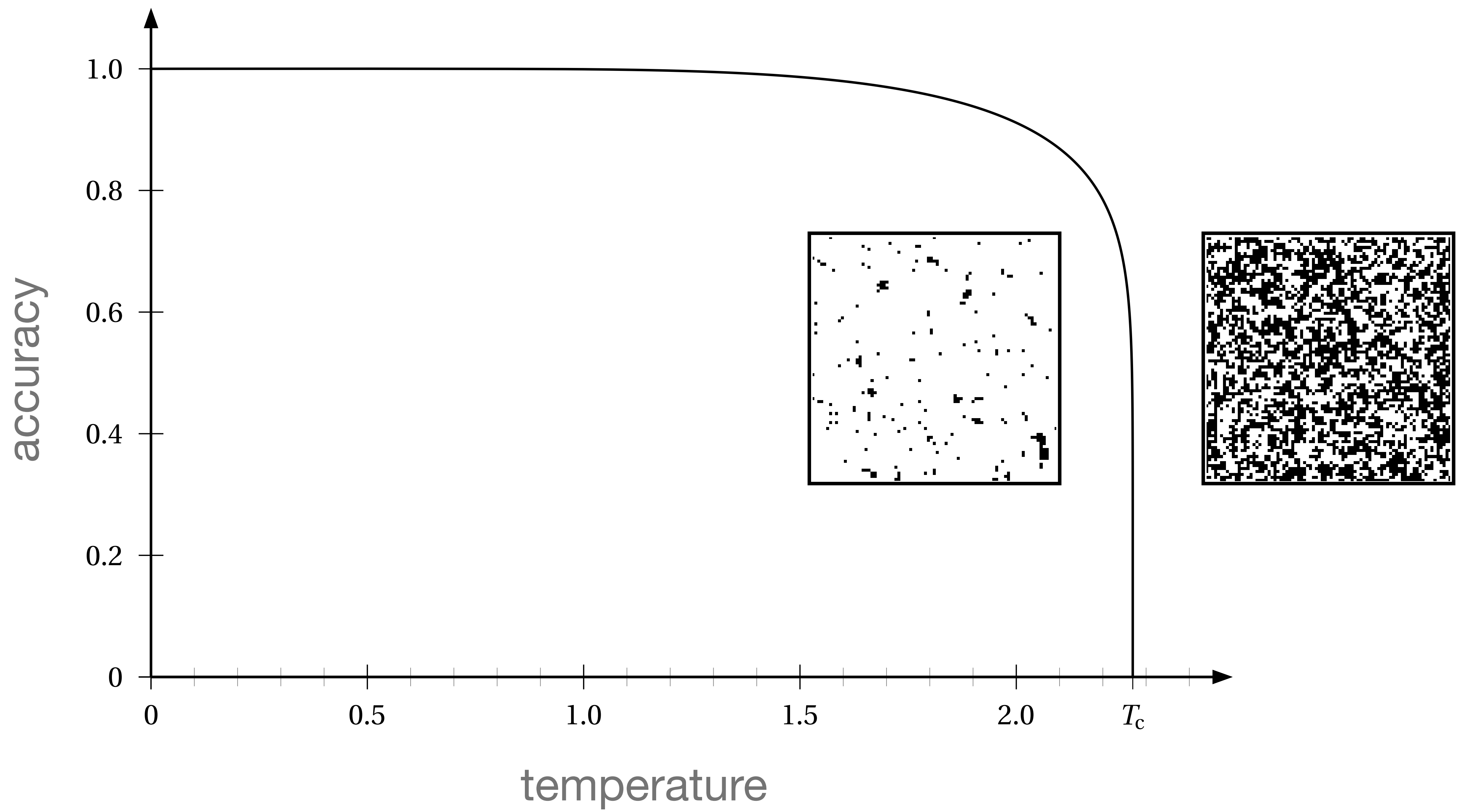
The analogy between physics and inference

but when the iron is heated beyond a critical temperature, it suddenly loses its ability to hold a magnetic field: the atoms become uncorrelated



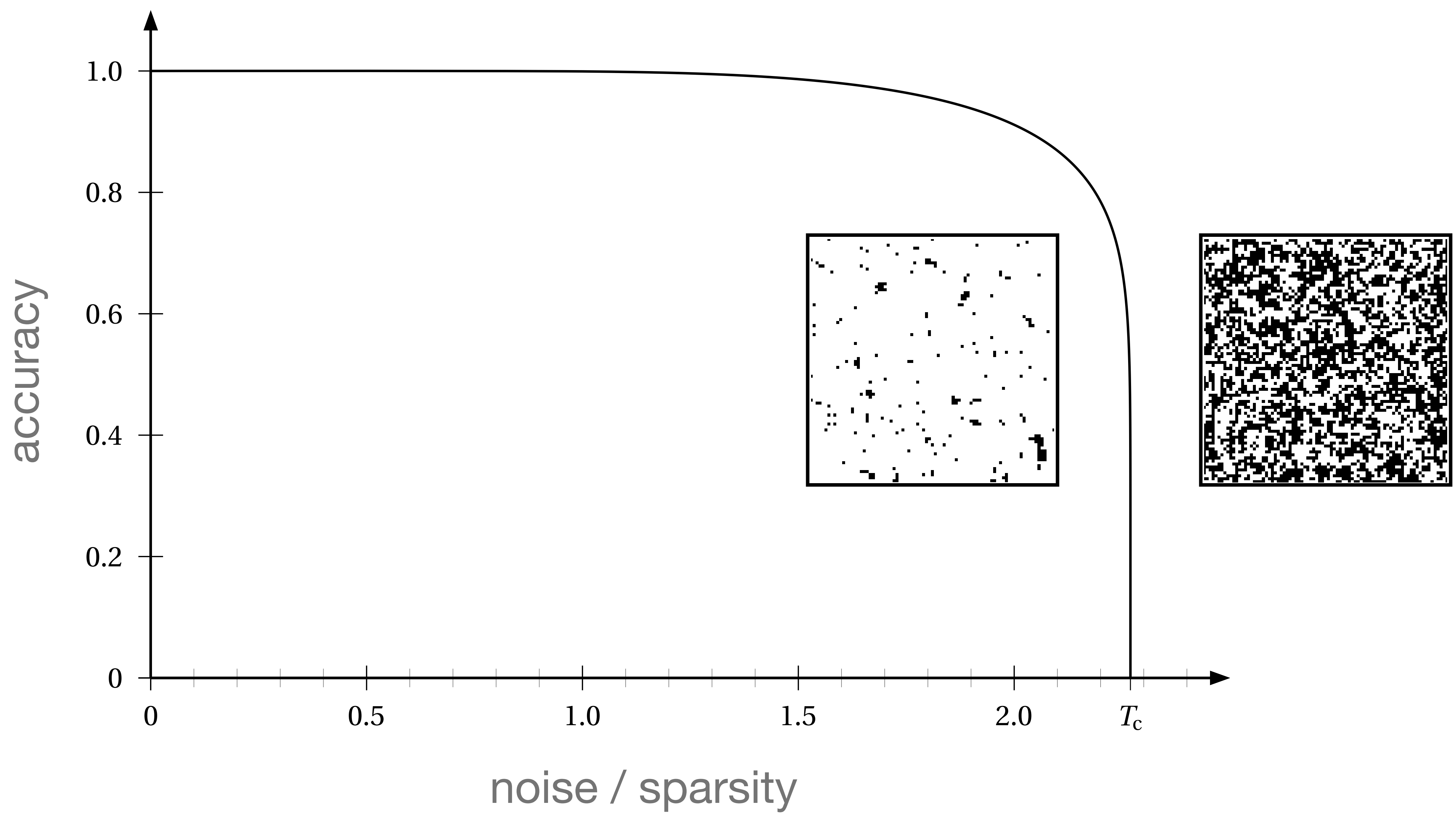
The analogy between physics and inference

but when the iron is heated beyond a critical temperature, it suddenly loses its ability to hold a magnetic field: the atoms become uncorrelated



The analogy between physics and inference

but when the iron is heated beyond a critical temperature, it suddenly loses its ability to hold a magnetic field: the atoms become uncorrelated



The analogy between physics and inference

The analogy between physics and inference

states of the atoms = hidden variables we want to infer

The analogy between physics and inference

states of the atoms = hidden variables we want to infer

fields and interactions between them = data (and our model)

The analogy between physics and inference

states of the atoms = hidden variables we want to infer

fields and interactions between them = data (and our model)

equilibrium distribution = posterior distribution given the data

The analogy between physics and inference

states of the atoms = hidden variables we want to infer

fields and interactions between them = data (and our model)

equilibrium distribution = posterior distribution given the data

temperature = noise

The analogy between physics and inference

states of the atoms = hidden variables we want to infer

fields and interactions between them = data (and our model)

equilibrium distribution = posterior distribution given the data

temperature = noise

magnetization = accuracy

The analogy between physics and inference

states of the atoms = hidden variables we want to infer

fields and interactions between them = data (and our model)

equilibrium distribution = posterior distribution given the data

temperature = noise

magnetization = accuracy

when data is too noisy or too sparse, we can suddenly lose our ability to find patterns in it

The analogy between physics and inference

states of the atoms = hidden variables we want to infer

fields and interactions between them = data (and our model)

equilibrium distribution = posterior distribution given the data

temperature = noise

magnetization = accuracy

when data is too noisy or too sparse, we can suddenly lose our ability to find patterns in it

where are these phase transitions?

The analogy between physics and inference

states of the atoms = hidden variables we want to infer

fields and interactions between them = data (and our model)

equilibrium distribution = posterior distribution given the data

temperature = noise

magnetization = accuracy

when data is too noisy or too sparse, we can suddenly lose our ability to find patterns in it

where are these phase transitions?

are there algorithms that succeed all the way up to these transitions?

The stochastic block model

The stochastic block model

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

The stochastic block model

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

$k \times k$ matrix p of connection probabilities

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}

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 $p_{rr} > p_{rs}$

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 $i \rightarrow j$ with probability p_{rs}

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

lots of games to play: can learn p, predict missing links, fill in missing labels...

The stochastic block model

nodes have discrete labels: k "groups" or types of nodes

$k \times k$ matrix p of connection probabilities

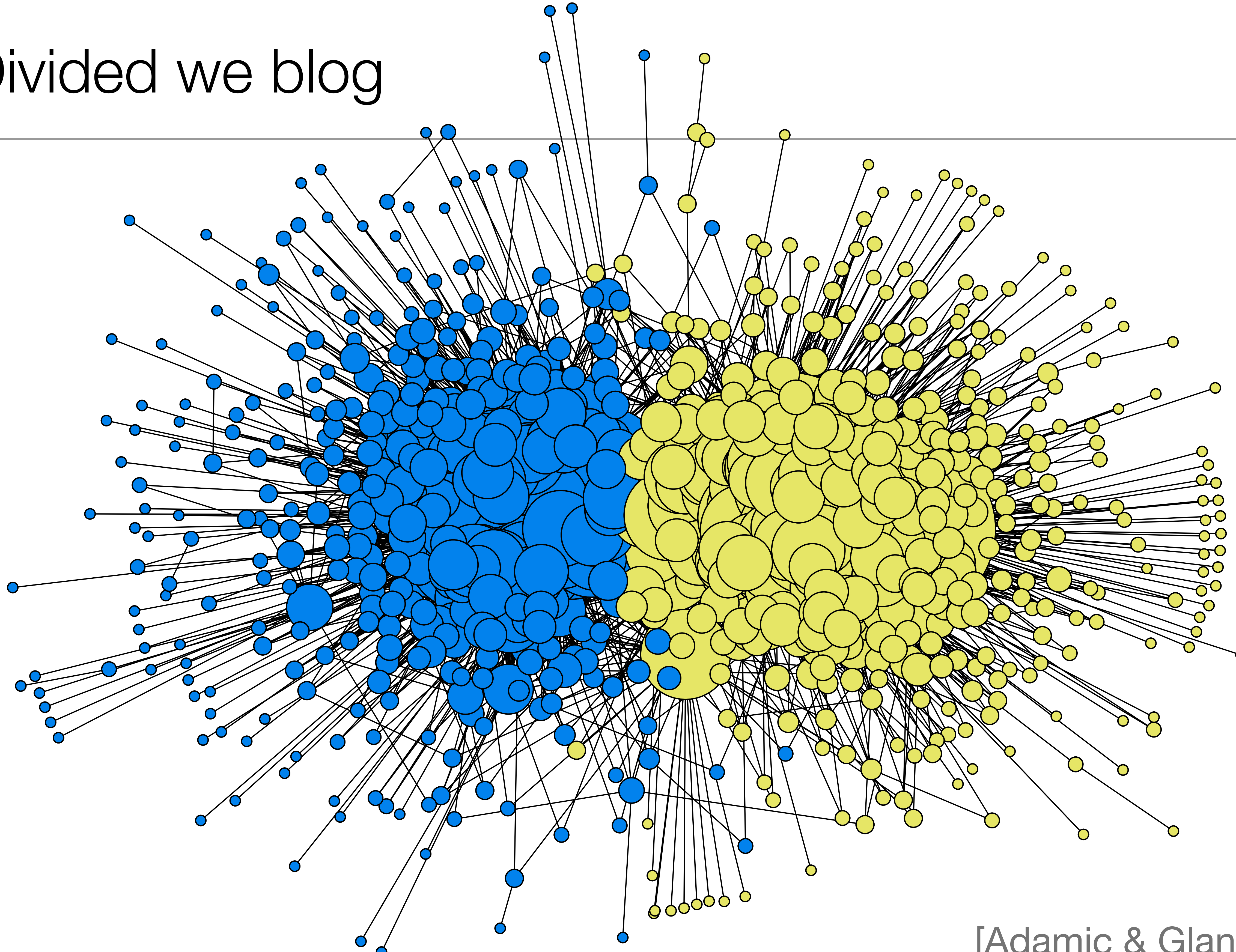
if i is type r and j is type s, there is a link $i \rightarrow j$ with probability p_{rs}

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

lots of games to play: can learn p, predict missing links, fill in missing labels...

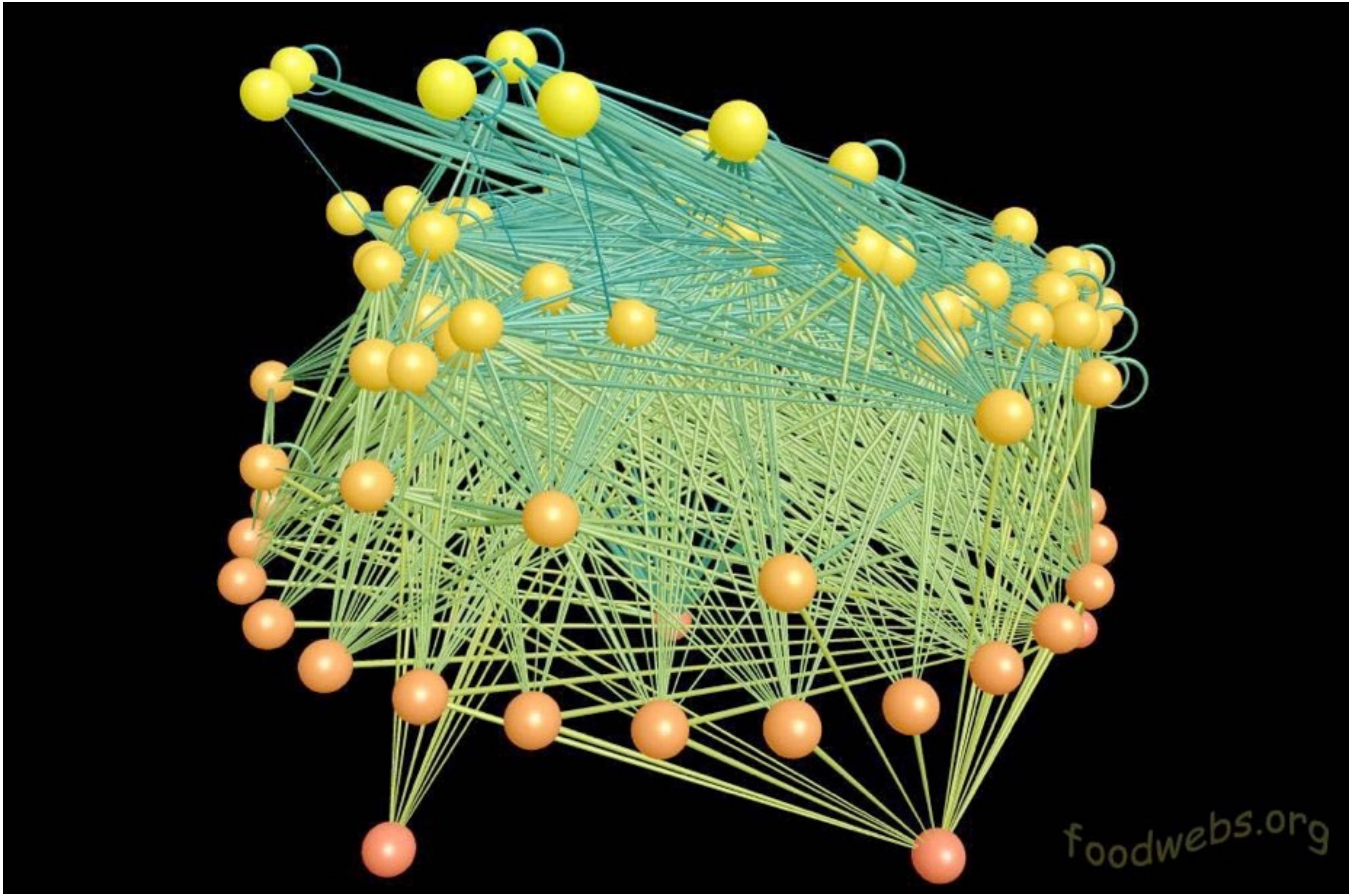
this talk: given the graph G, find the labels!

Divided we blog

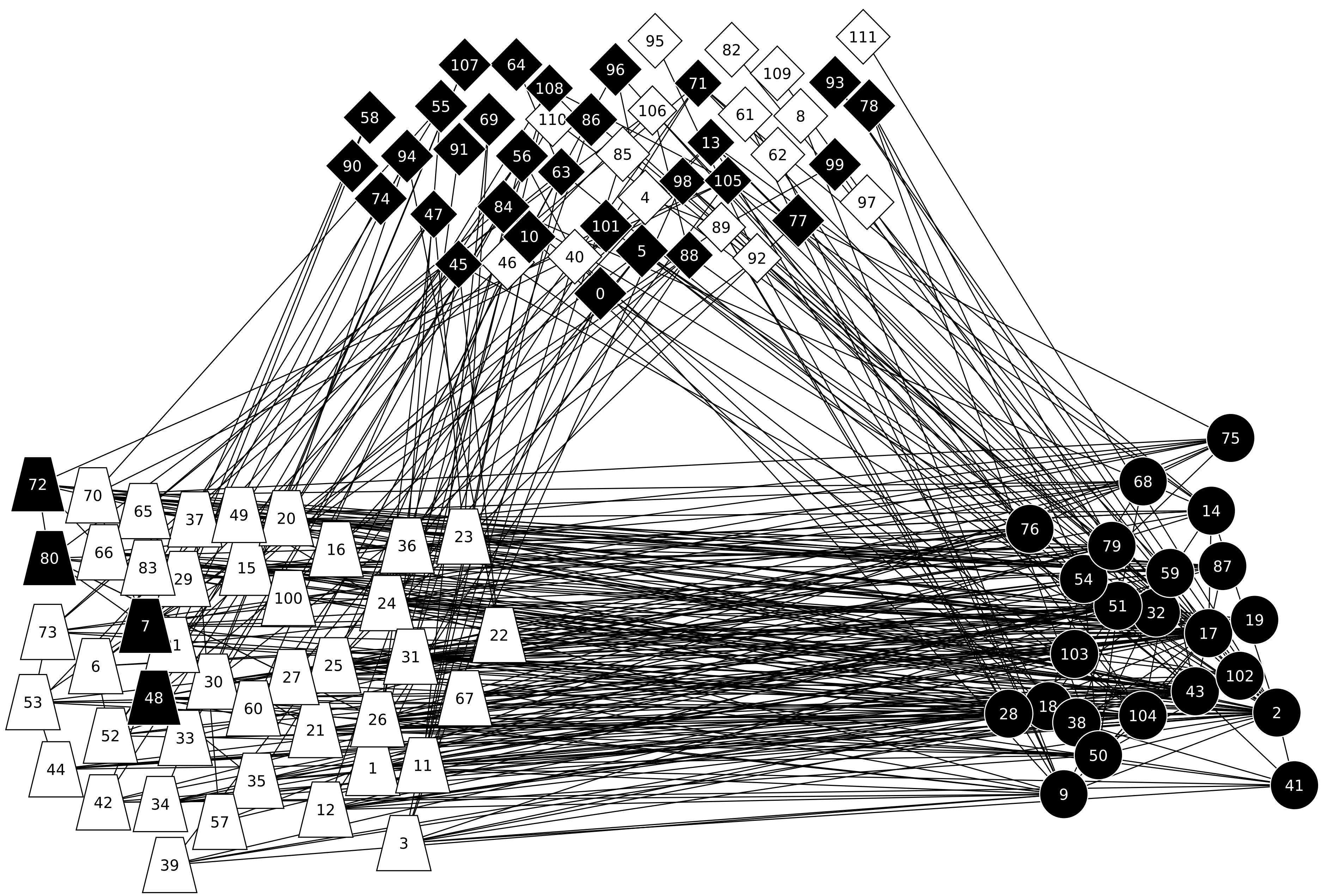


[Adamic & Glance]

Who eats whom



I record that I was born on a Friday



Some cases of interest

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

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

Inferring the block model scalably

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

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:

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]

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]

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.]

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]

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

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 model selection, and reveals phase transitions in the detectability of community structure

Likelihood and energy

Likelihood and energy

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

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

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,

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

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 $\beta=1/T$ (Boltzmann) and E is the energy,

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

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

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 β=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

Analogies with statistical physics: a glossary

probability of G given t $\quad P(G \mid t)$	

Analogies with statistical physics: a glossary

probability of *G* given *t*	$P(G\,\vert\,t)$	$\mathrm{e}^{-\beta E(t)}$	(β=1 for now)

Analogies with statistical physics: a glossary

probability of G given t	$P(G\,\|\,t)$	$e^{-\beta E(t)}$	(β=1 for now)
	$-\log P(G\,\|\,t)$		

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

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)	$\operatorname*{argmax}_t P(G \mid t)$		

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

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* “evidence”	$\sum_{t \in \{1,\ldots,k\}^n} P(G, t)$		

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* "evidence"	$\sum_{t \in \{1,\ldots,k\}^n} P(G, t)$	Z	partition function

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 "evidence"	$\sum_{t \in \{1,\ldots,k\}^n} P(G, t)$	Z	partition function
	$-\log \sum_t P(G \mid t)$		

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 "evidence"	$\sum_{t \in \{1,\ldots,k\}^n} P(G, t)$	Z	partition function
	$-\log \sum_t P(G \mid t)$	$F = -\log Z$	free energy

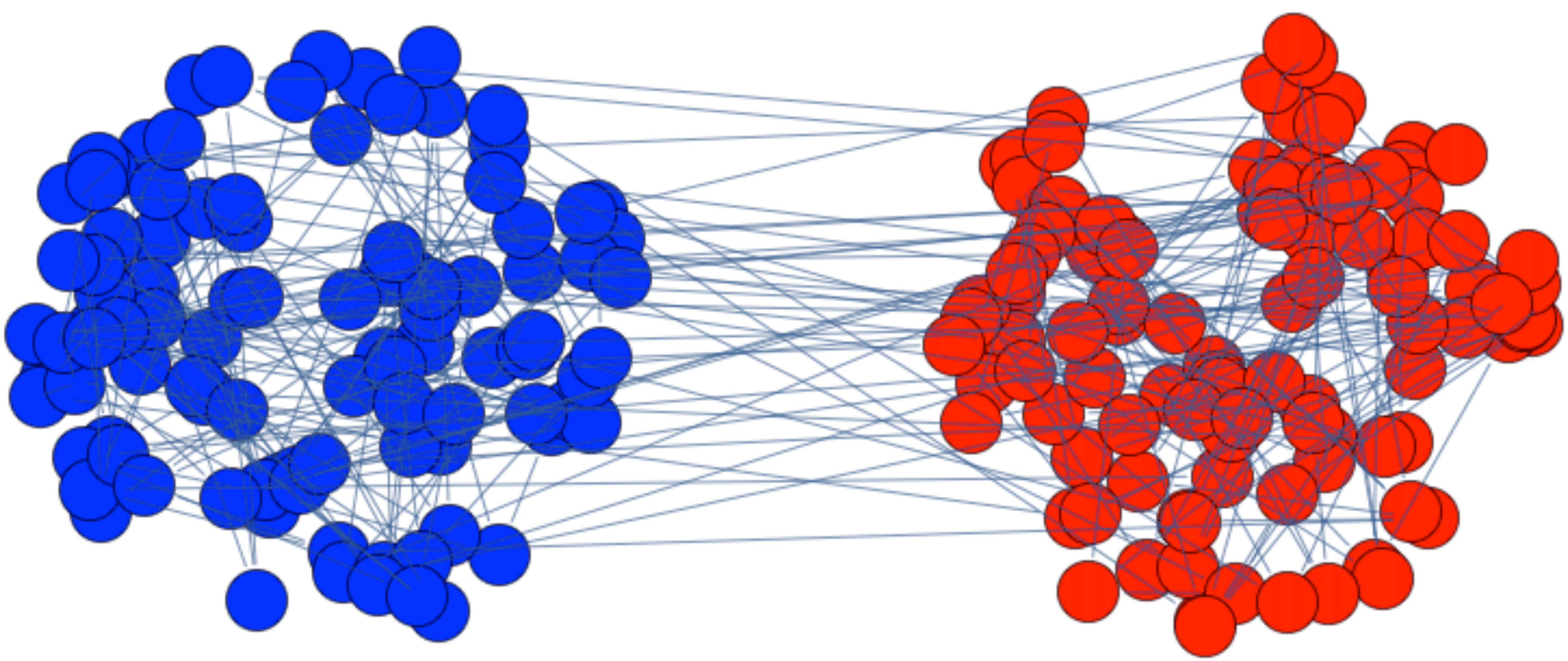
Analogies with statistical physics: a glossary

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)	$\underset{t}{\operatorname{argmax}}\, P(G\mid t)$	$\underset{t}{\operatorname{argmin}}\, E(t)$	ground state
total probability of *G* “evidence”	$\sum_{t\in\{1,\dots,k\}^n} P(G,t)$	Z	partition function
	$-\log \sum_t P(G\mid t)$	$F=-\log Z$	free energy
Gibbs distribution	$P(t\mid G)=\dfrac{P(G\mid t)}{\sum_{t'} P(G\mid t')}$		

Analogies with statistical physics: a glossary

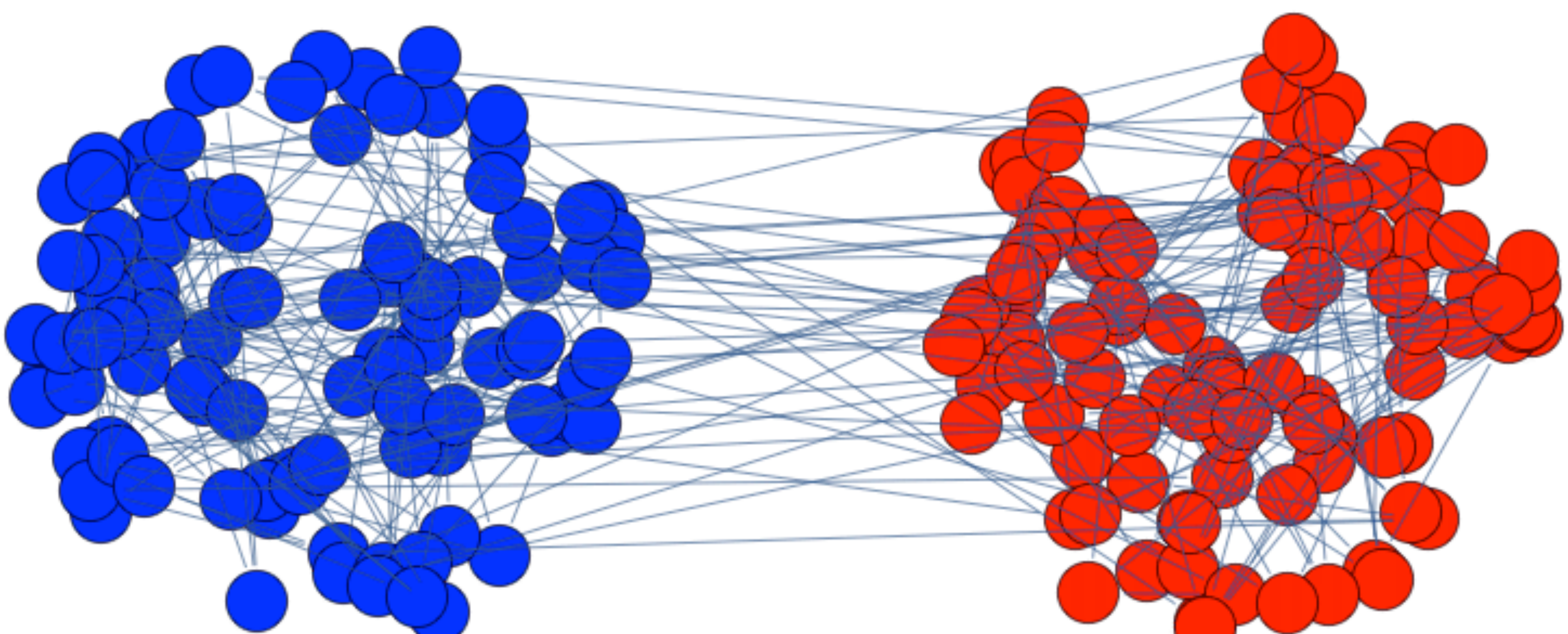
probability of *G* given *t*	$P(G \mid t)$	$e^{-\beta E(t)}$	(β=1 for now)
	$-\log P(G \mid t)$	$E(t)$	energy
most likely labeling (MAP)	$\underset{t}{\operatorname{argmax}} P(G \mid t)$	$\underset{t}{\operatorname{argmin}} E(t)$	ground state
total probability of *G* “evidence”	$\sum_{t \in \{1,\ldots,k\}^n}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

What's the best labeling?



What’s the best labeling?

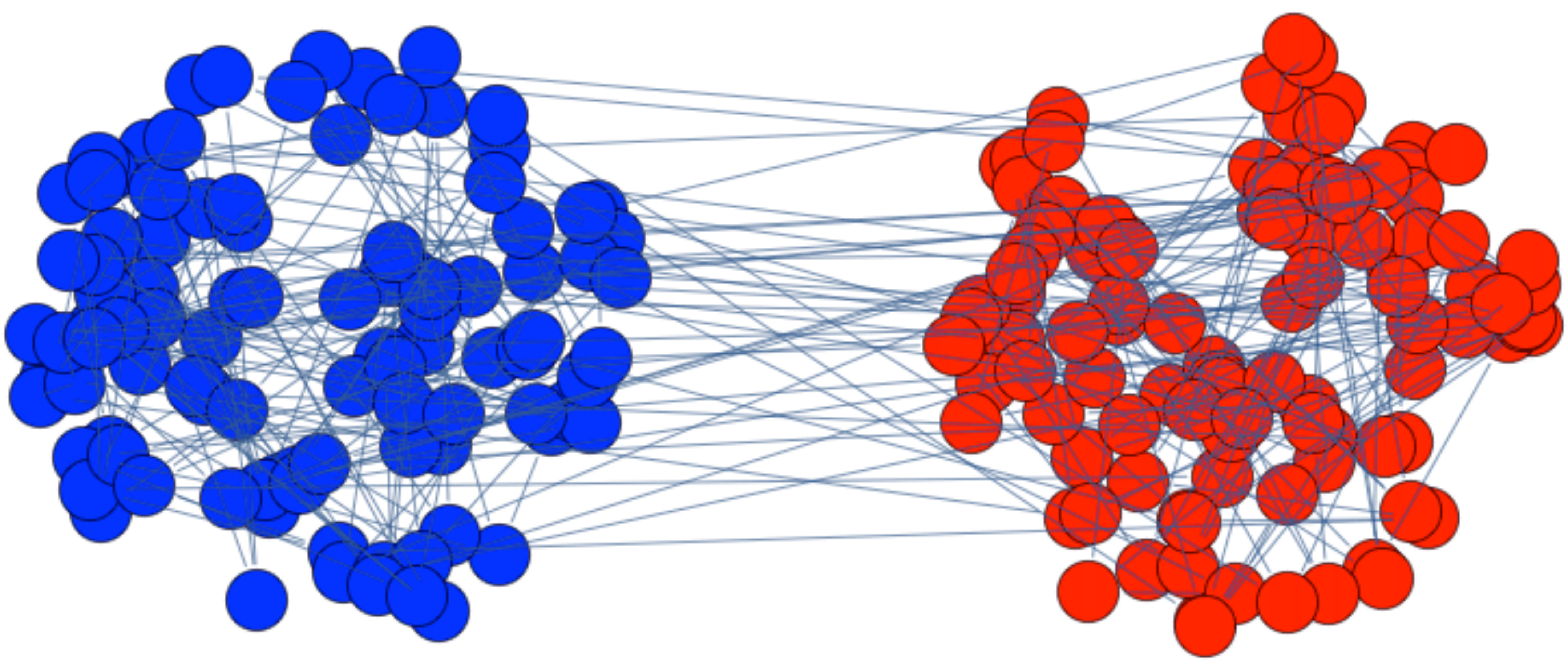
the most likely labeling is the *ground state*: it maximizes $P(G|t)$



What’s the best labeling?

the most likely labeling 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]

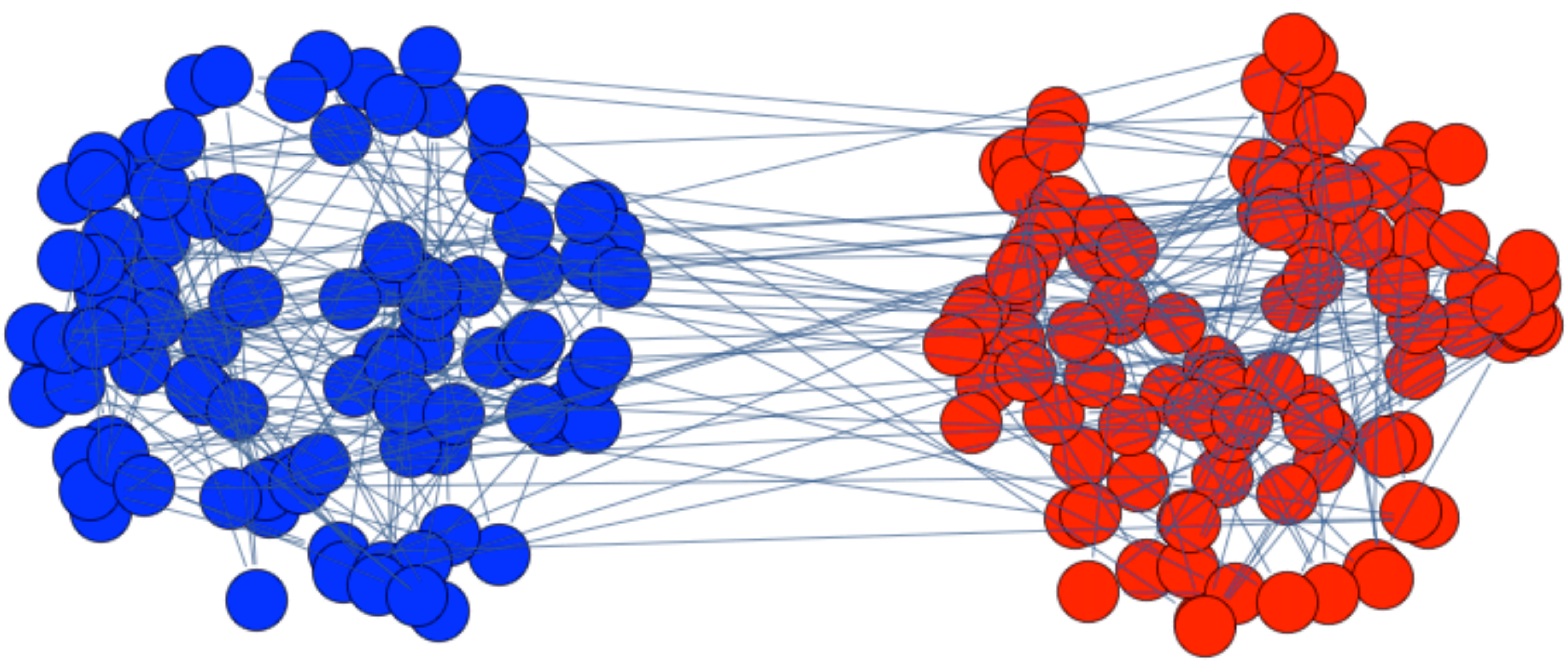


What's the best labeling?

the most likely labeling 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!

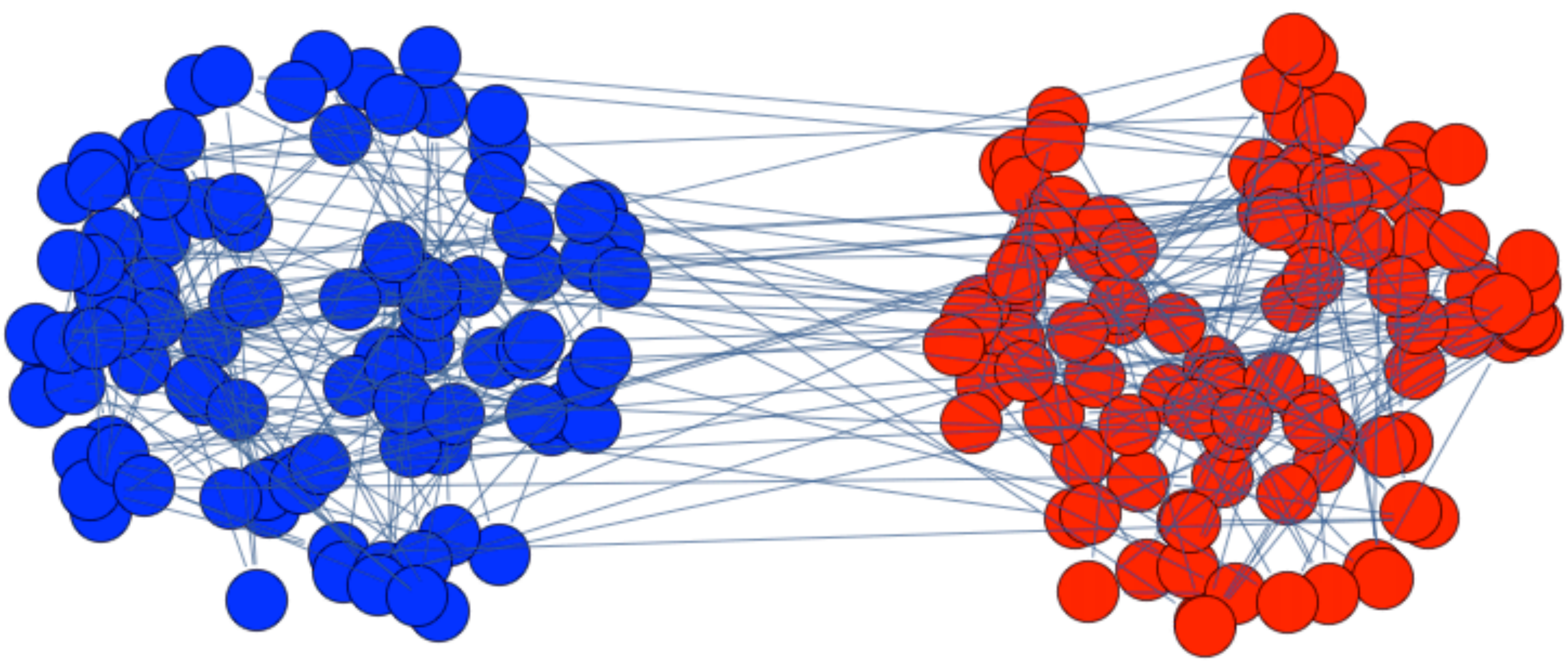


What's the best labeling?

the most likely labeling 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!



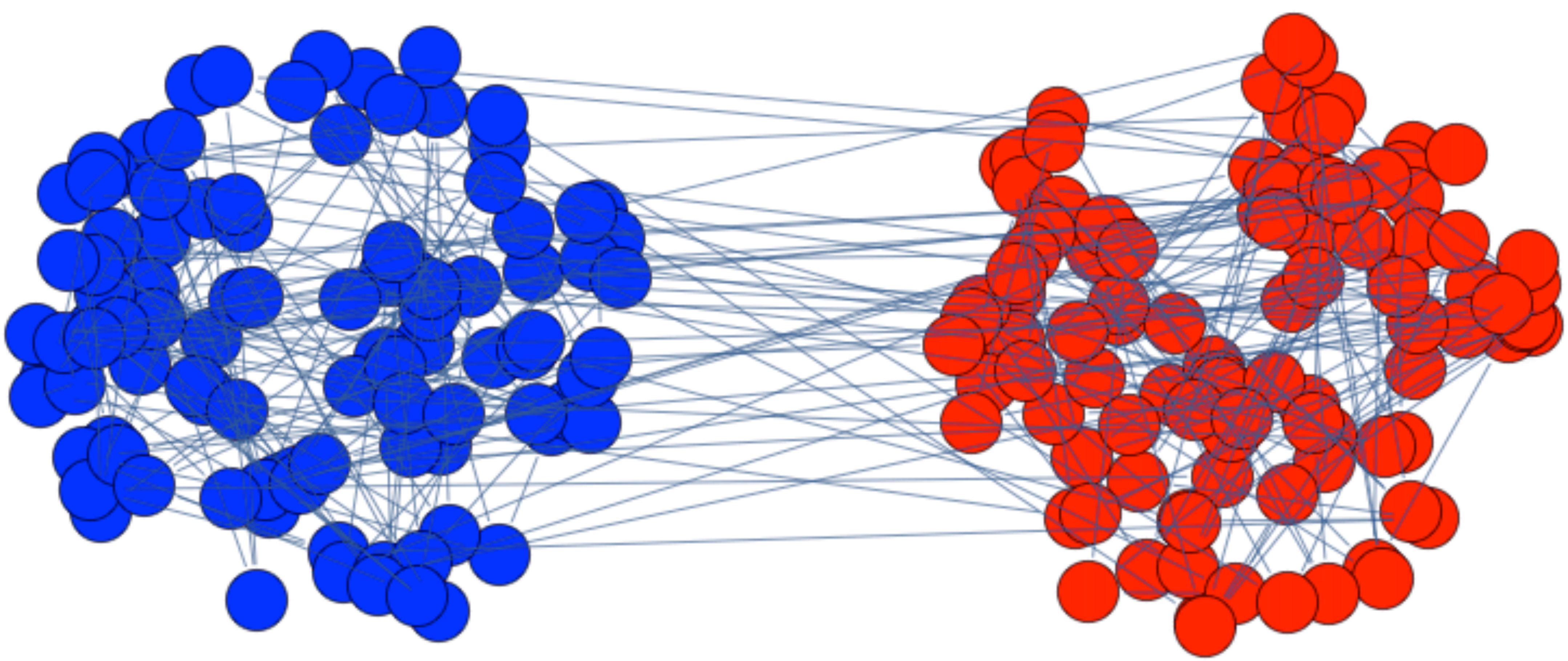
What's the best labeling?

the most likely labeling 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

Statistical significance vs. overfitting

we don’t just want the best fit!

Statistical significance vs. overfitting

we don’t just want the best fit!

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

Statistical significance vs. overfitting

we don't just want the best fit!

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

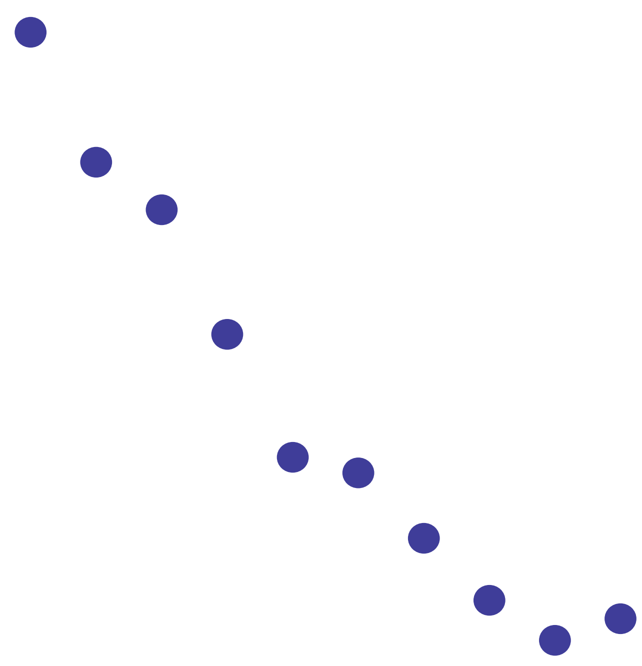
sometimes the patterns we find aren't really there:

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:

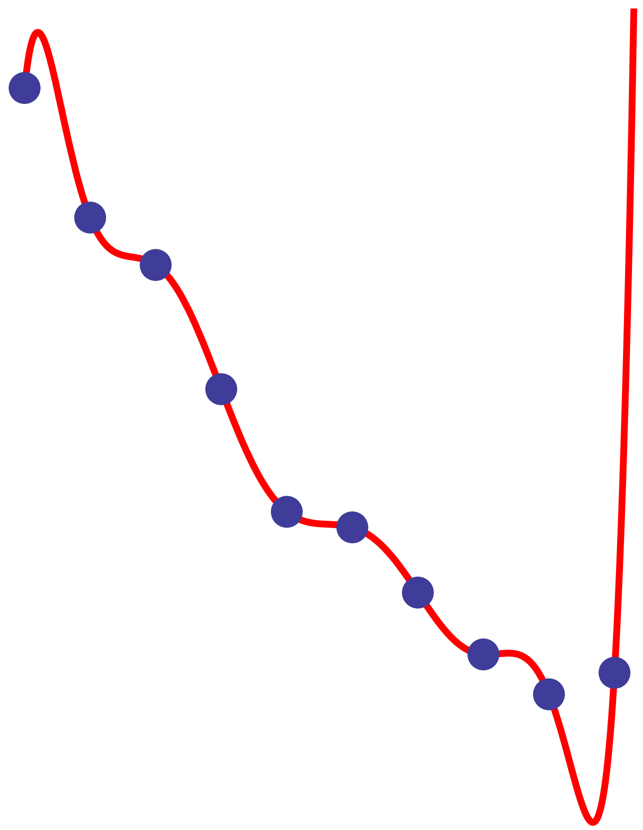


Statistical significance vs. overfitting

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:

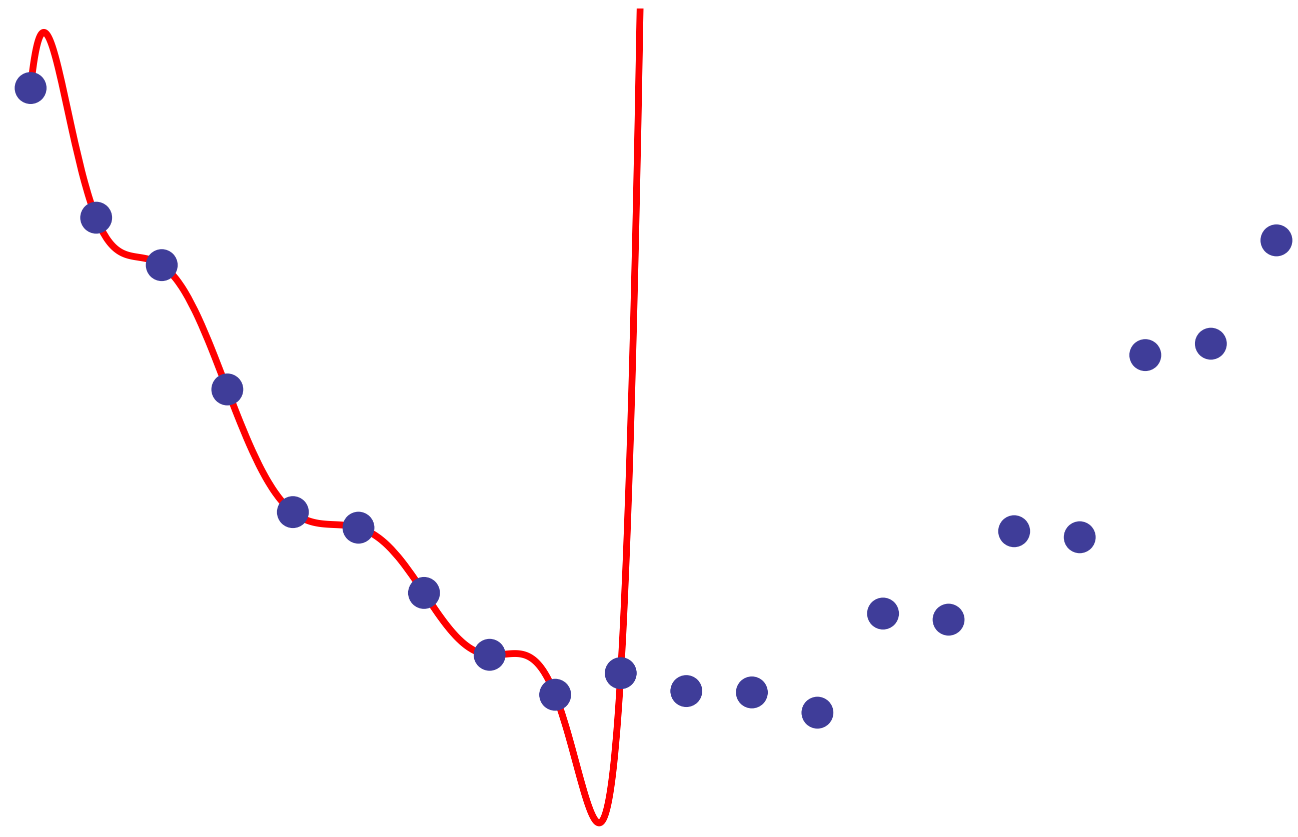


Statistical significance vs. overfitting

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:

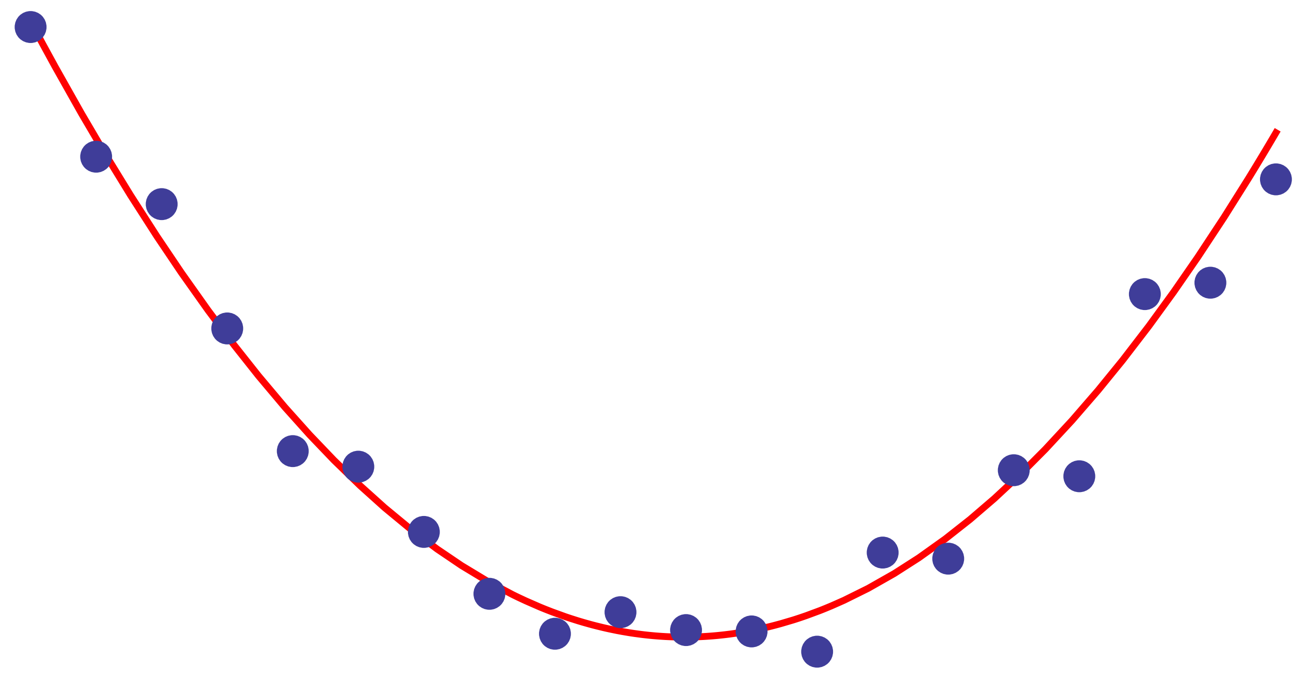


Statistical significance vs. overfitting

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:

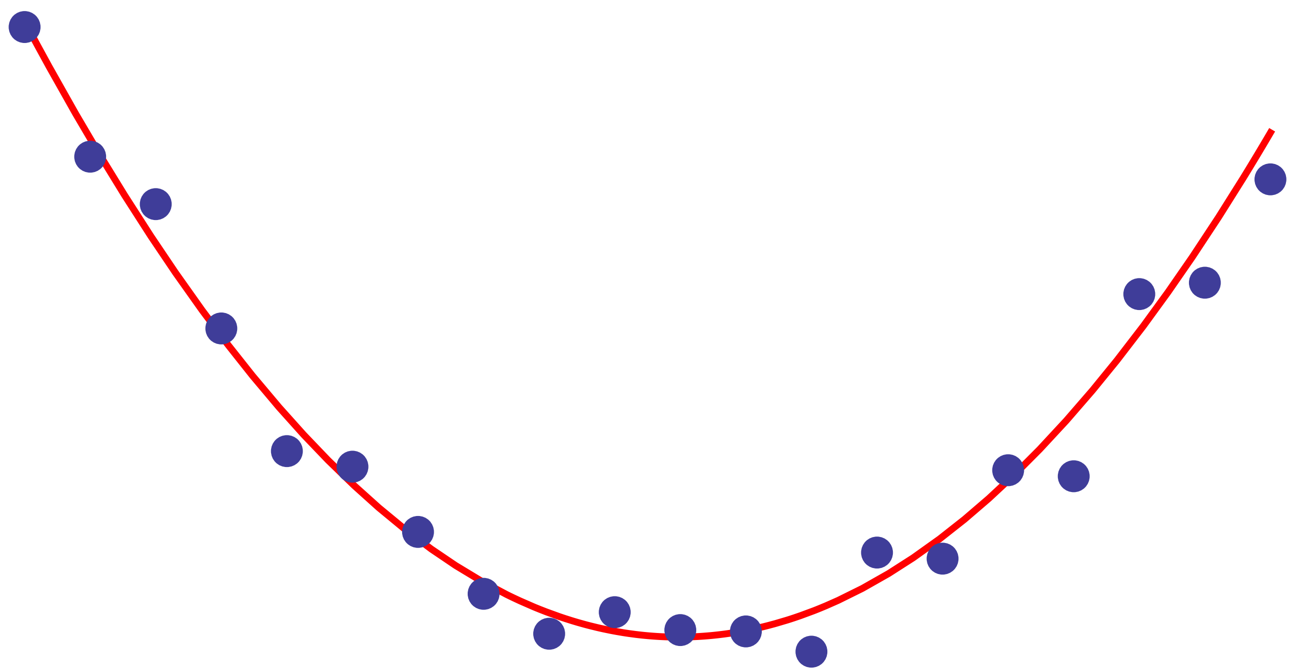


Statistical significance vs. overfitting

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

Statistical significance and the energy landscape

Statistical significance and the energy landscape

explore the landscape of models, not just the best one

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

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

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?

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 agreement

Consensus and marginals

Consensus and marginals

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

Consensus and marginals

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

assign each node to its most-likely label

Consensus and marginals

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

assign each node to its most-likely label

this maximizes the expected fraction of nodes labeled correctly

Consensus and marginals

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

assign each node to its most-likely label

this maximizes the expected fraction of nodes labeled correctly

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

Consensus and marginals

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

assign each node to its most-likely label

this 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

Model selection and free energy

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

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:

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

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

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, $F = E - TS$

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)

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, $F = E - TS$

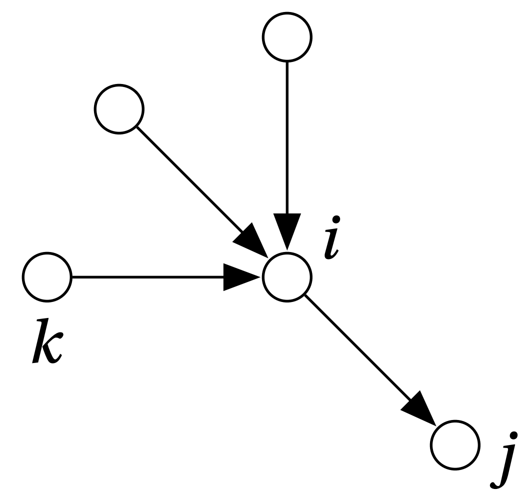
minimizing F = low energy (high probability) + high entropy (many good solutions)

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

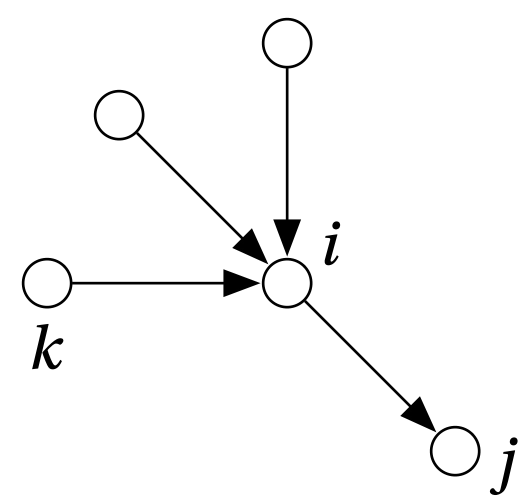
But how can we compute marginals and free energies?

But how can we compute marginals and free energies?
Monte Carlo is too slow!

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

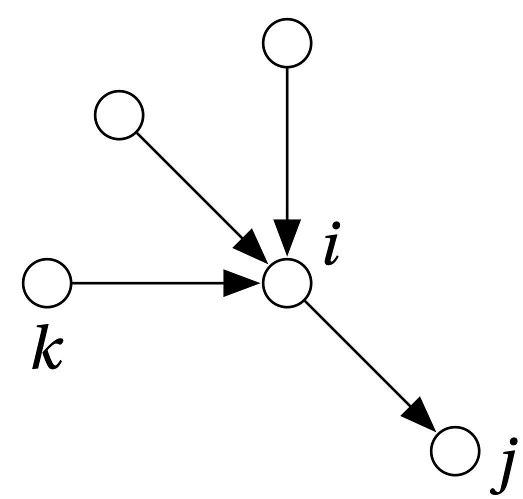


Belief propagation (a.k.a. the cavity method)



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

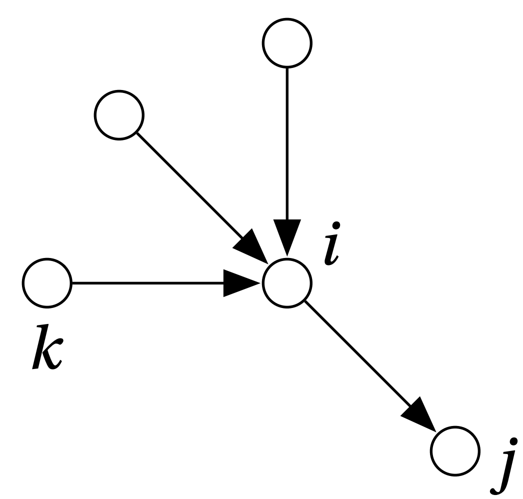
Belief propagation (a.k.a. the cavity method)



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

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

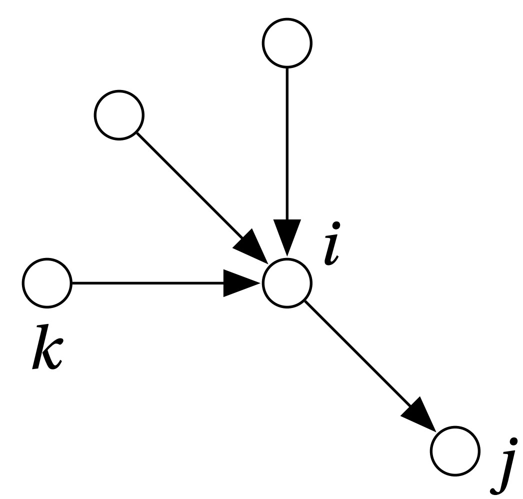


each node *i* sends a “message” to each of its neighbors *j*, giving *i*’s marginal distribution based on its other neighbors *k*

avoids an “echo chamber” between pairs of nodes

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

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



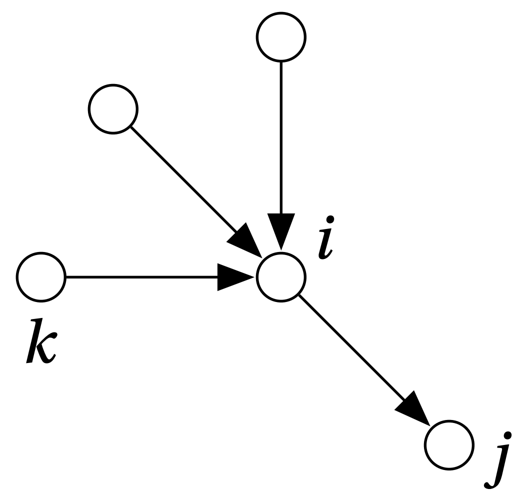
each node *i* sends a “message” to each of its neighbors *j*, giving *i*’s marginal distribution based on its other neighbors *k*

avoids an “echo chamber” between pairs of nodes

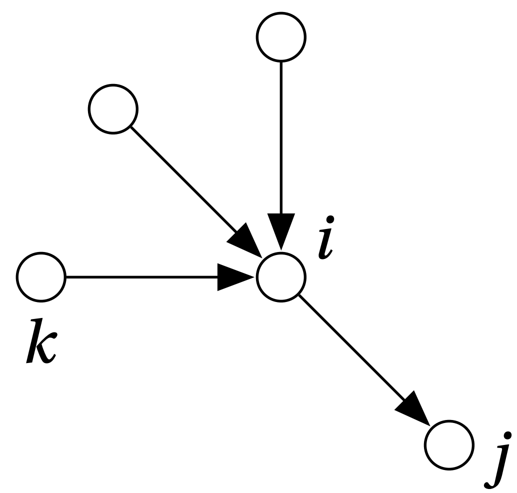
update until we reach a fixed point (how many iterations? does it converge?)

fixed point returns estimated marginals and the Bethe free energy

Updating the beliefs

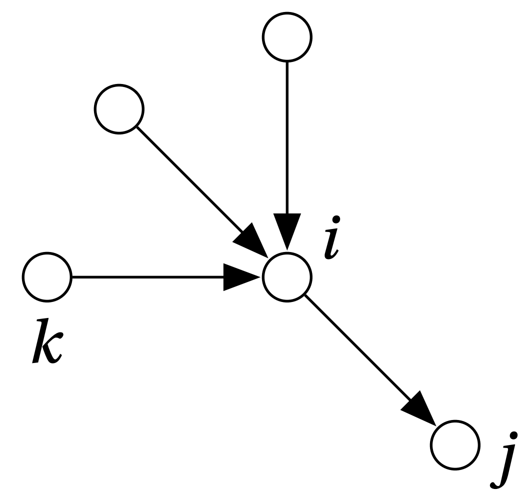


Updating the beliefs



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

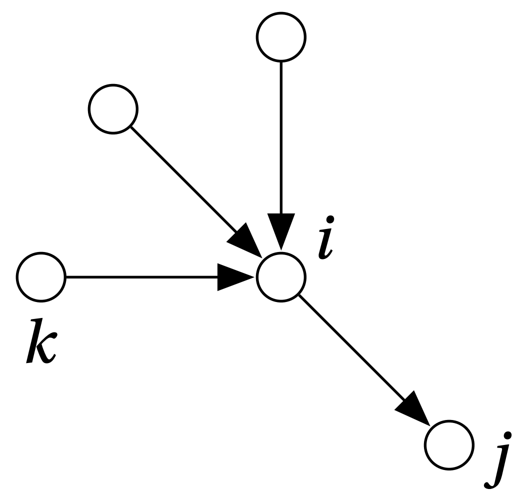
Updating the beliefs



conditional independence

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

Updating the beliefs

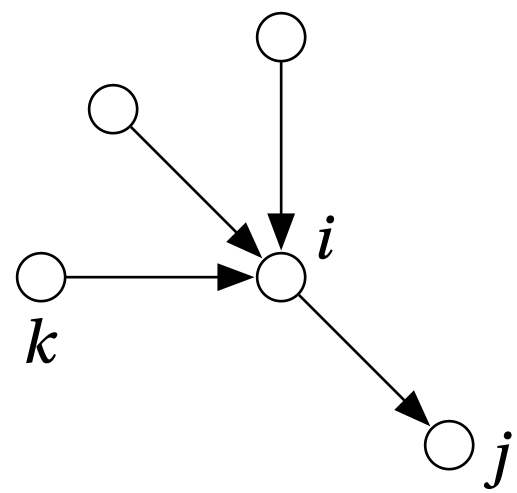


WARNING: EXACT ONLY ON TREES

conditional independence

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

Updating the beliefs



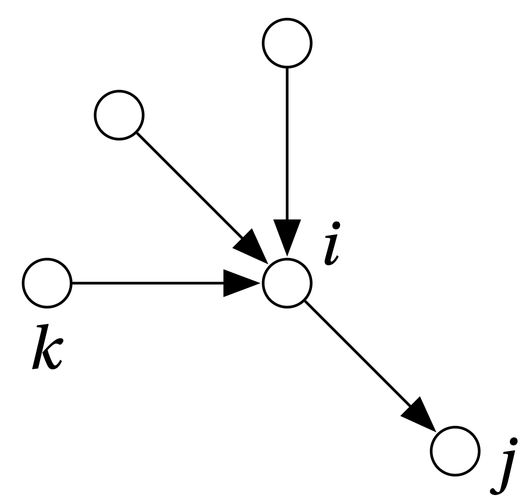
WARNING: EXACT ONLY ON TREES

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!

Updating the beliefs



WARNING: EXACT ONLY ON TREES

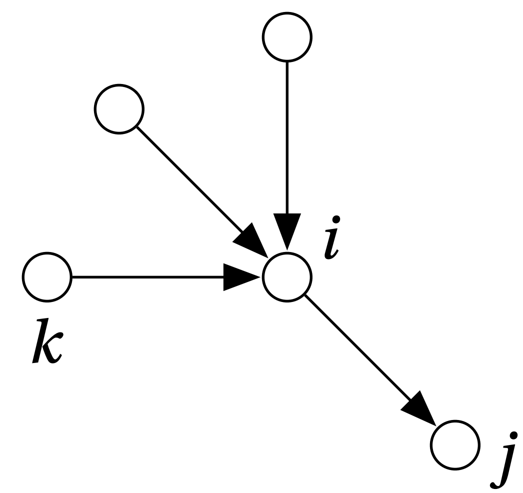
conditional independence

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

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*

Updating the beliefs



WARNING: EXACT ONLY ON TREES

conditional independence

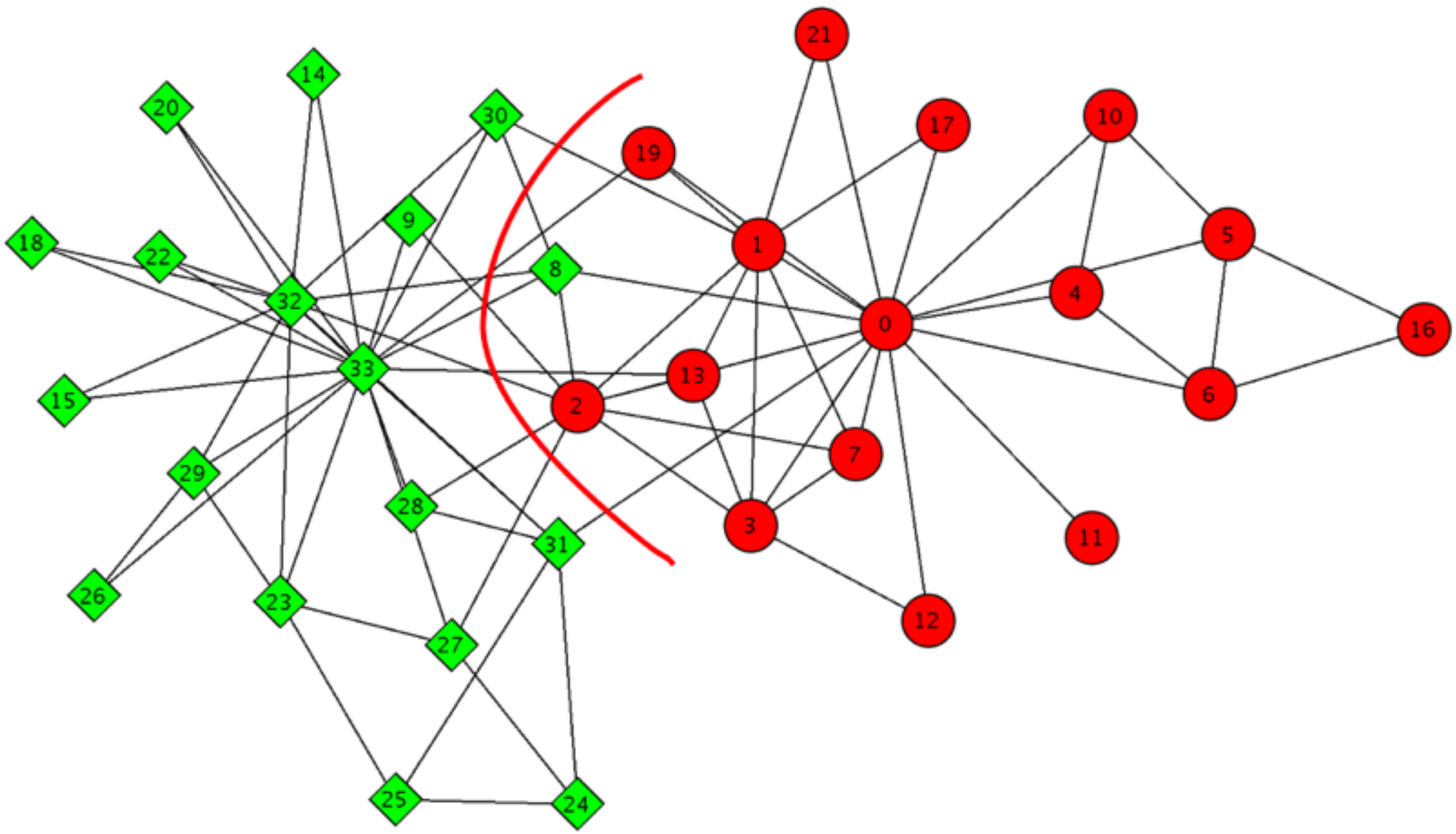
$$\mu_s^{i\to j} = \frac{1}{Z^{i\to j}} q_s \prod_{\substack{k\neq j \\ (i,k)\in E}} \sum_r \mu_r^{k\to i} p_{rs} \times \prod_{\substack{k\neq j \\ (i,k)\notin E}} \sum_r \mu_r^{k\to i} (1-p_{rs})$$

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

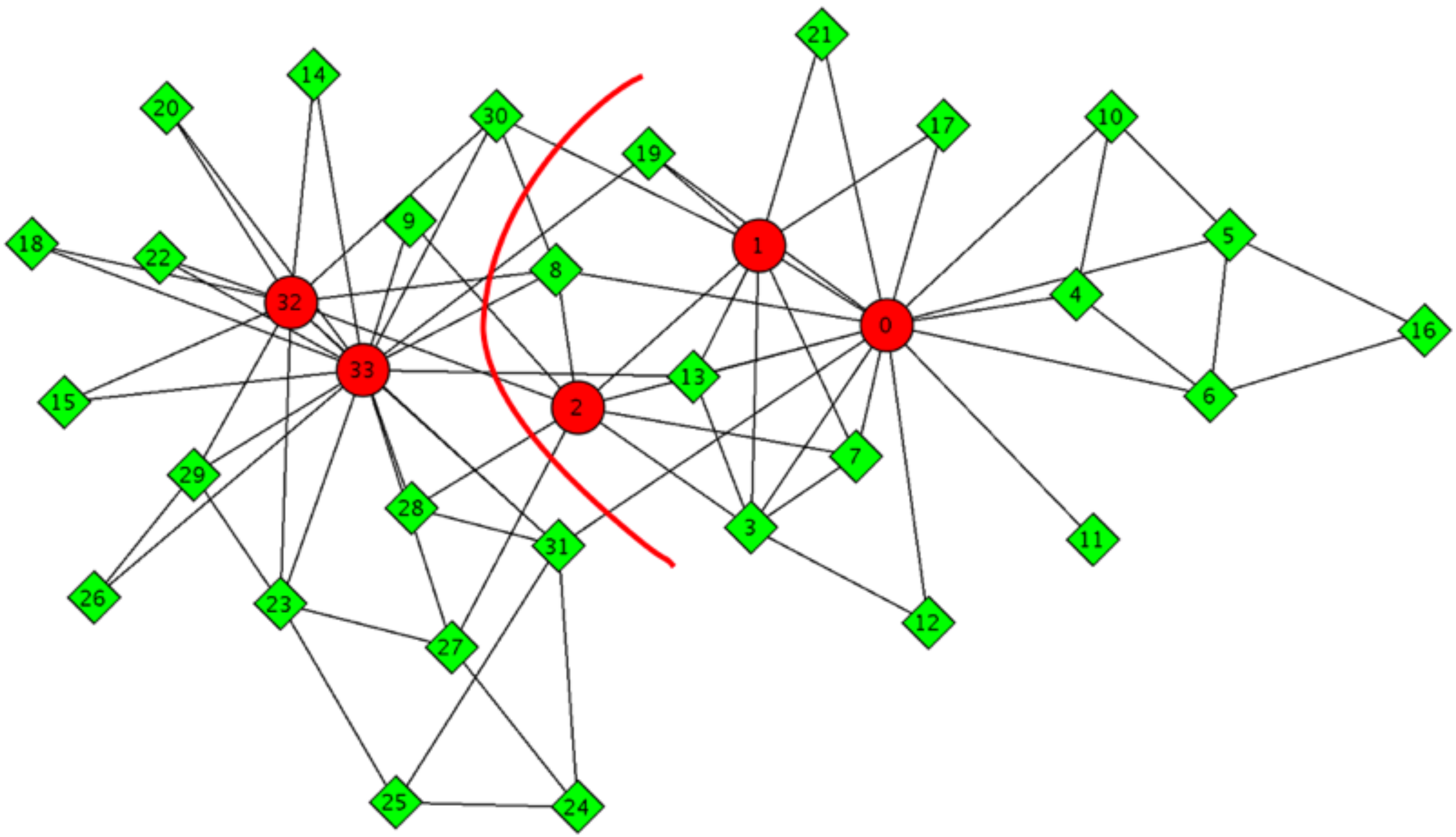
sparse case: can simplify by assuming that $\mu_r^{k\to i} = \mu_r^k$ for all non-neighbors *i*

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

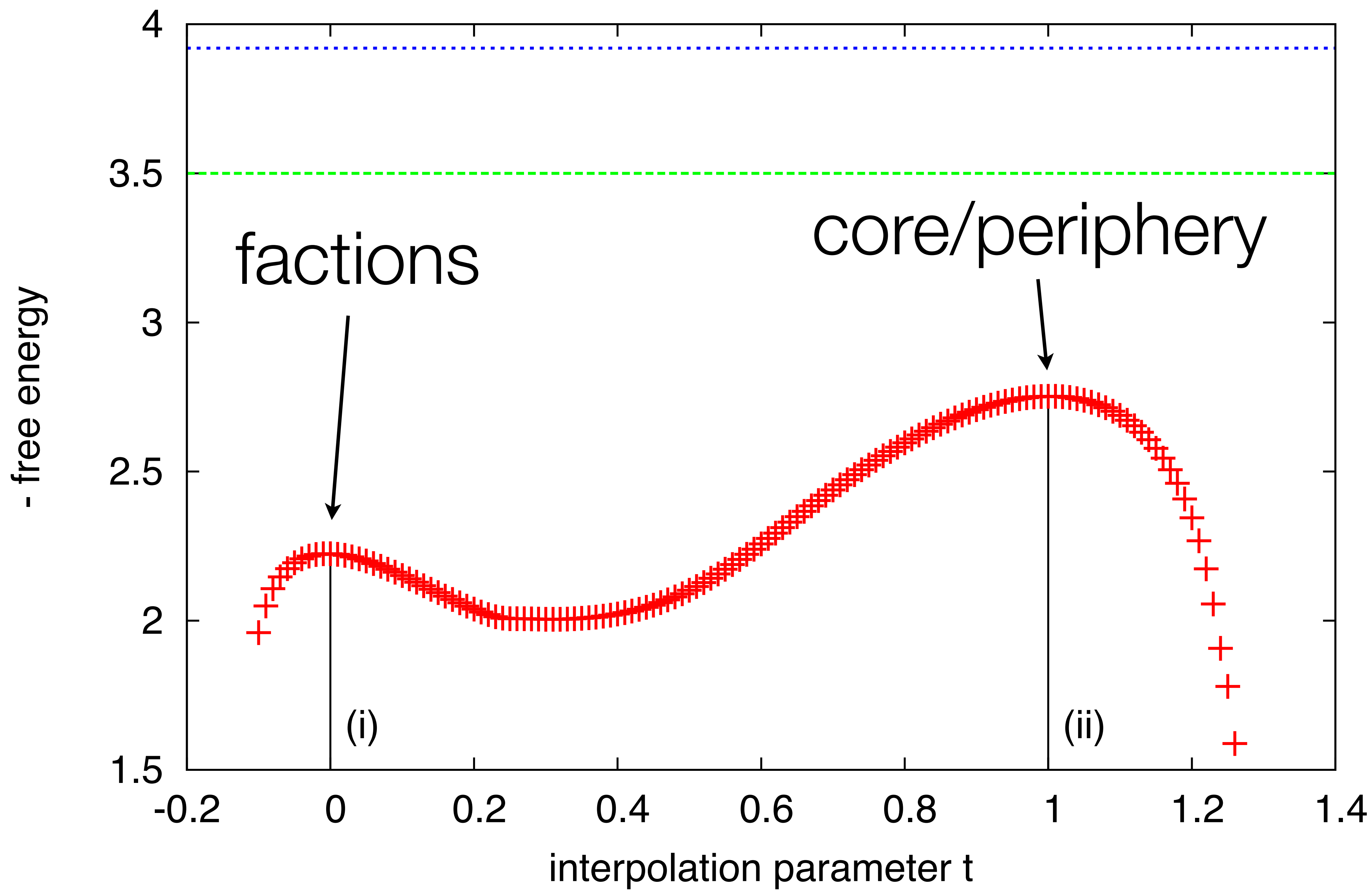
Zachary's Karate Club: Two factions



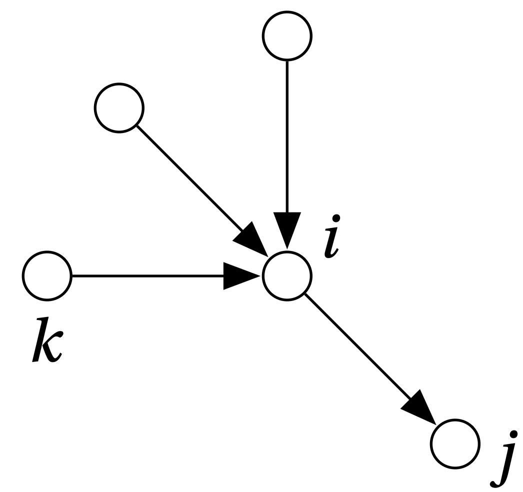
Zachary's Karate Club: Core-periphery



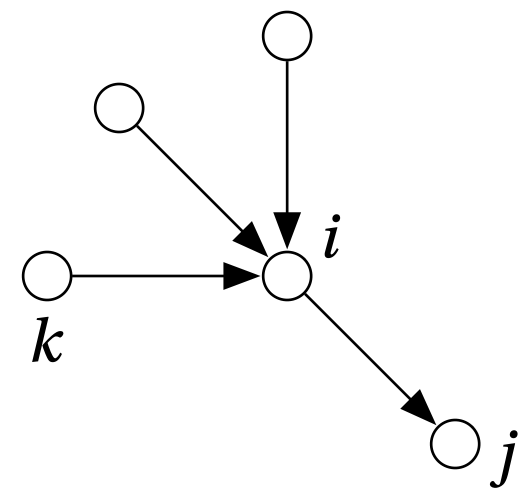
Two local optima in free energy



The double life of Belief Propagation

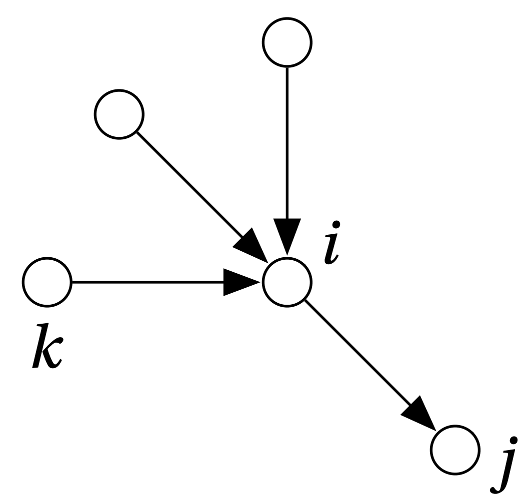


The double life of Belief Propagation



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

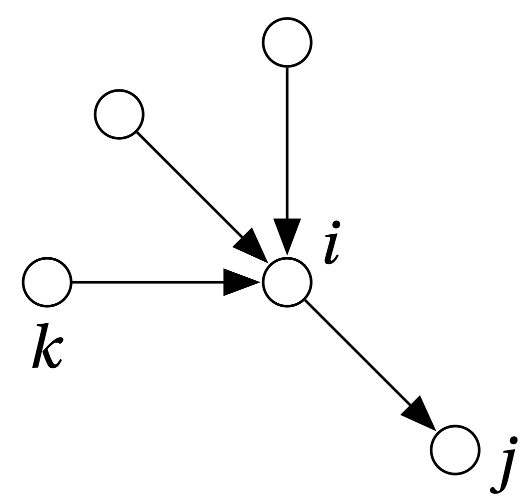
The double life of Belief Propagation



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

The double life of Belief Propagation



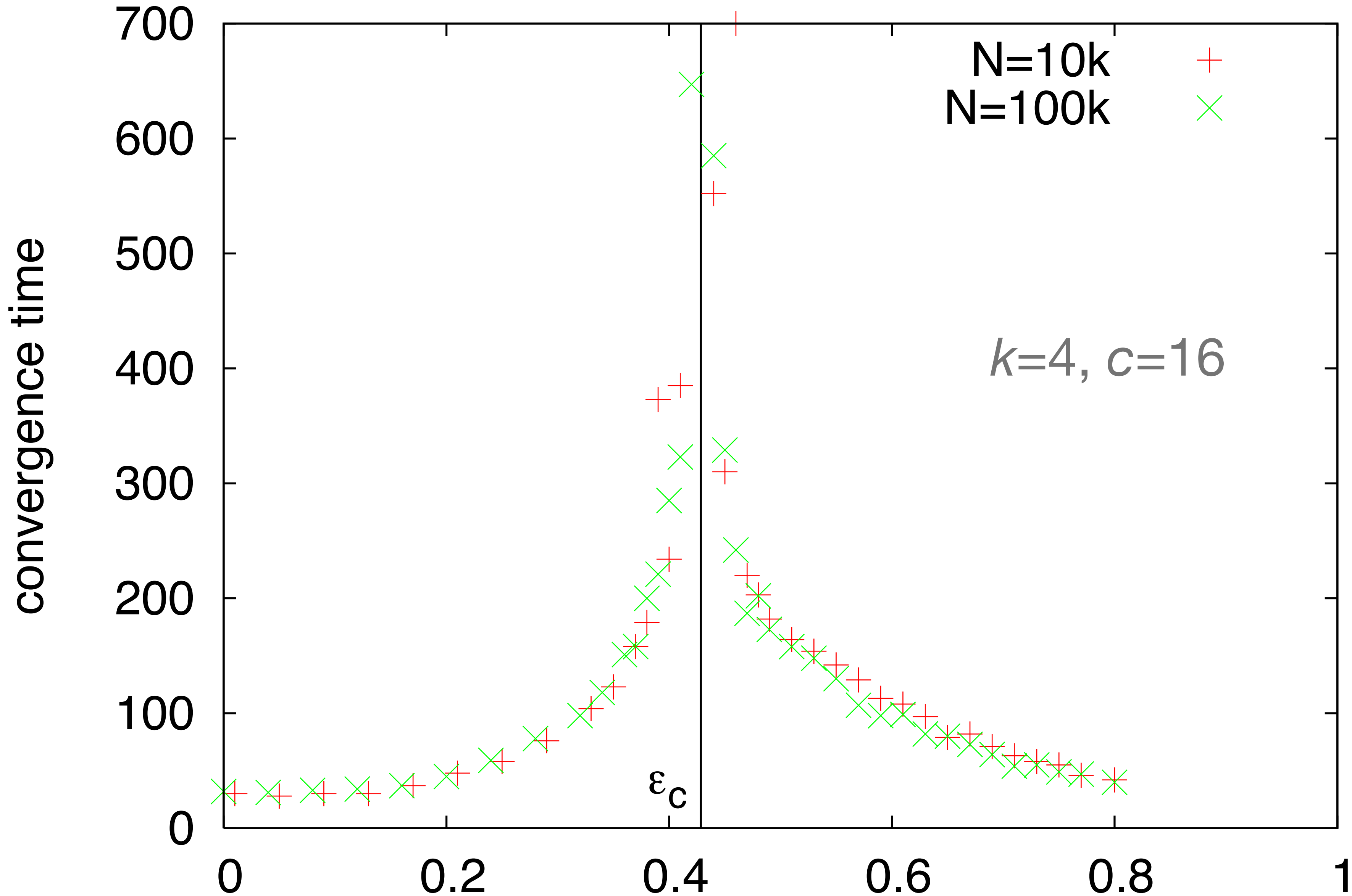
BP is a fast algorithm we can run on real networks...

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

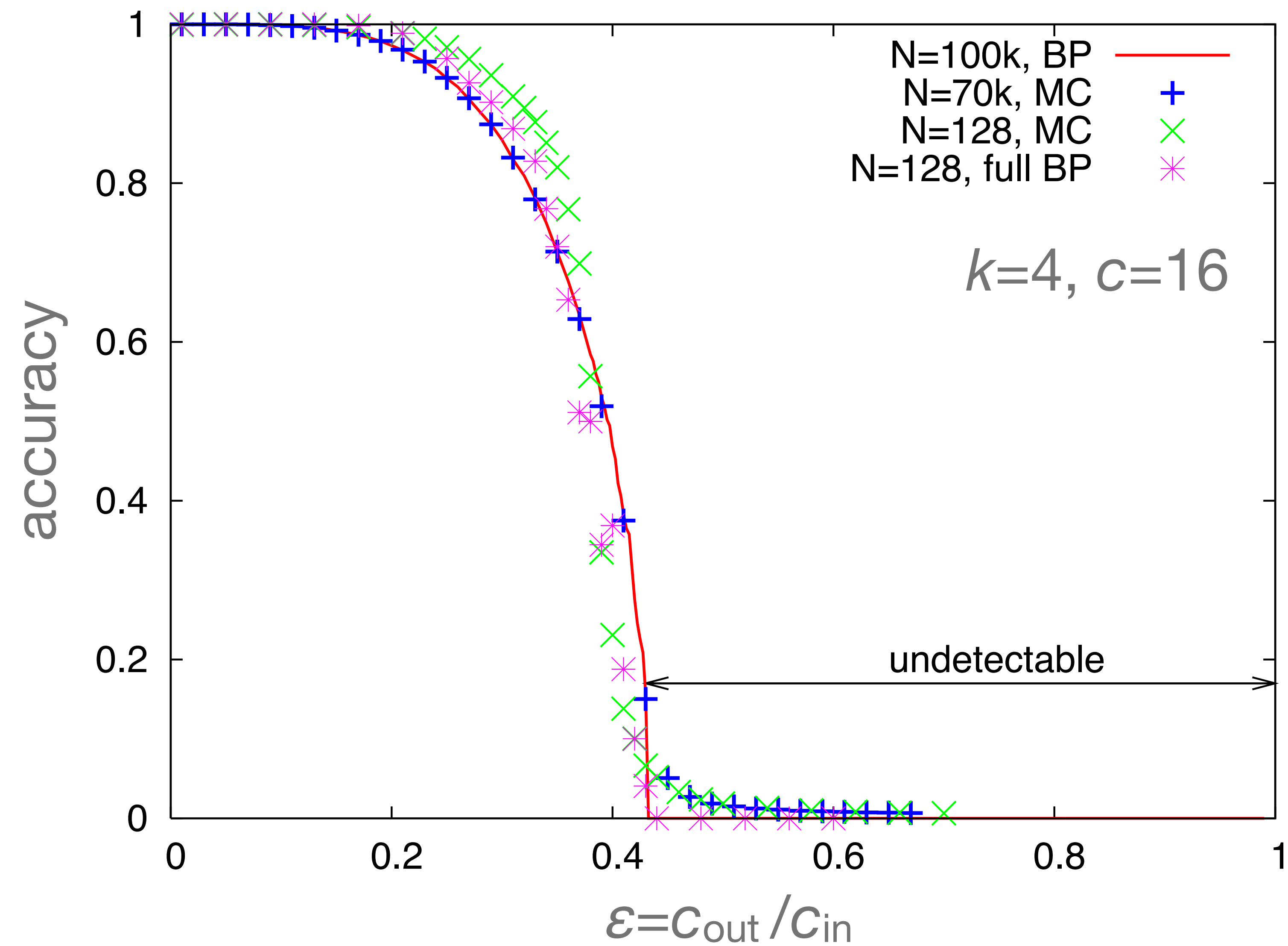
BP convergence time: nearly size-independent, but with critical slowing down at a phase transition

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



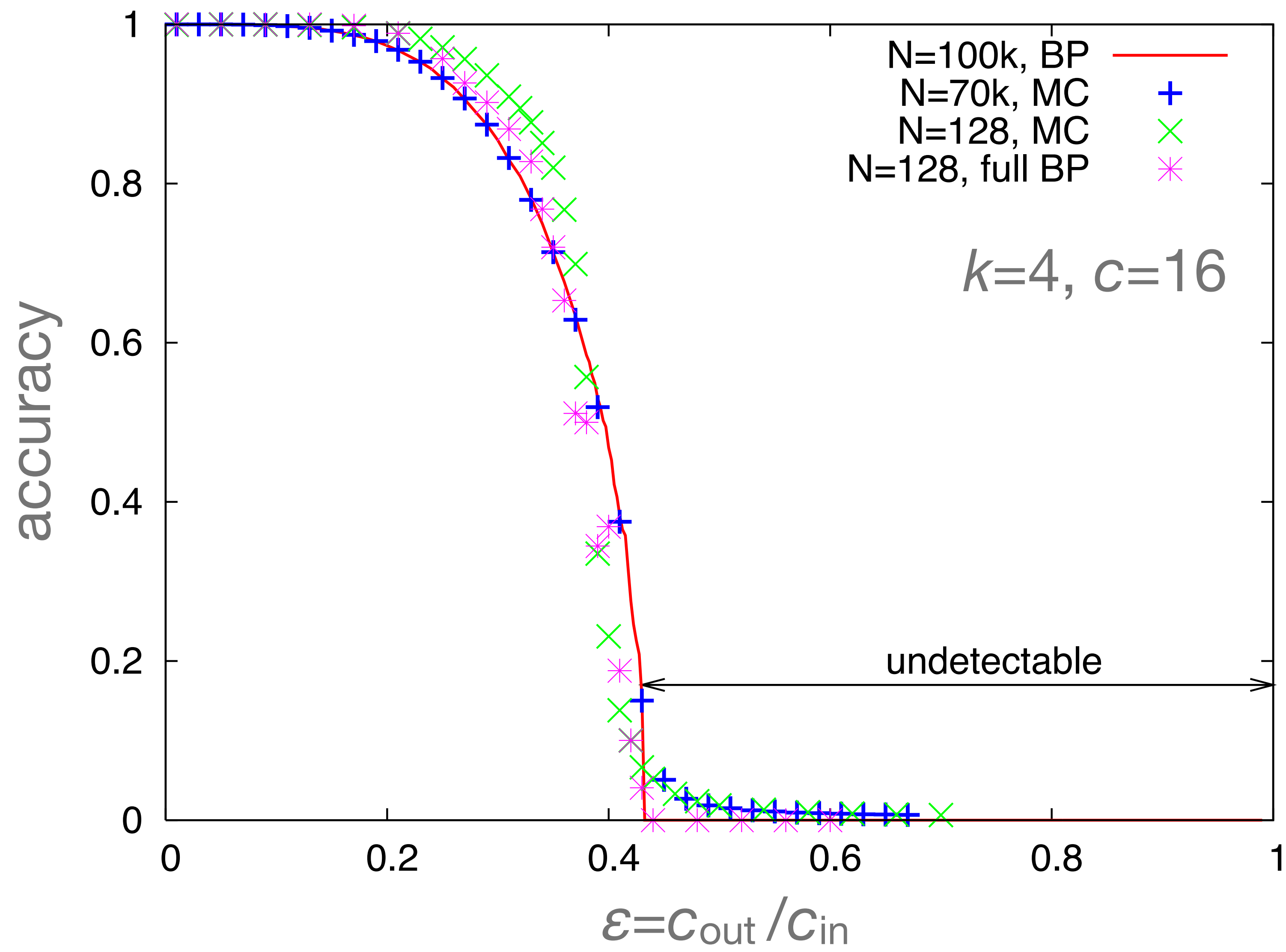
strong communities $\epsilon=c_{out}/c_{in}$ random graph

A phase transition: detectable to undetectable communities



A phase transition: detectable to undetectable communities

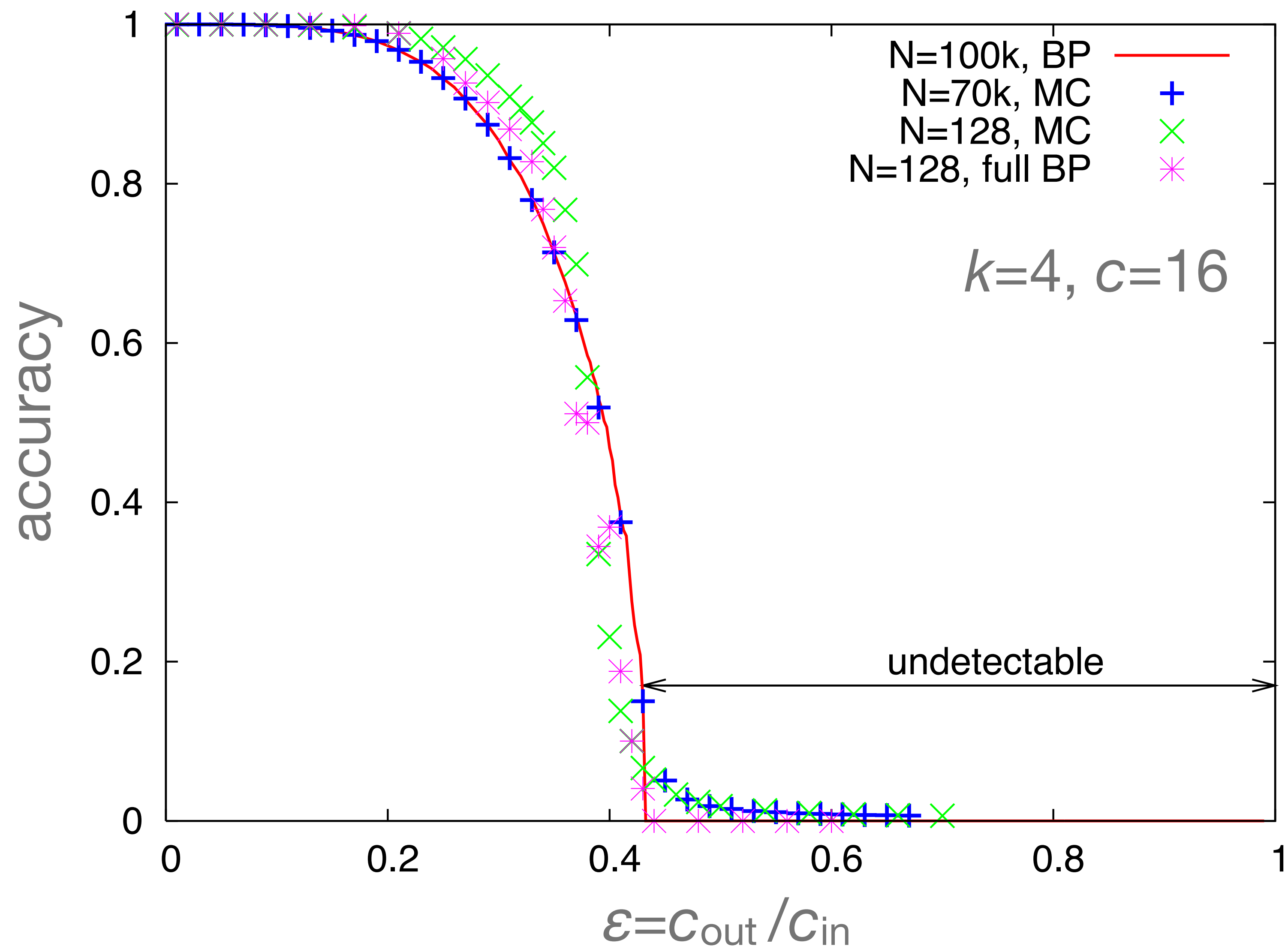
when c_{out} / c_{in} is small enough,
BP can find the communities



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!



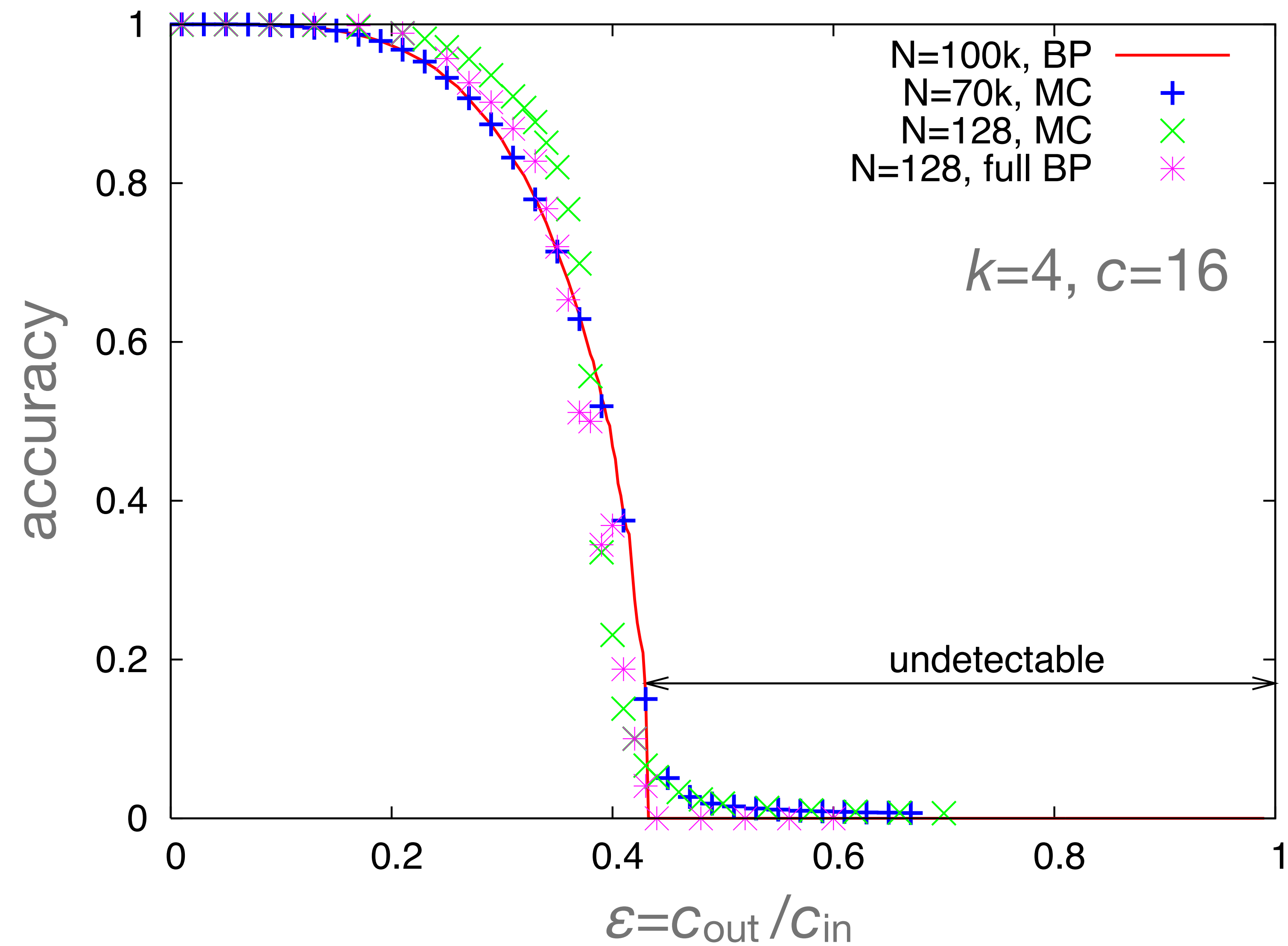
A phase transition: detectable to undetectable communities

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

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



A phase transition: detectable to undetectable communities

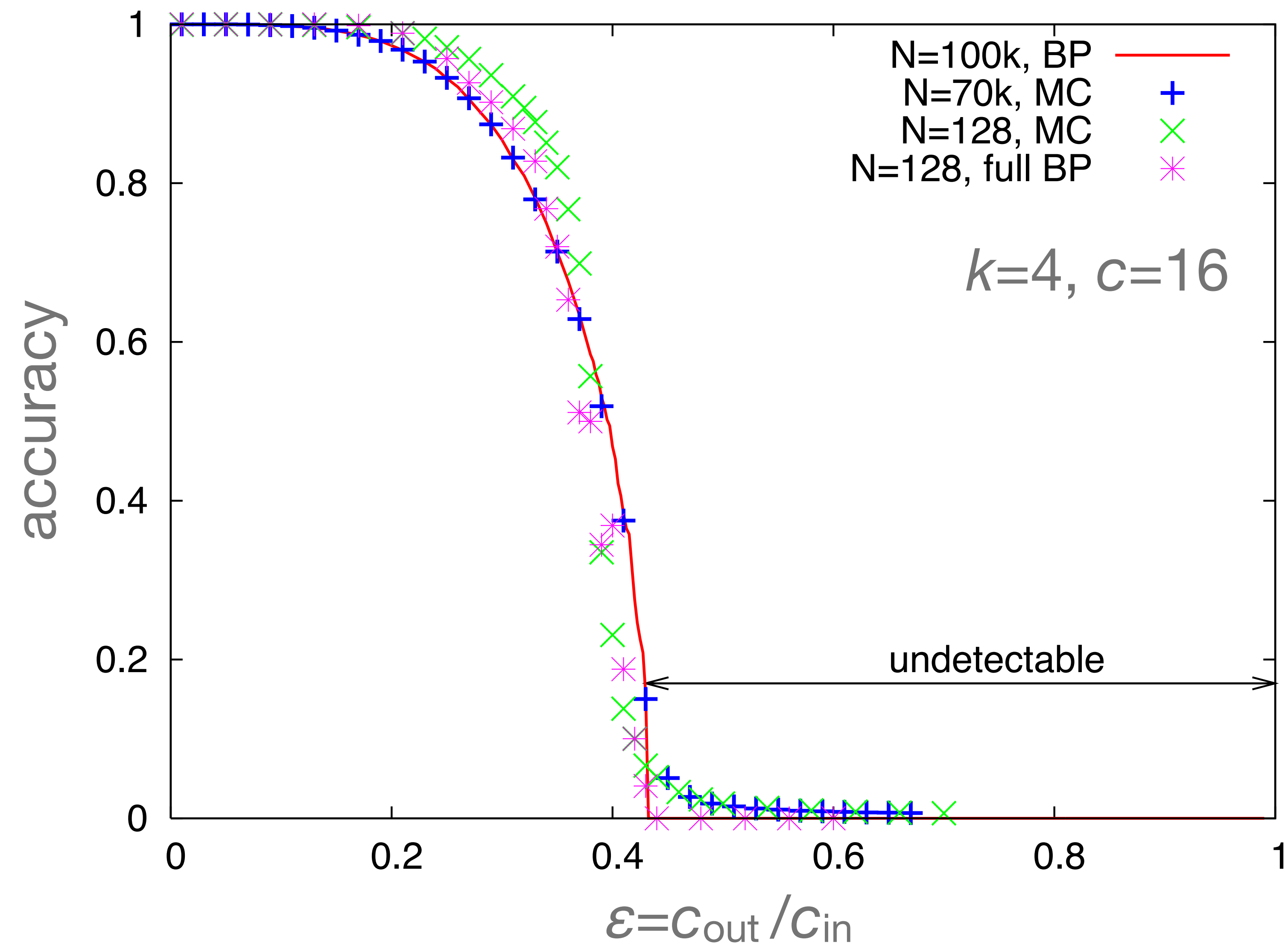
when c_{out}/c_{in} is small enough, BP can find the communities

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

for 2 groups, the threshold is at

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

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



A phase transition: detectable to undetectable communities

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

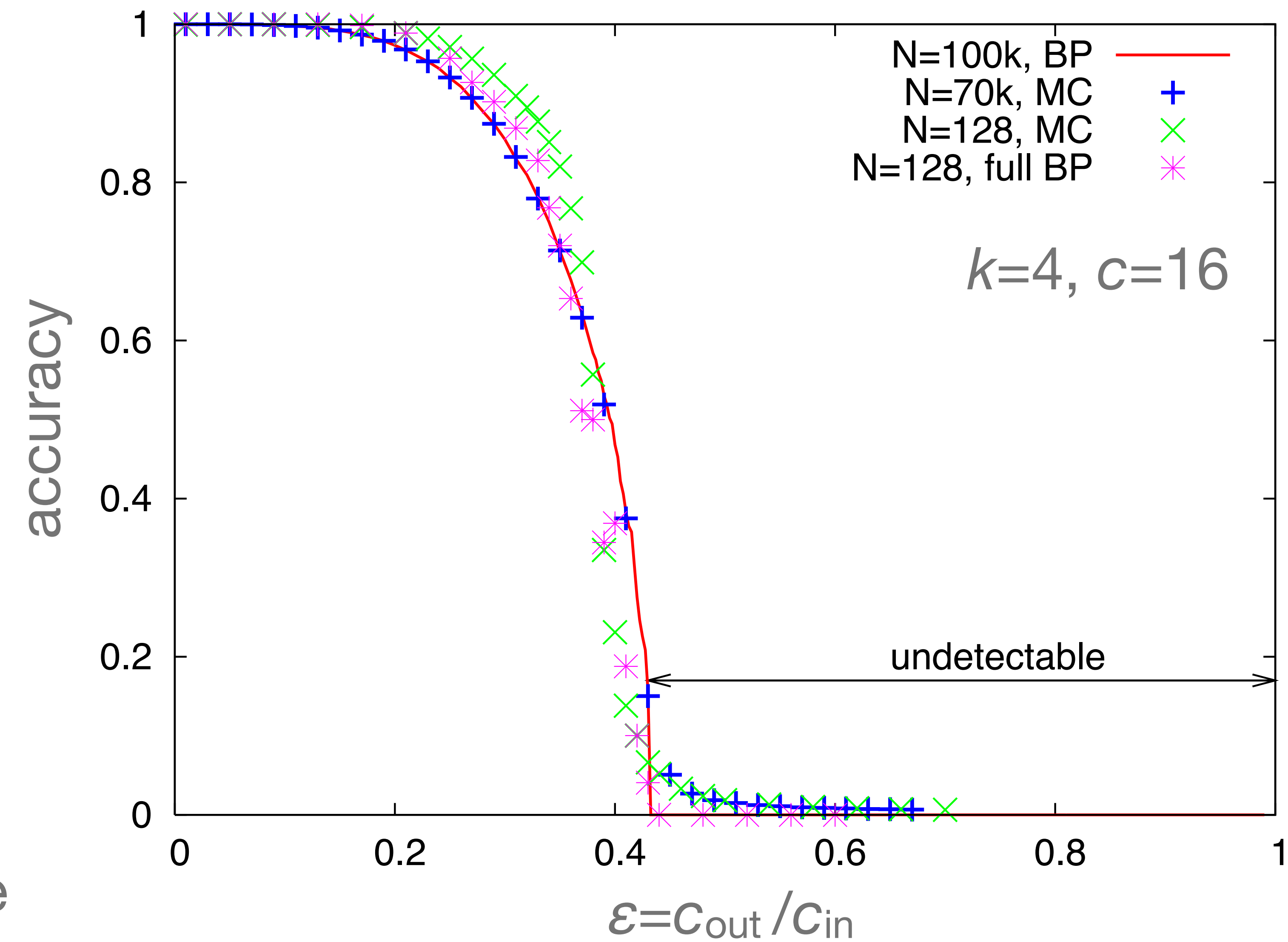
there is a regime where it can't, and no algorithm can!

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



A phase transition: detectable to undetectable communities

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

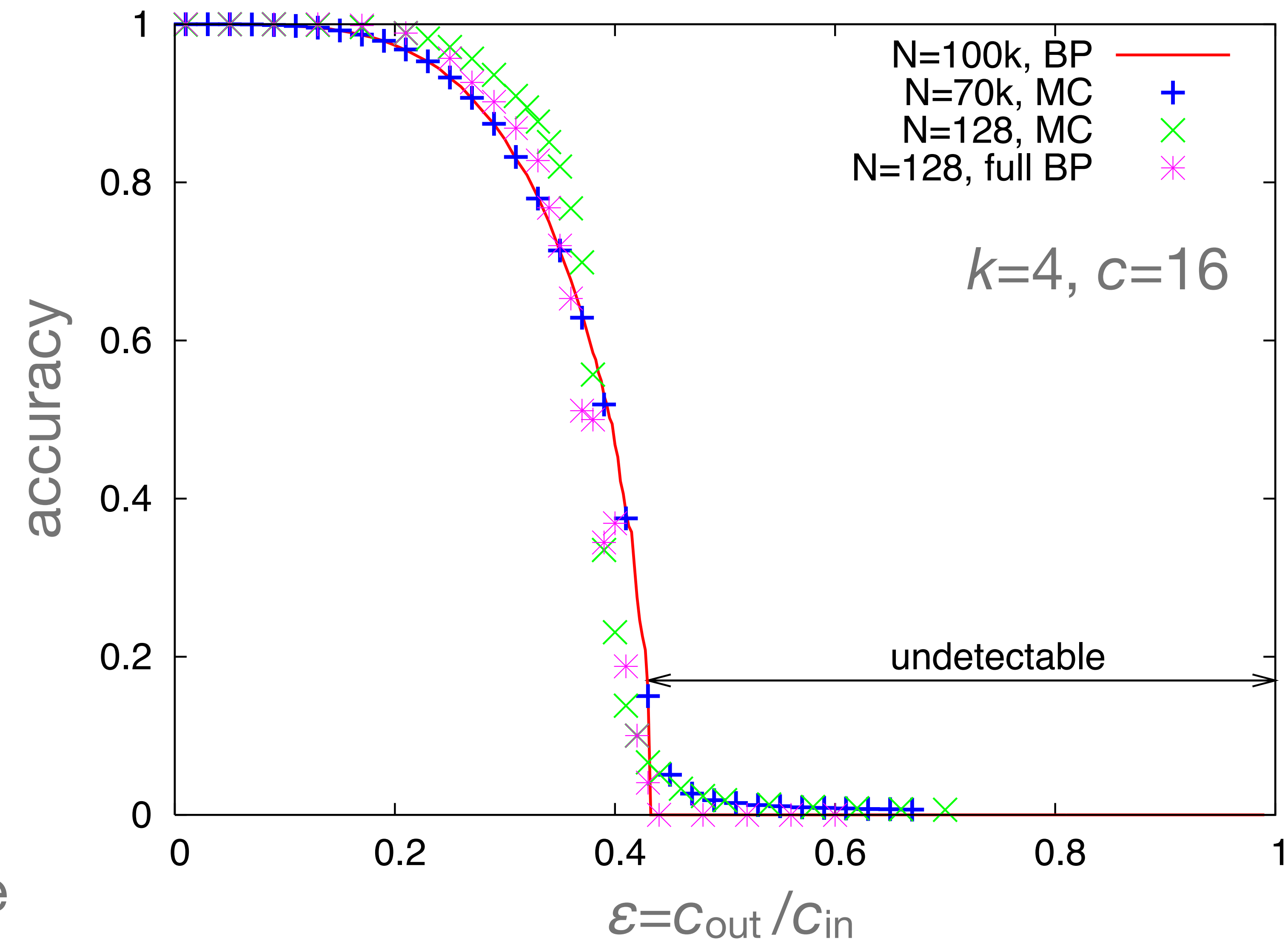
there is a regime where it can't, and no algorithm can!

for 2 groups, the threshold is at

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

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

at the transition, it becomes stable



conjectured by [Decelle, Krzakala, Moore, Zdeborová, '11]

A phase transition:
detectable to undetectable communities

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

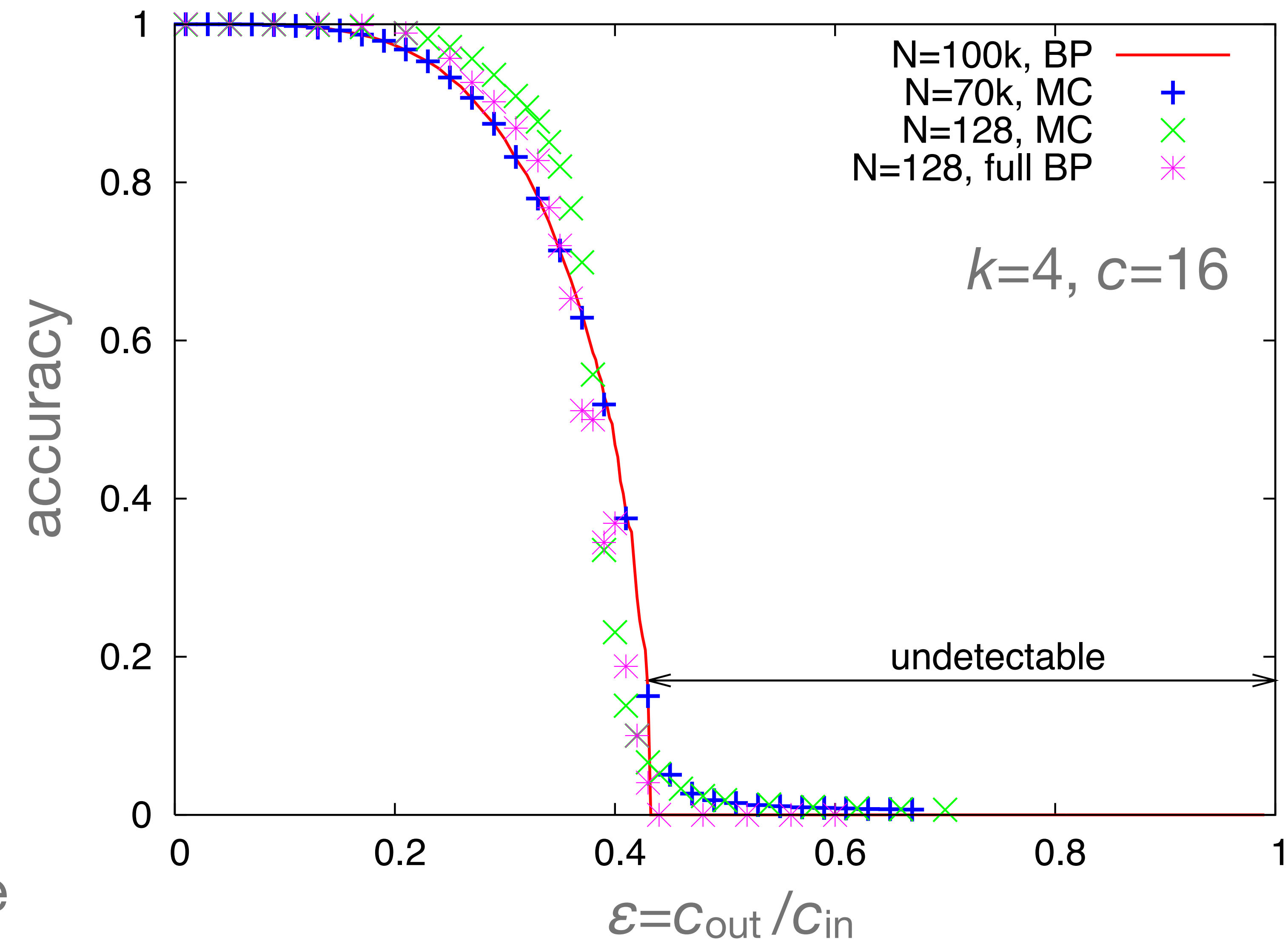
there is a regime where it can't, and no algorithm can!

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]

A phase transition:
detectable to undetectable communities

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

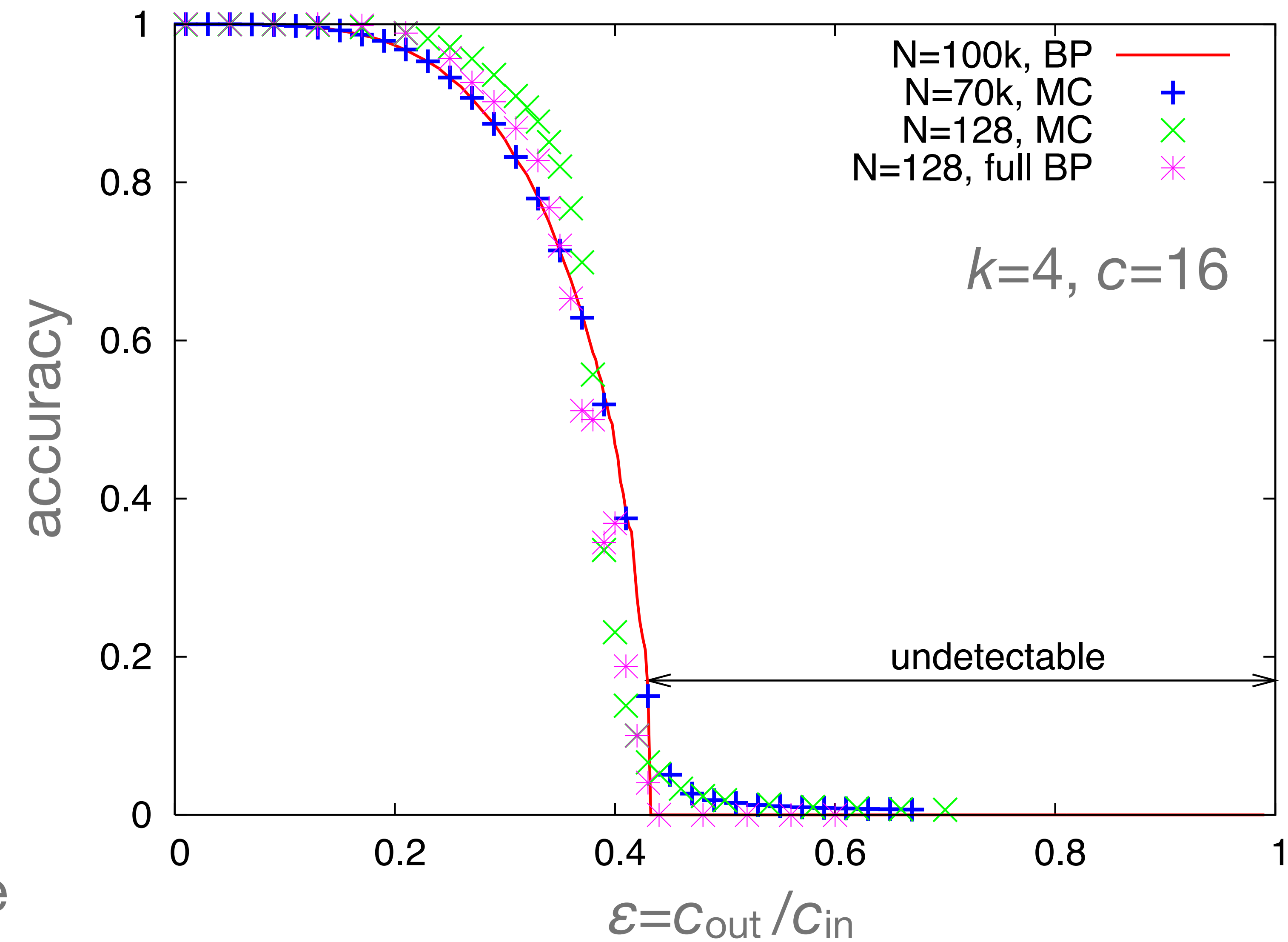
there is a regime where it can't, and no algorithm can!

for 2 groups, the threshold is at

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

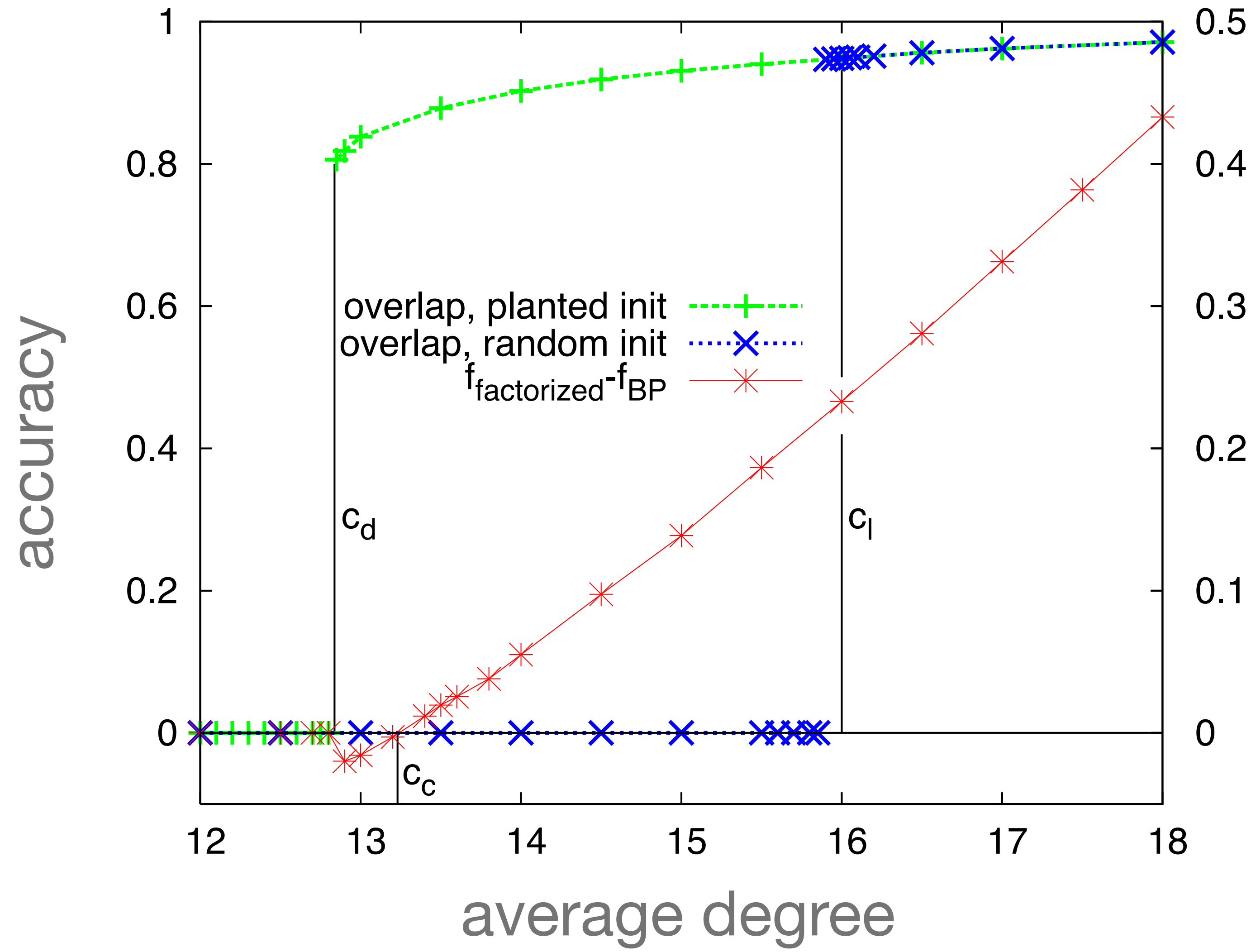
there is a fixed point where all nodes have uniform marginals...

at the transition, it becomes stable

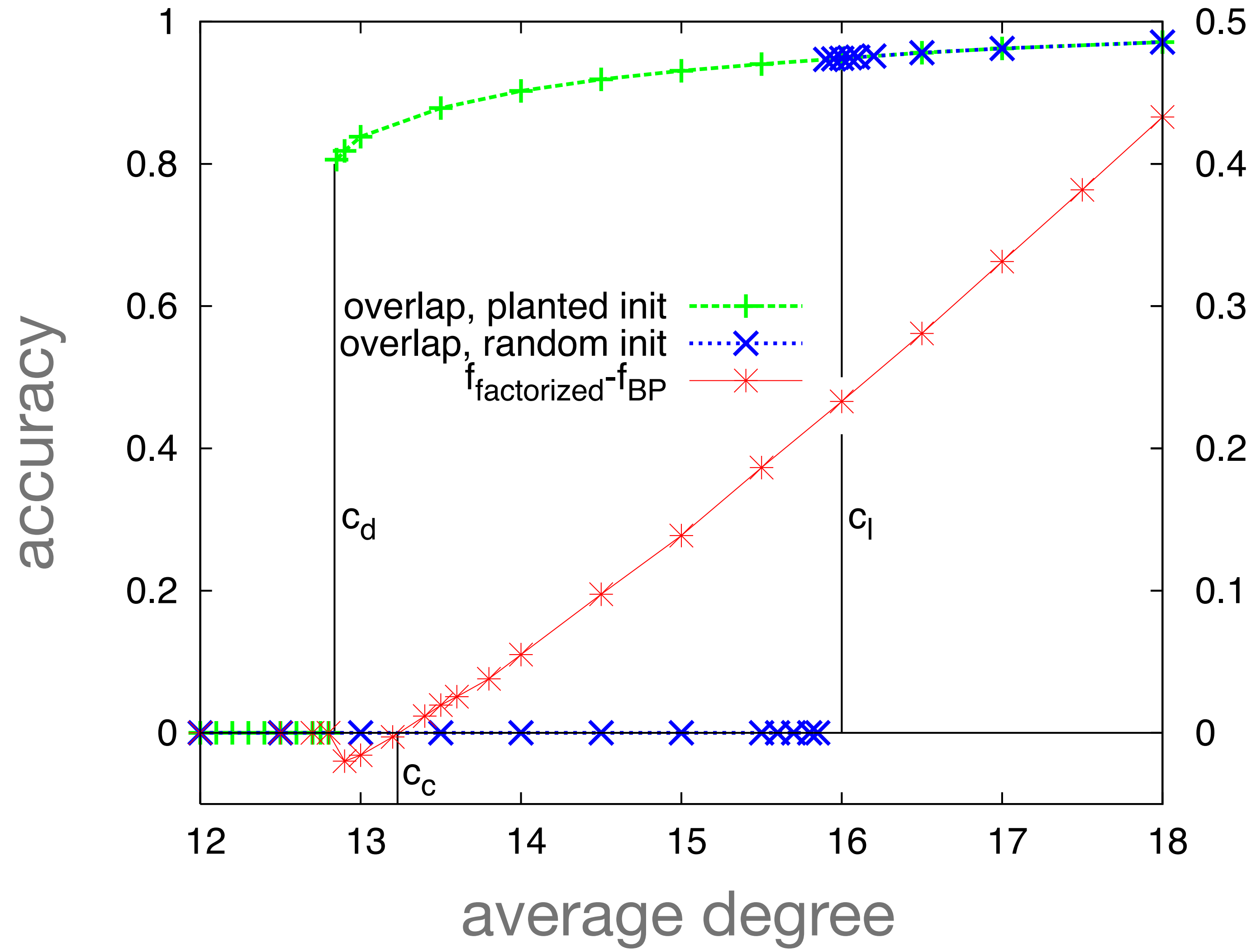


conjectured by [Decelle, Krzakala, Moore, Zdeborová, '11]
proved by [Mossel, Neeman, Sly, `13; Massoulié '13]
for $k > 2$ groups, much less is known rigorously...

Another regime: detectable but hard

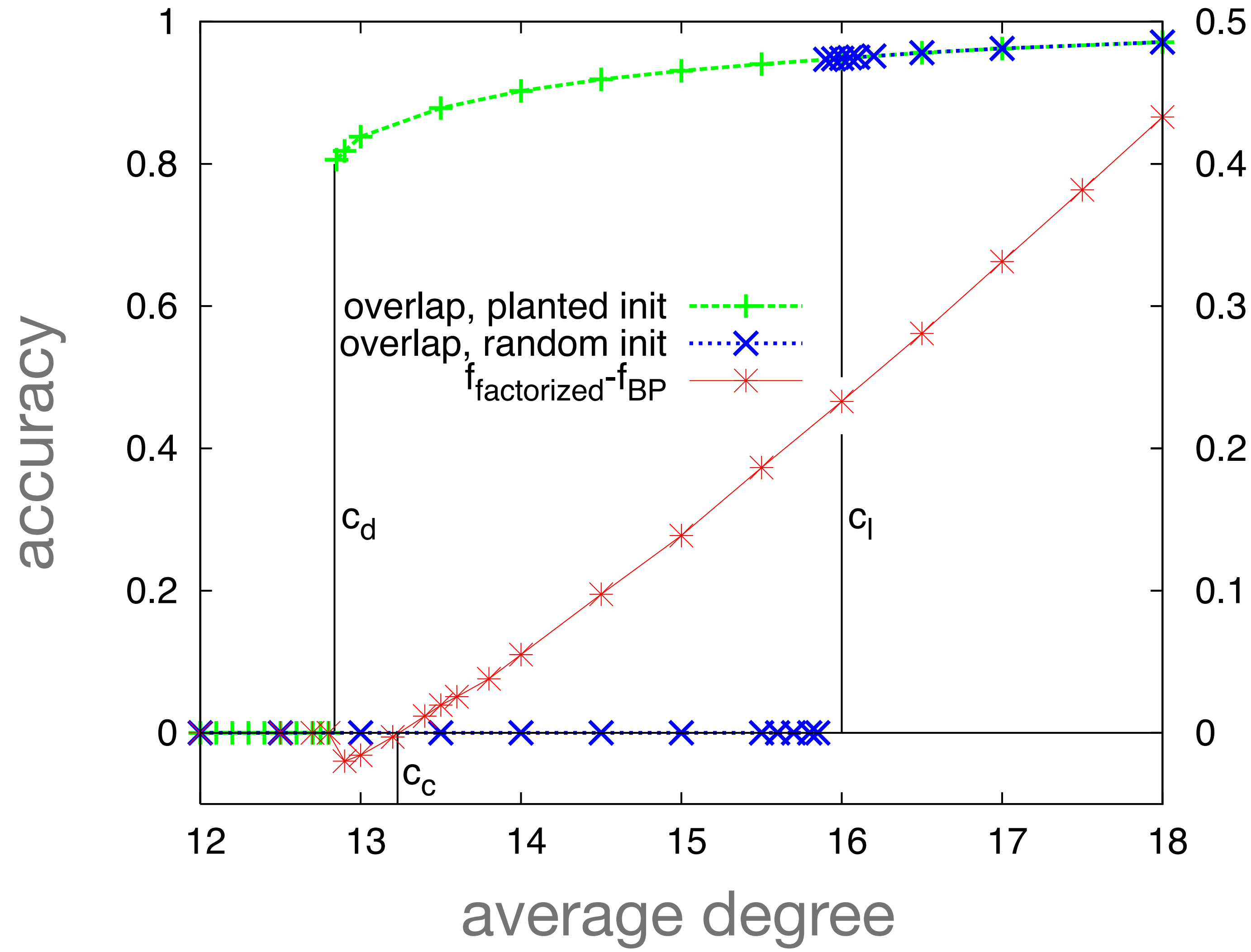


Another regime: detectable but hard



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

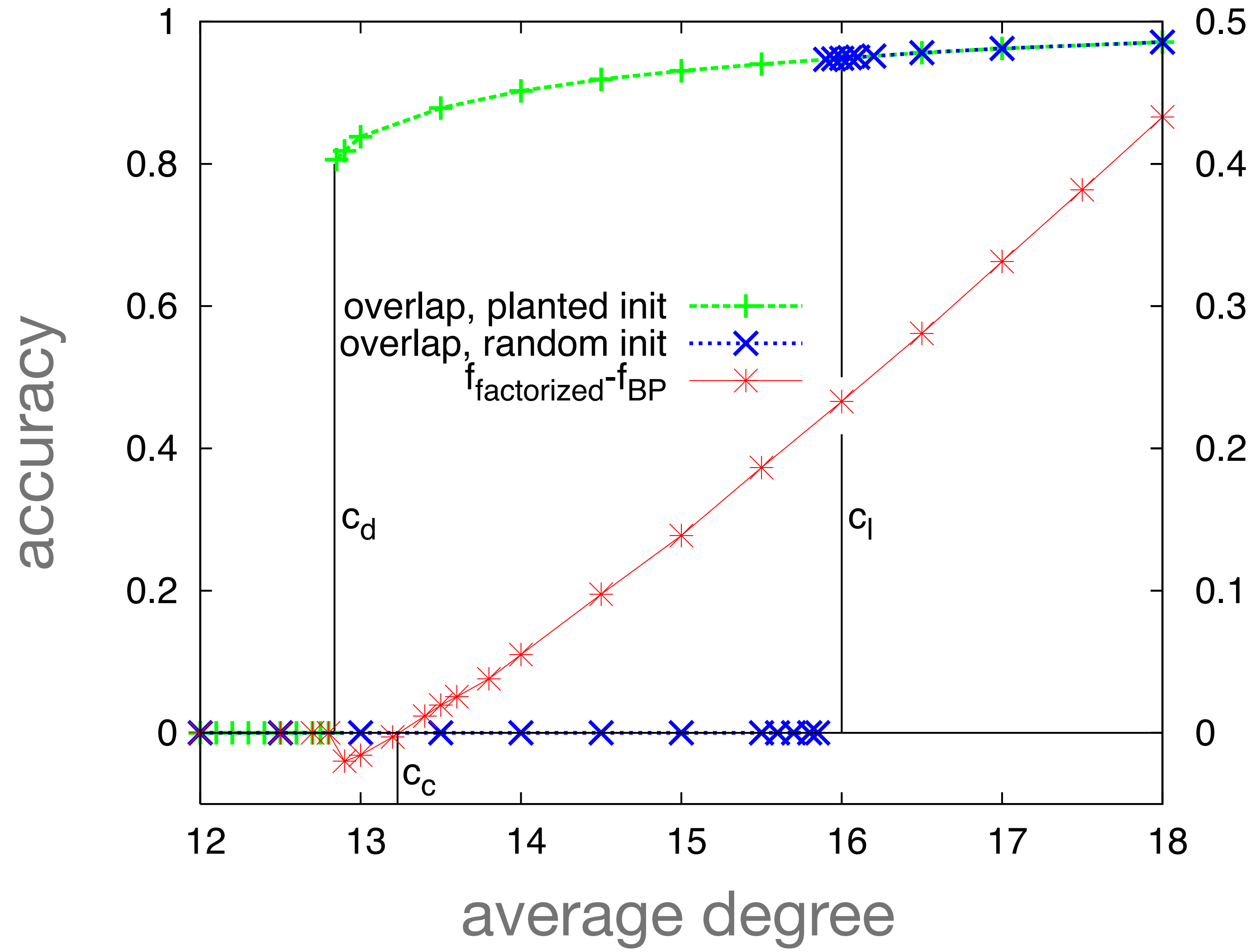
Another regime: detectable but hard



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

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

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 below the Kesten-Stigum bound [Abbe+Sandon, Banks+Moore] but we believe it’s computationally hard

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

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

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

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

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

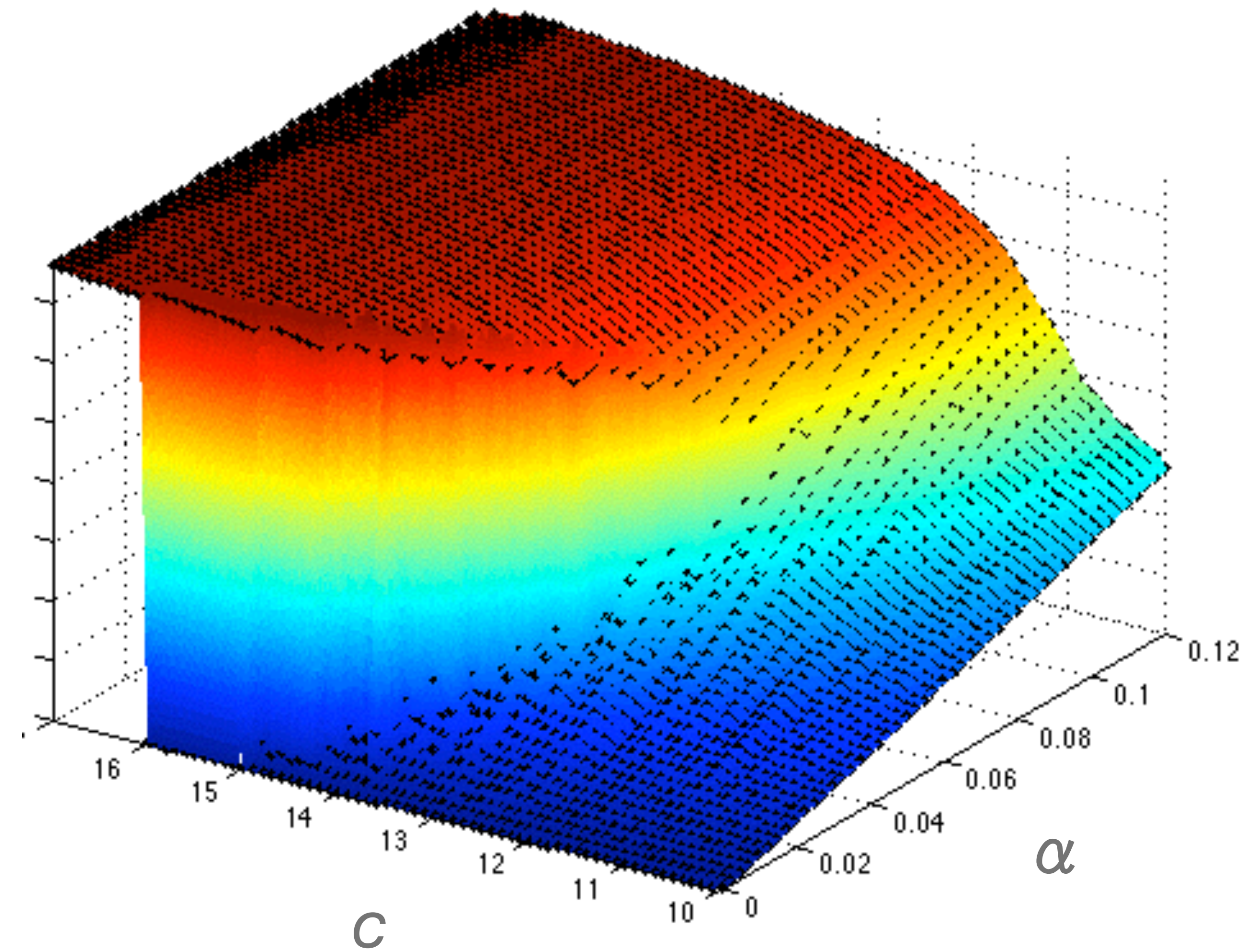
can we extend this information to the rest of the graph?

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



[Zhang, Moore, Zdeborová '14]

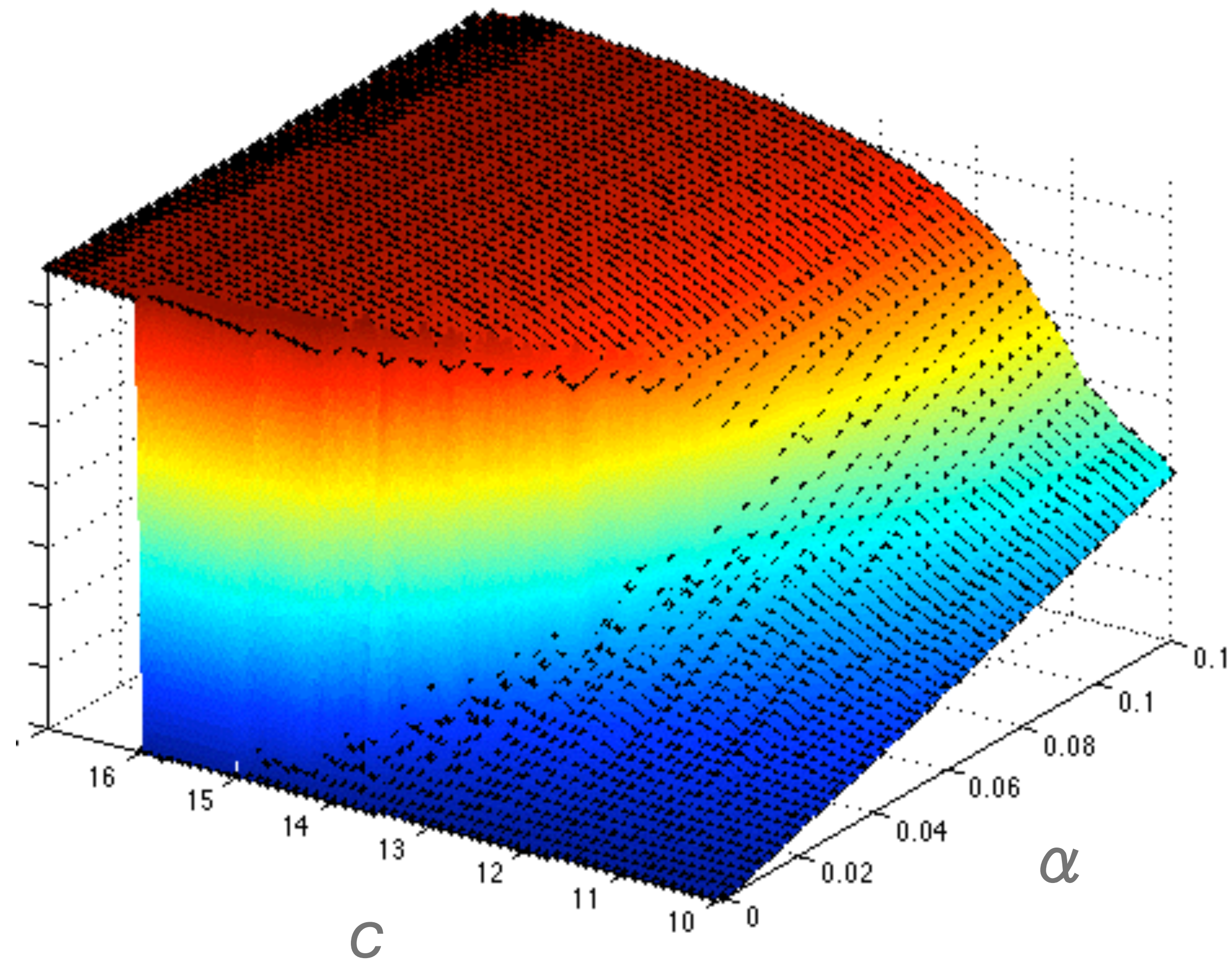
Phase transitions with metadata: what if we know some labels?

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

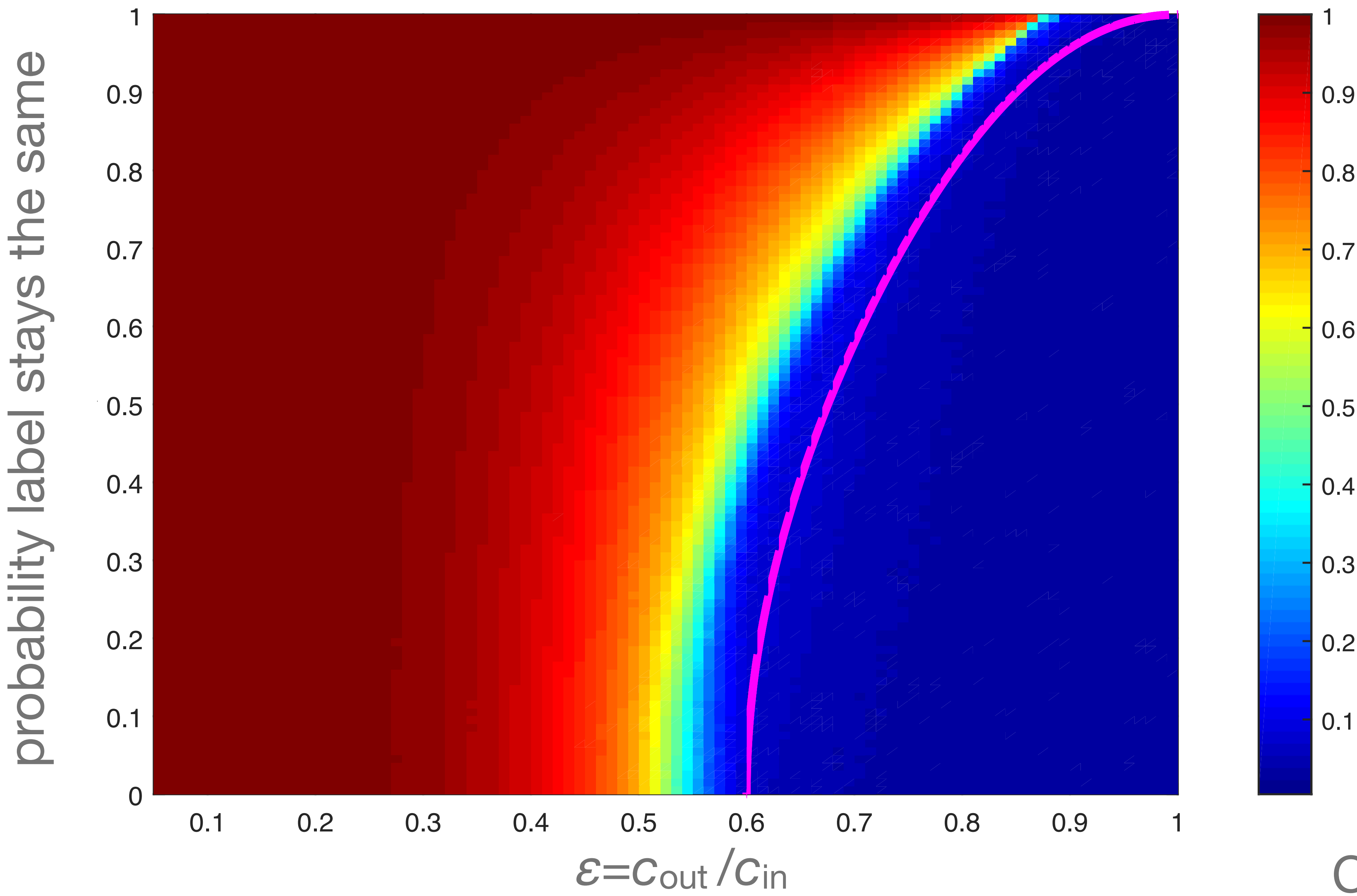
Dynamic networks

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

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?

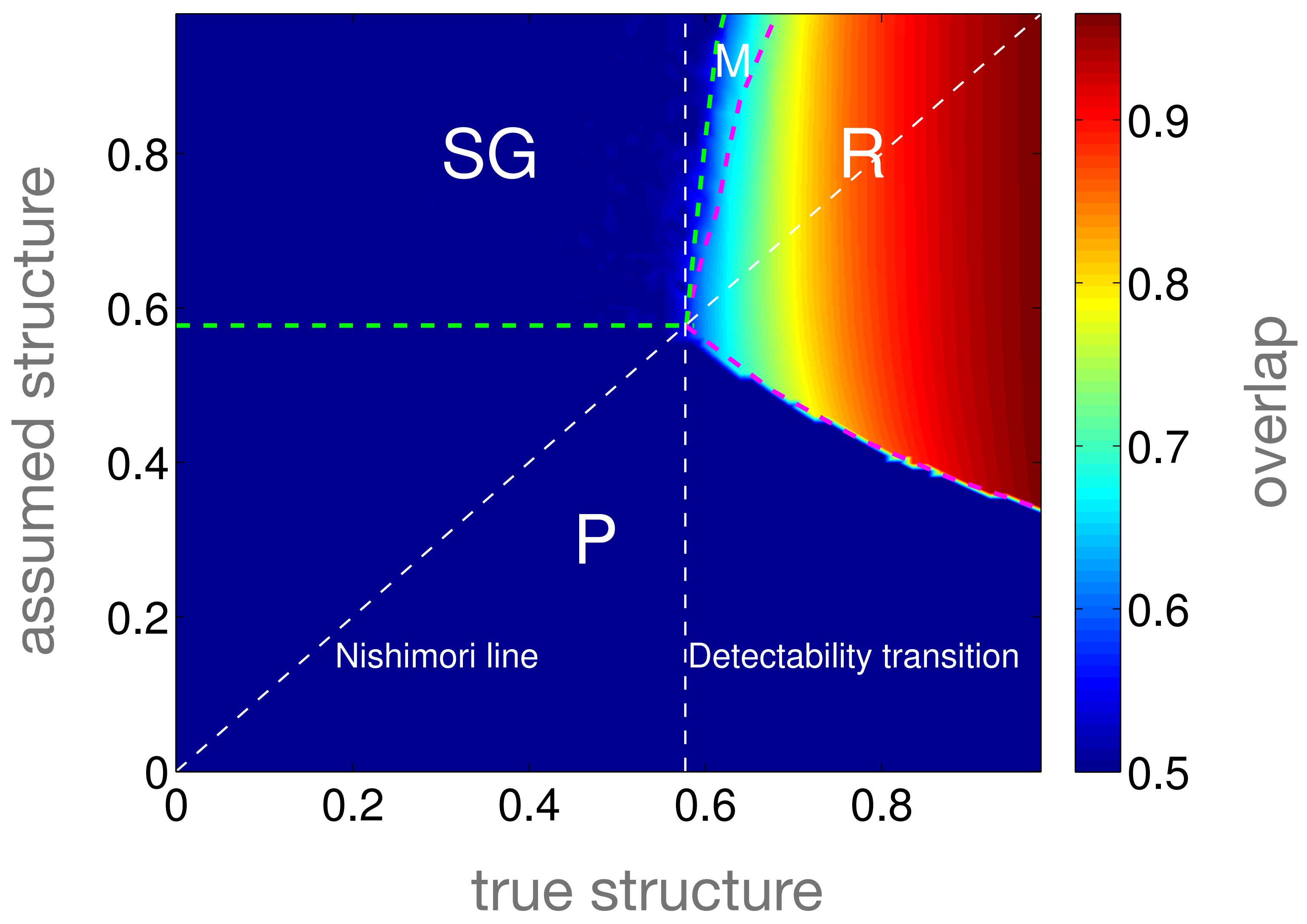
What if we don’t know how strong the structure is?

lower temperature = greedier algorithm = assume stronger structure

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]

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

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...

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

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):

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
multiplexer
buses
microinstructions
microprograms
```

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
multiplexer
buses
microinstructions
microprograms
```

```
testing
debugging
emulator
error
traces
embedding
jumps
halting
```

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
- multiplexer
- buses
- microinstructions
- microprograms

- testing
- debugging
- emulator
- error
- traces
- embedding
- jumps
- halting

- power
- reset
- frequencies
- pulses
- voltages
- sensing
- driving
- oscillators

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	protection
multiplexer	debugging	reset	transparent
buses	emulator	frequencies	security
microinstructions	error	pulses	multi-tasking
microprograms	traces	voltages	encryption
	embedding	sensing	restricting
	jumps	driving	
	halting	oscillators	

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

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):

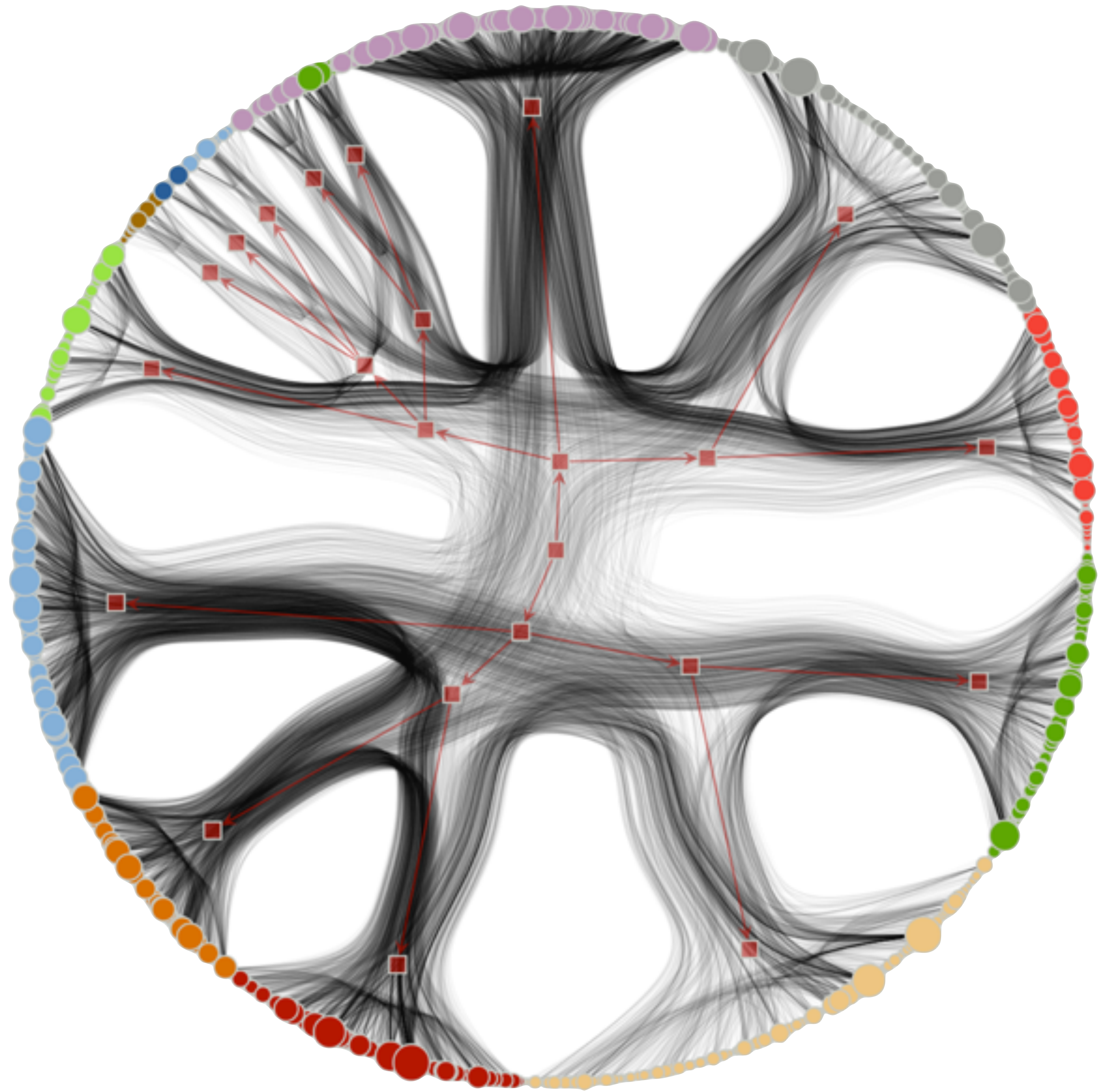
arithmetic	testing	power	protection	branching
multiplexer	debugging	reset	transparent	prediction
buses	emulator	frequencies	security	concurrence
microinstructions	error	pulses	multi-tasking	speculation
microprograms	traces	voltages	encryption	reordering
	embedding	sensing	restricting	
	jumps	driving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

using both text and links does better than either one alone

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

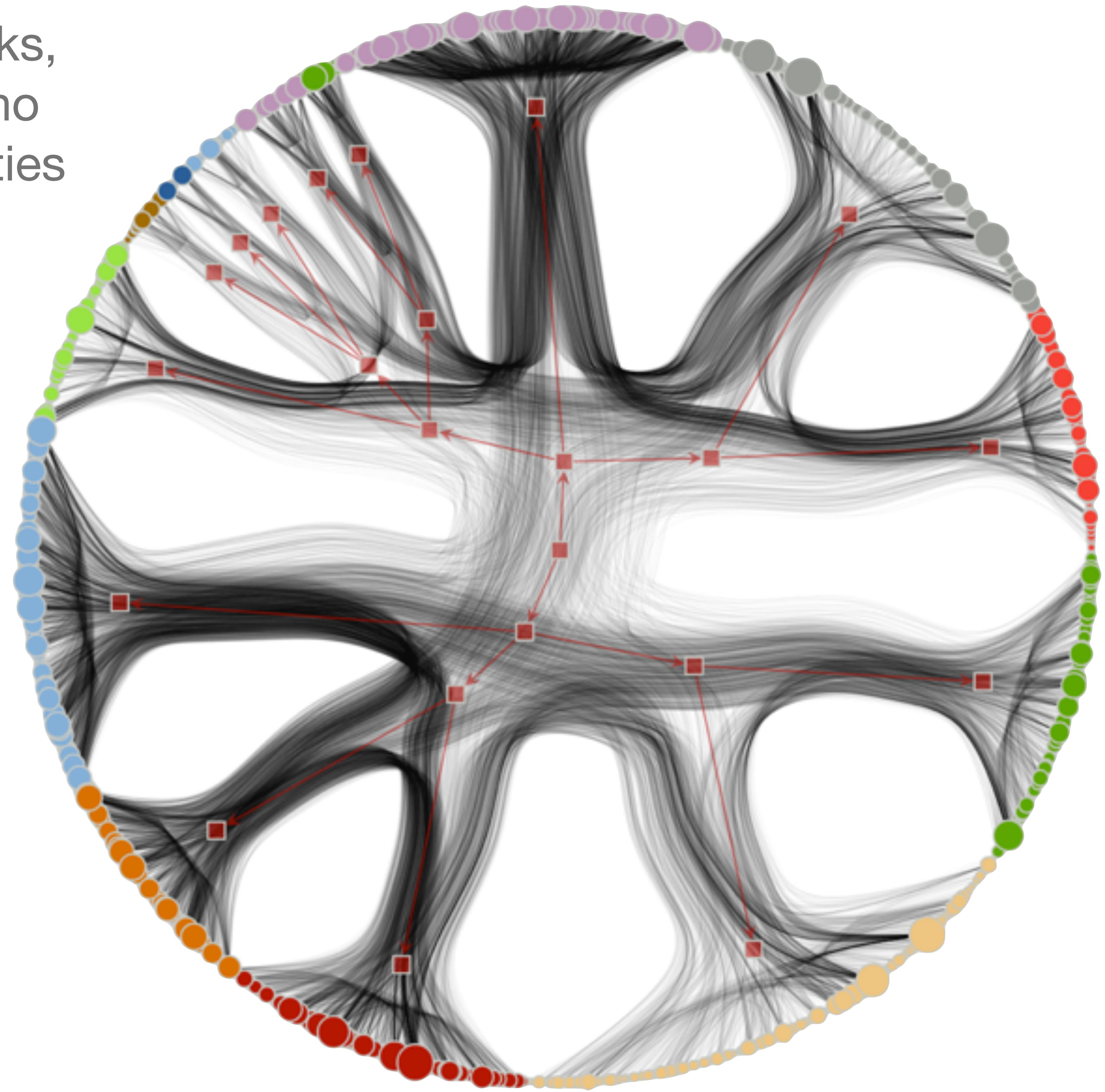
Hierarchical clustering

Hierarchical clustering



Hierarchical clustering

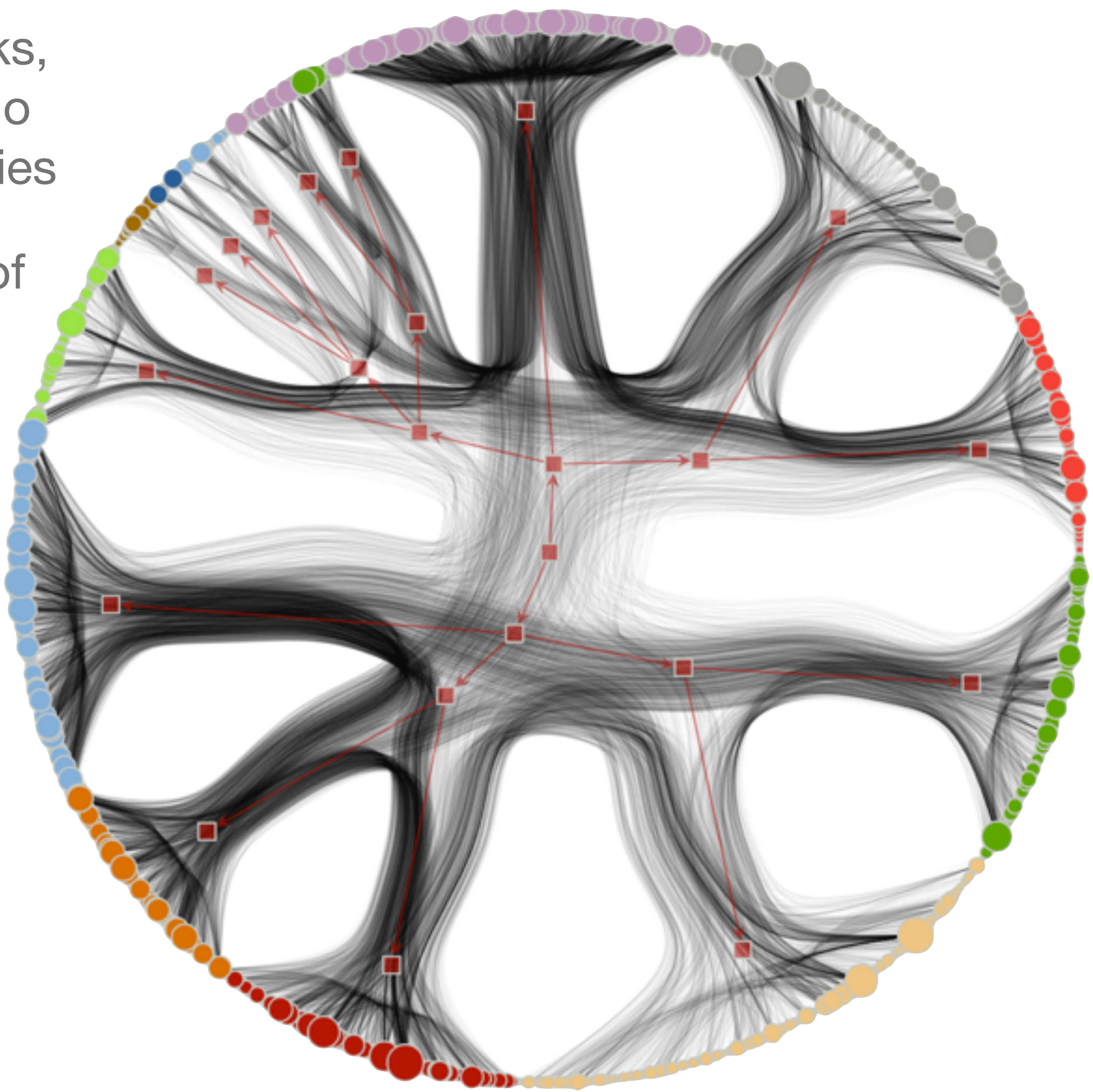
divide a network into subnetworks, until the remaining pieces have no statistically significant communities



Hierarchical clustering

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

reveals substructure in network of political blogs

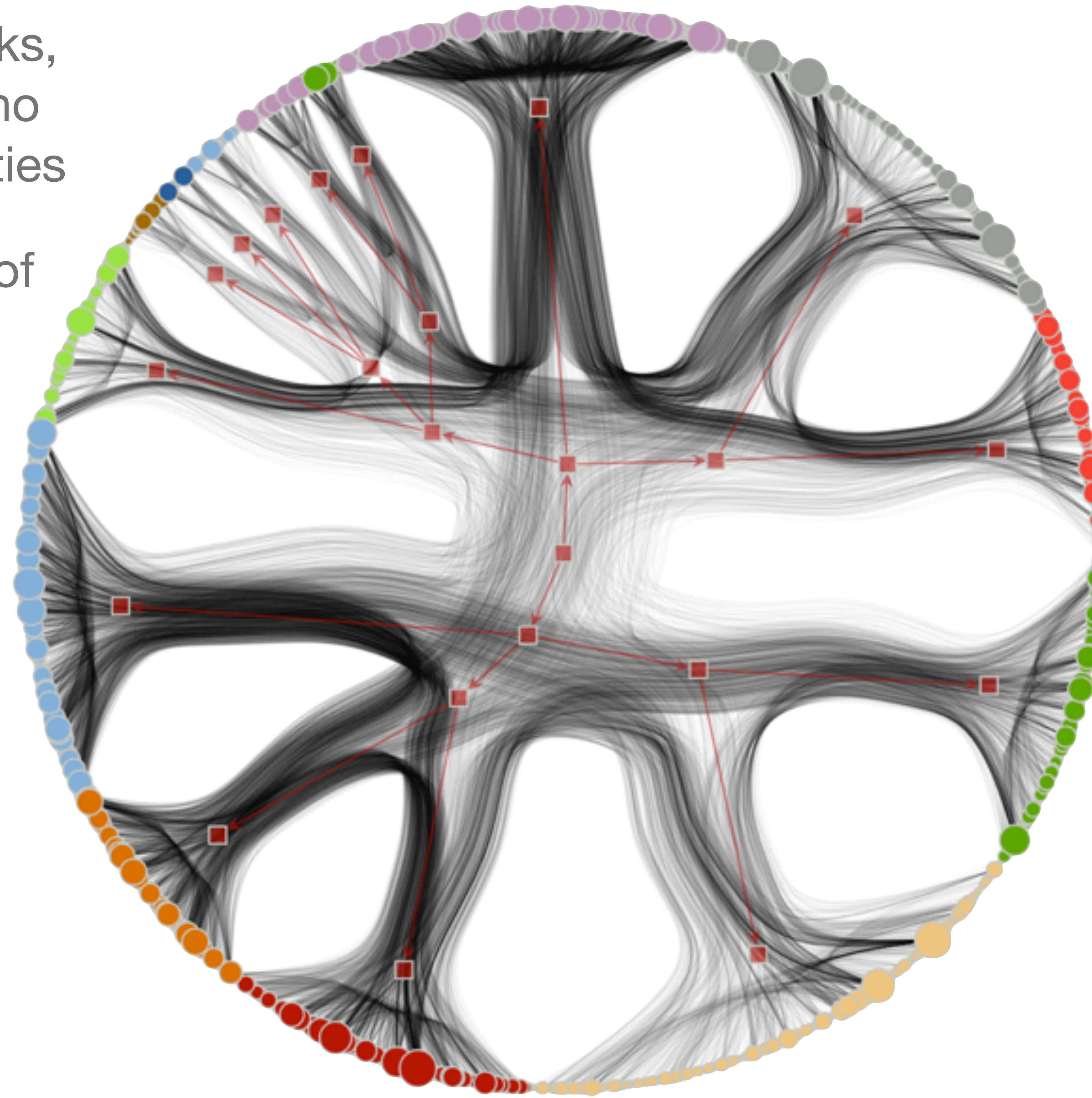


Hierarchical clustering

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

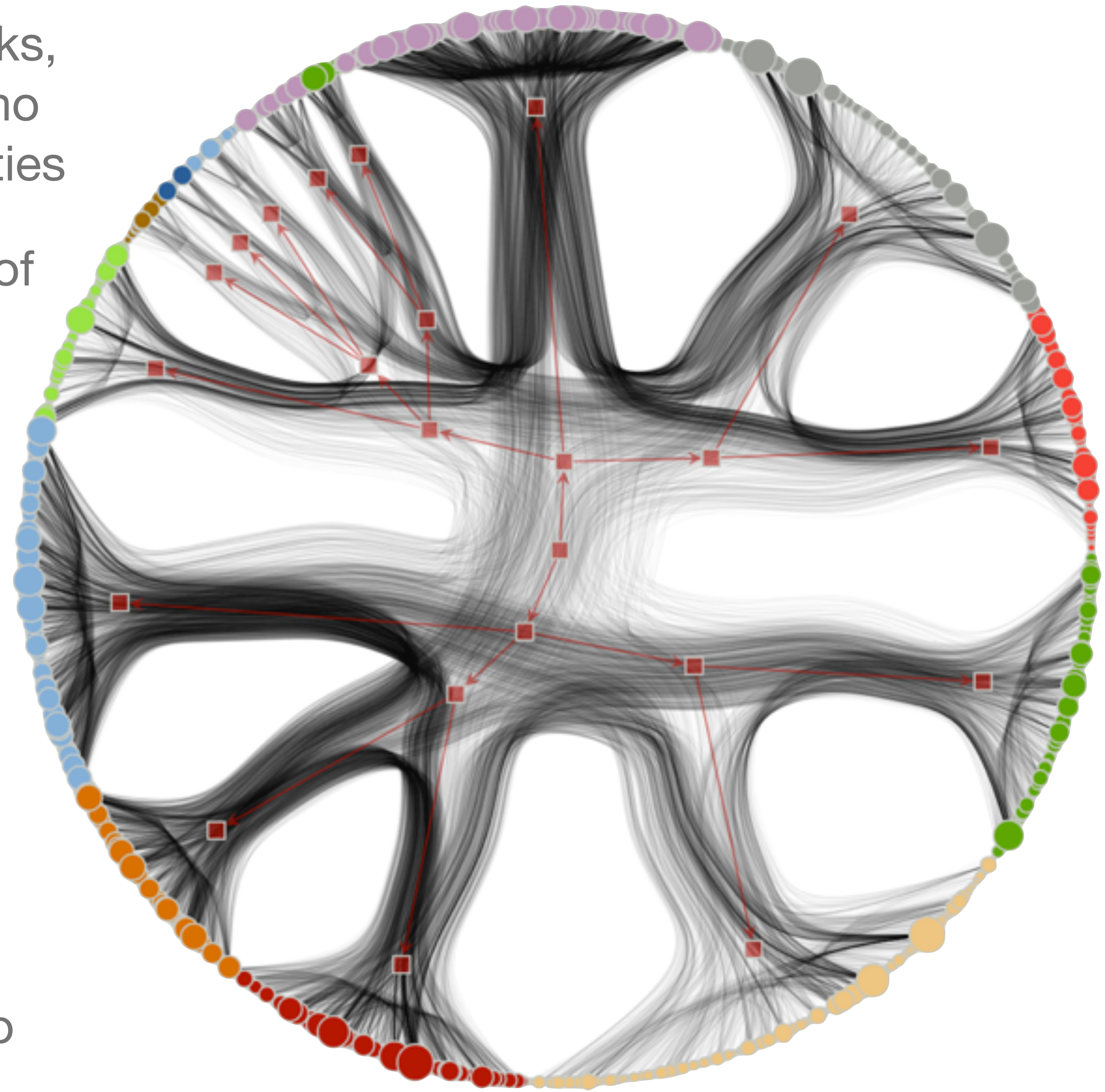


Hierarchical clustering

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

reveals substructure in network of political blogs

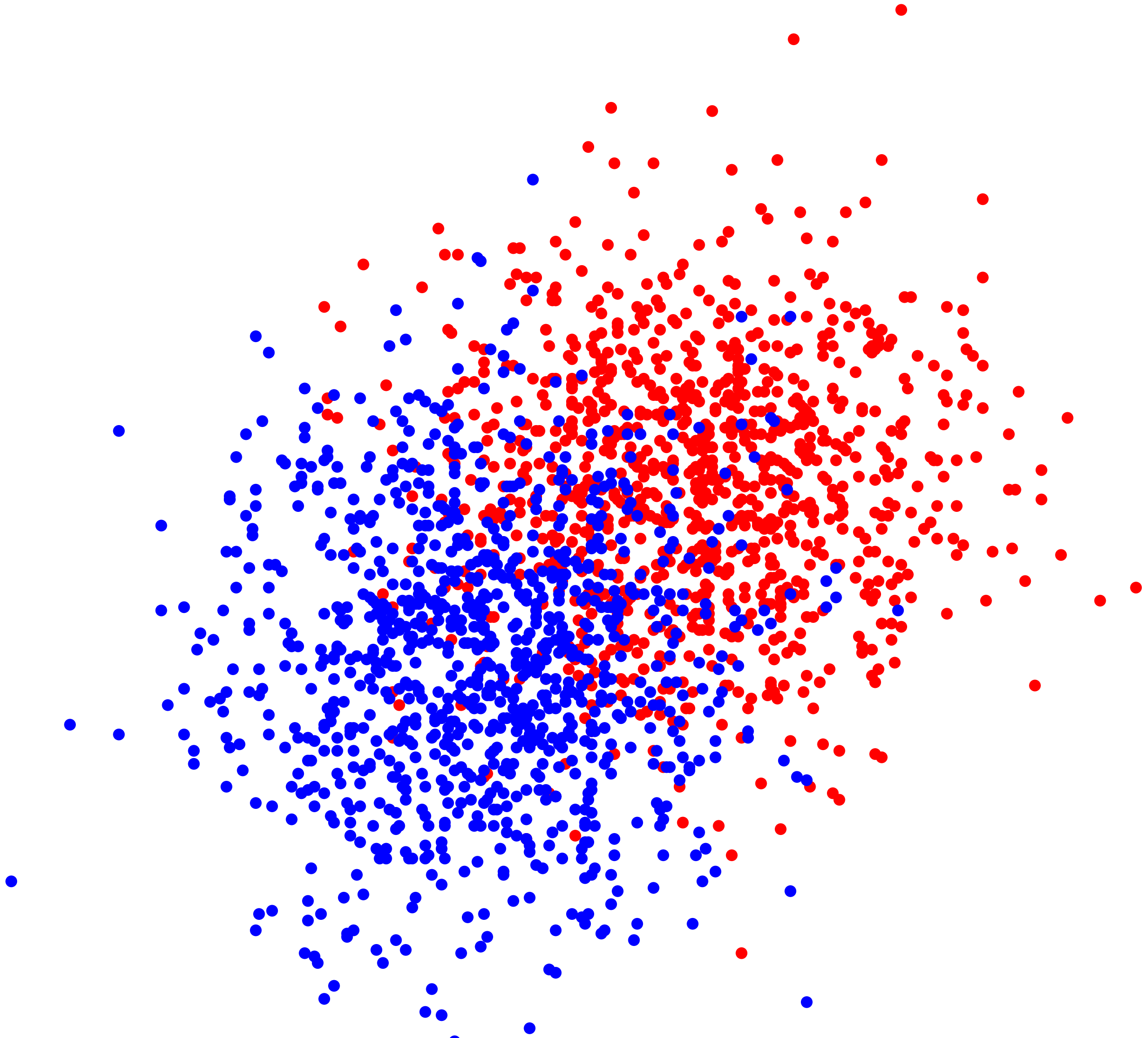
don’t maximize modularity!
the consensus of many high-modularity structures is better than the “best” one



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

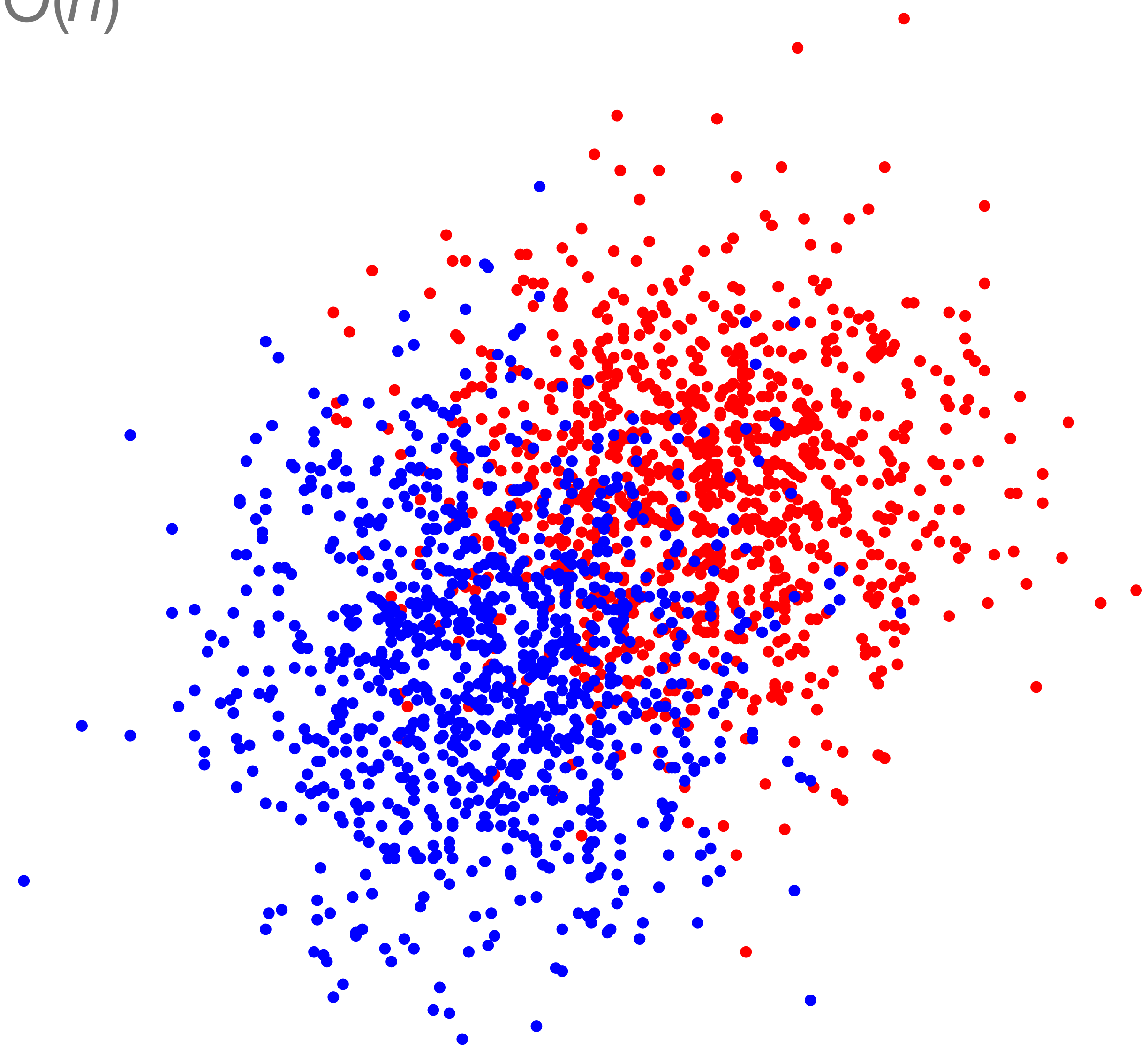
Clustering high-dimensional data

Sparse, high-dimensional clustering



Sparse, high-dimensional clustering

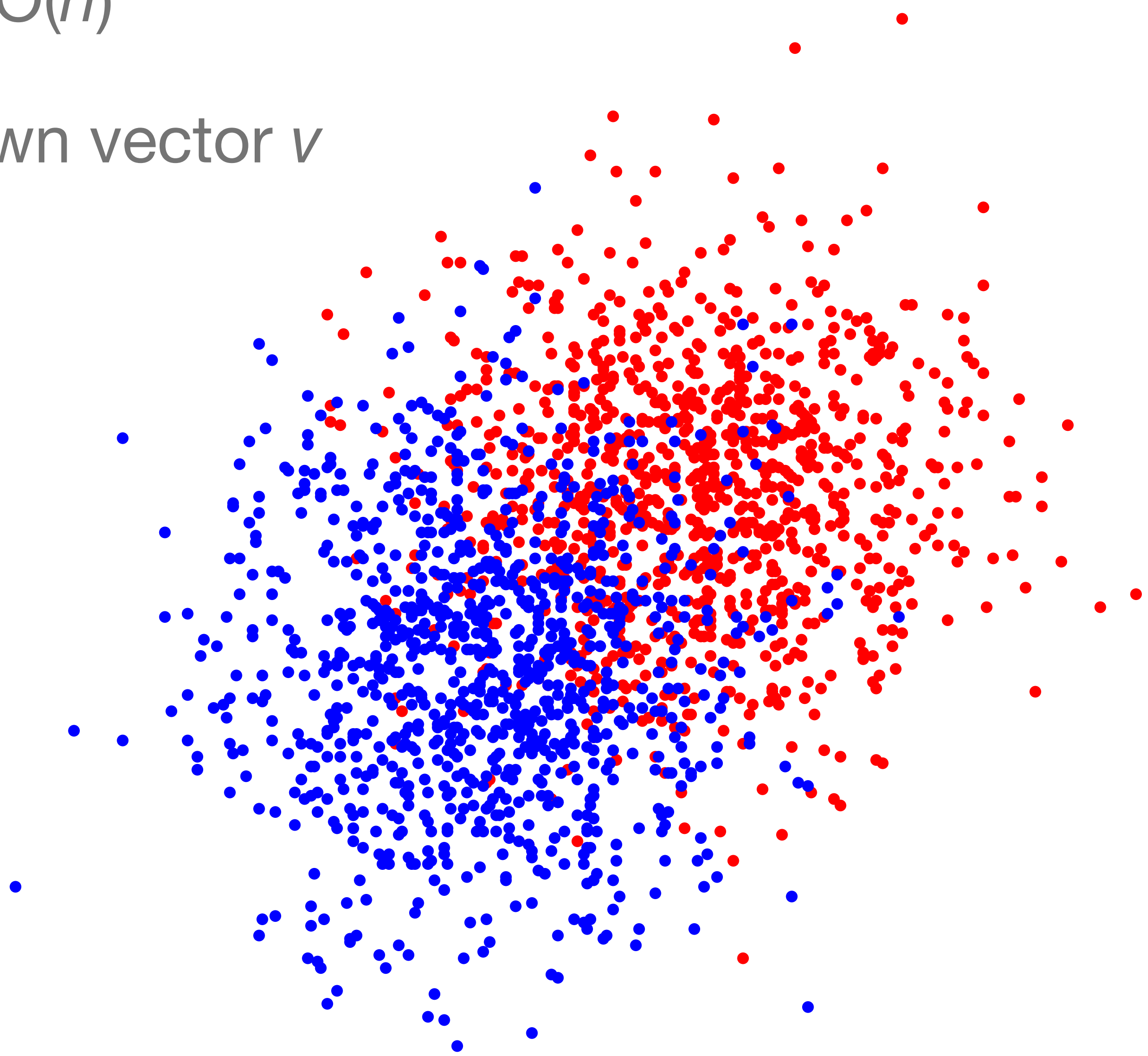
m points in *n*-dimensional space, where *m*=O(*n*)



Sparse, high-dimensional clustering

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

two clusters centered at ±v for some unknown vector v

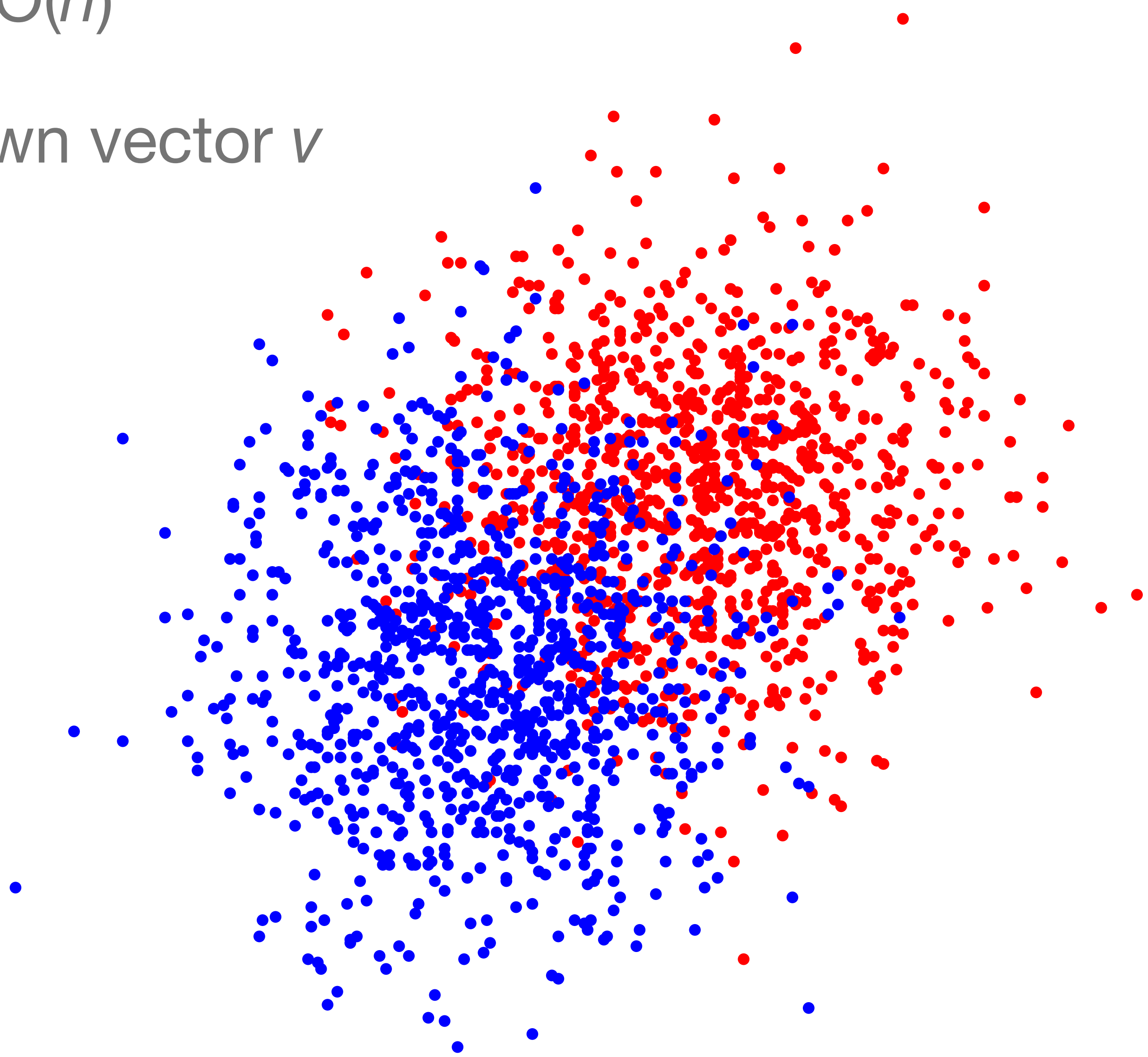


Sparse, high-dimensional clustering

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

two clusters centered at $\pm v$ for some unknown vector v

each point has a Gaussian noise vector u_i



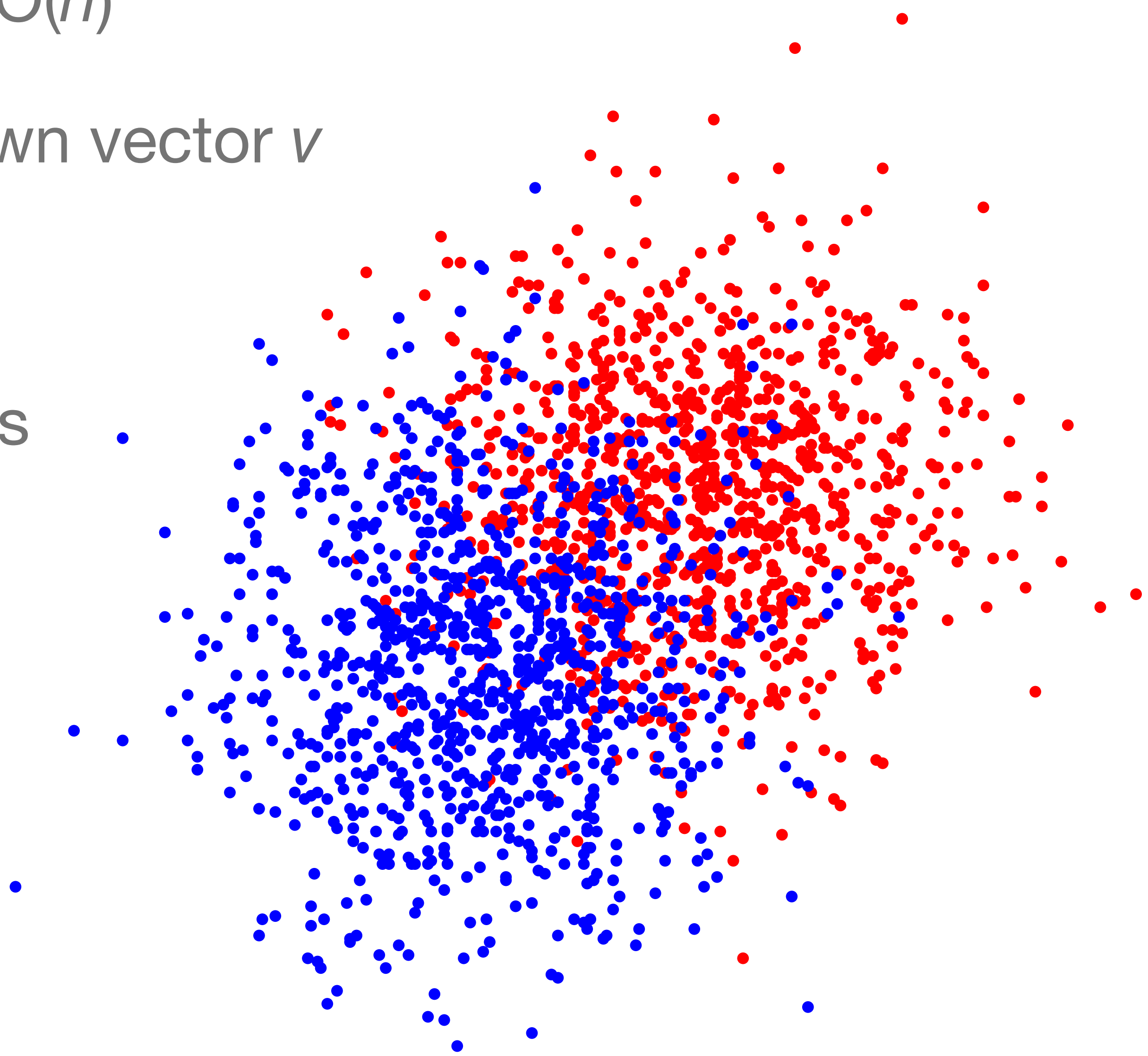
Sparse, high-dimensional clustering

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

two clusters centered at $\pm v$ for some unknown vector v

each point has a Gaussian noise vector u_i

$|v|=O(1)$ but u_i has variance 1 along each axis



Sparse, high-dimensional clustering

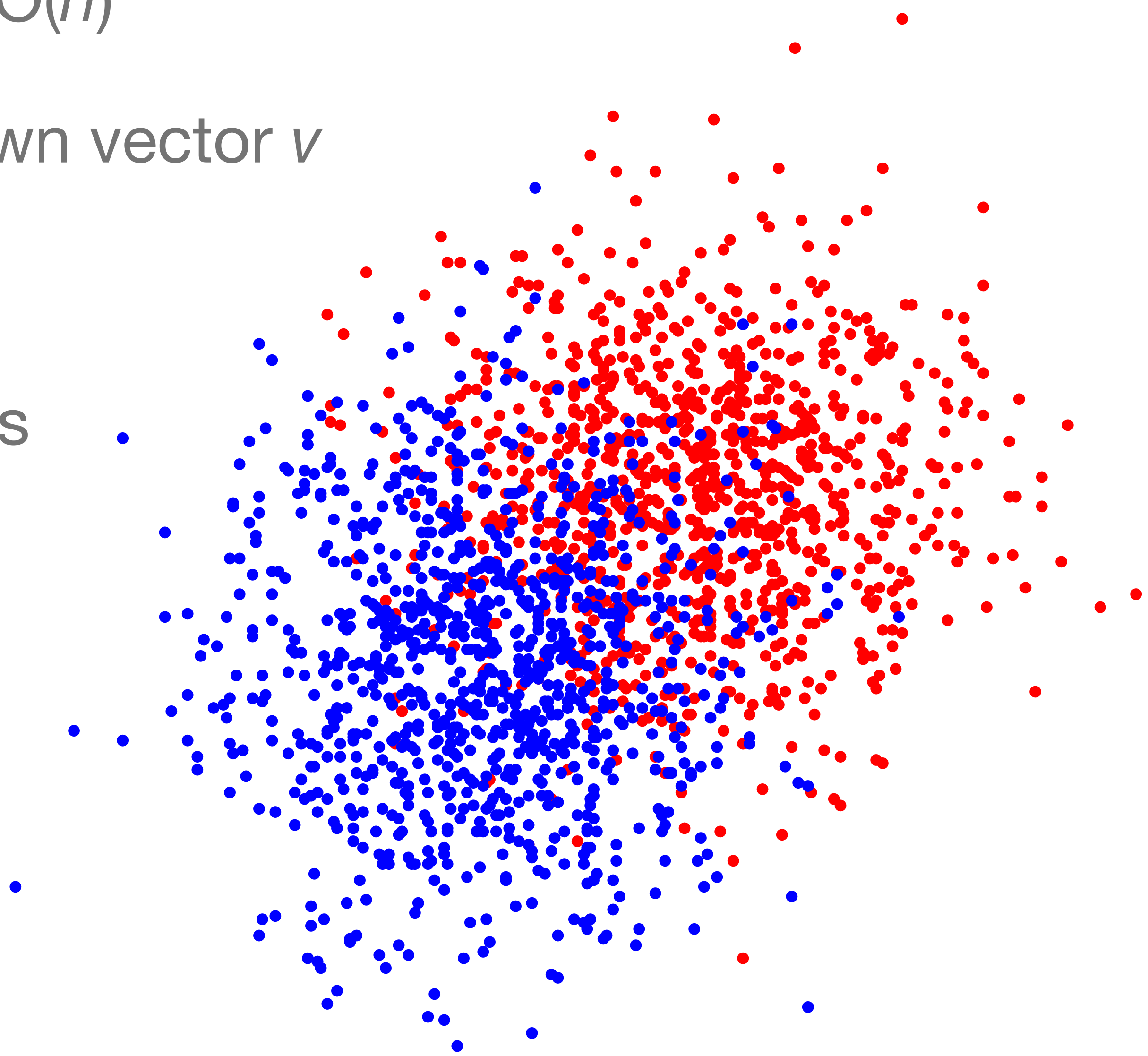
m points in n-dimensional space, where m=O(n)

two clusters centered at $\pm v$ for some unknown vector v

each point has a Gaussian noise vector u_i

$|v|=O(1)$ but u_i has variance 1 along each axis

when can we...



Sparse, high-dimensional clustering

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

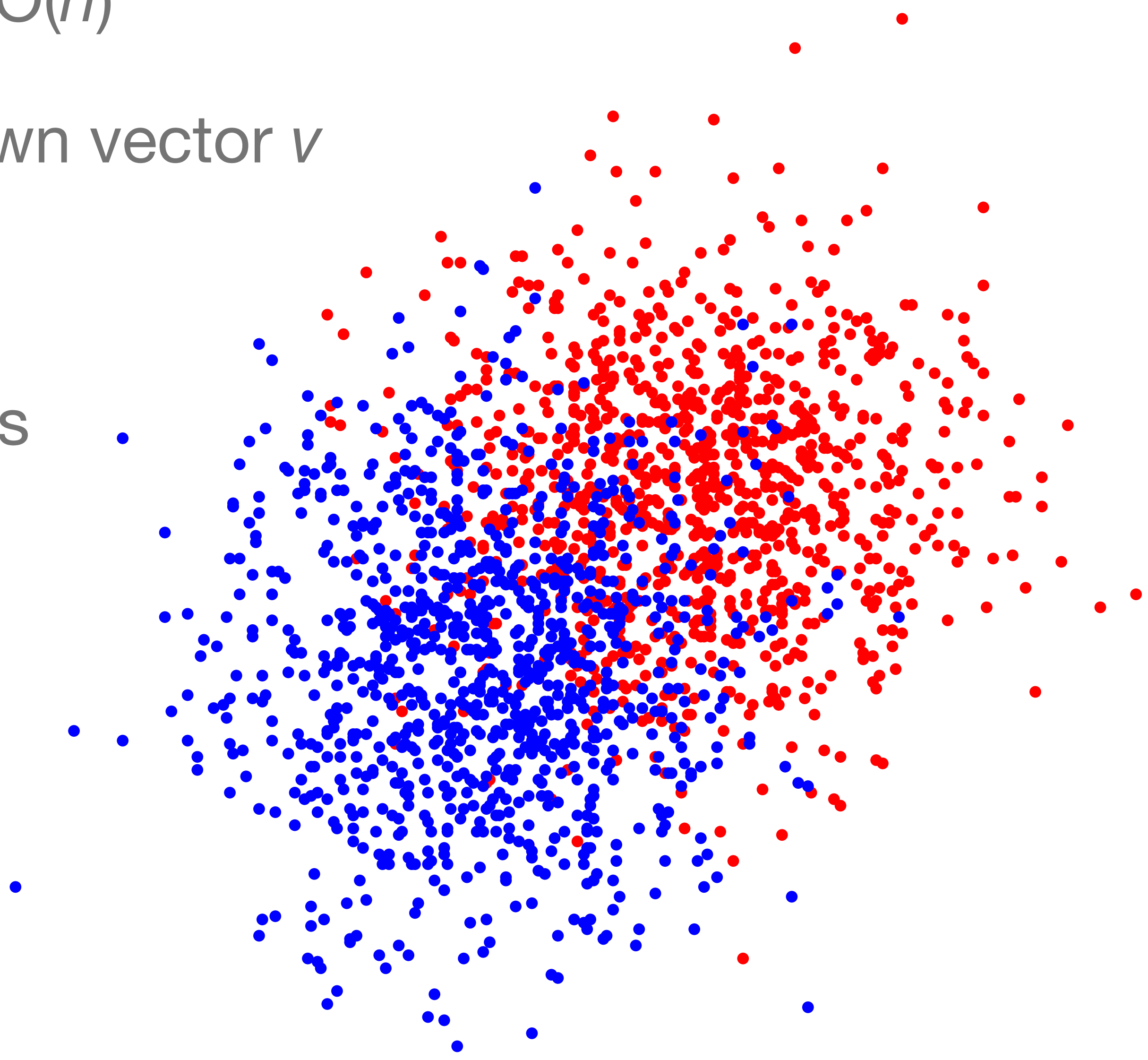
two clusters centered at $\pm v$ for some unknown vector v

each point has a Gaussian noise vector u_i

$|v|=O(1)$ but u_i has variance 1 along each axis

when can we...

find v?



Sparse, high-dimensional clustering

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

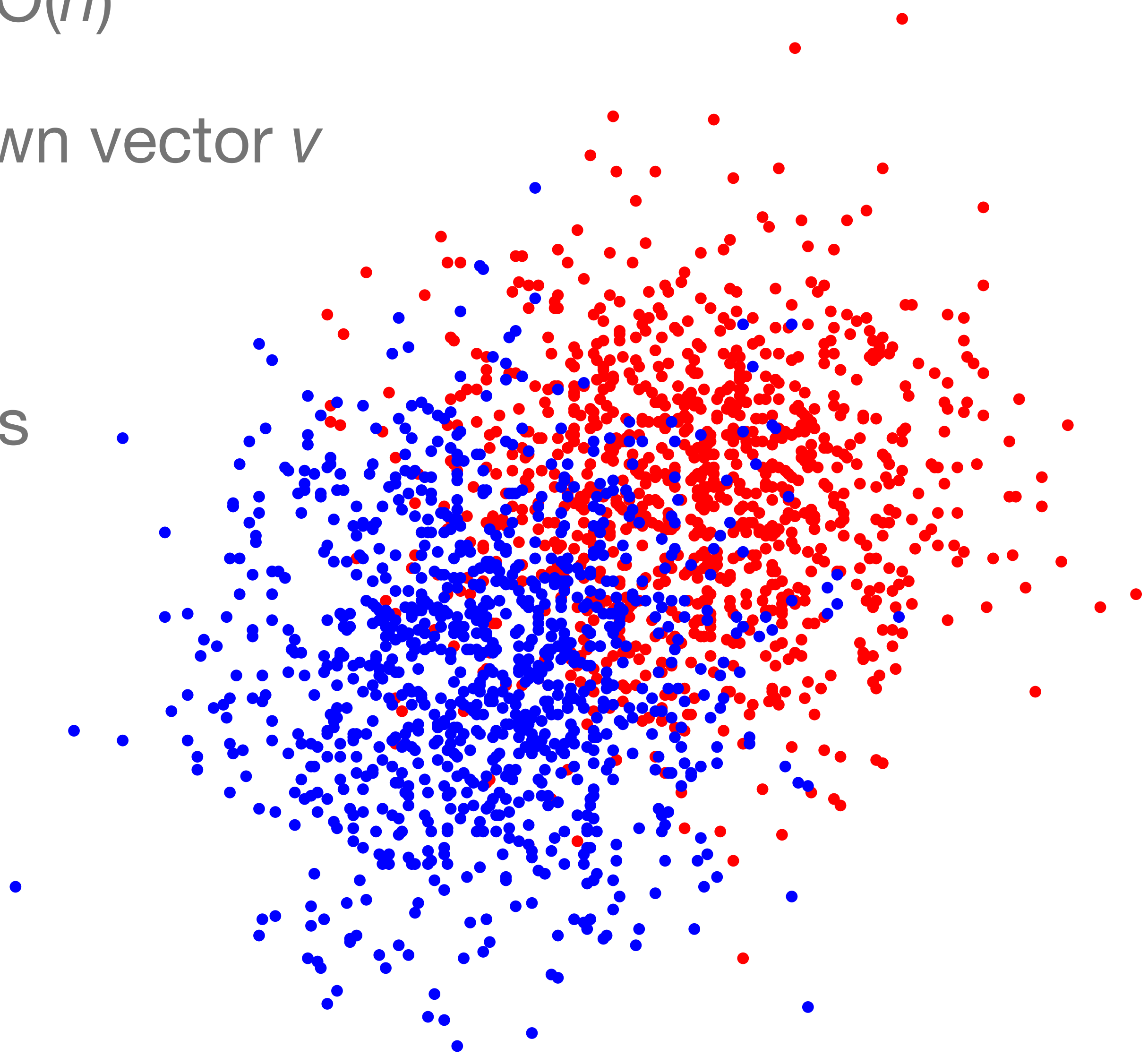
two clusters centered at $\pm v$ for some unknown vector v

each point has a Gaussian noise vector u_i

$|v|=O(1)$ but u_i has variance 1 along each axis

when can we...

- find v?
- label the points?



Sparse, high-dimensional clustering

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

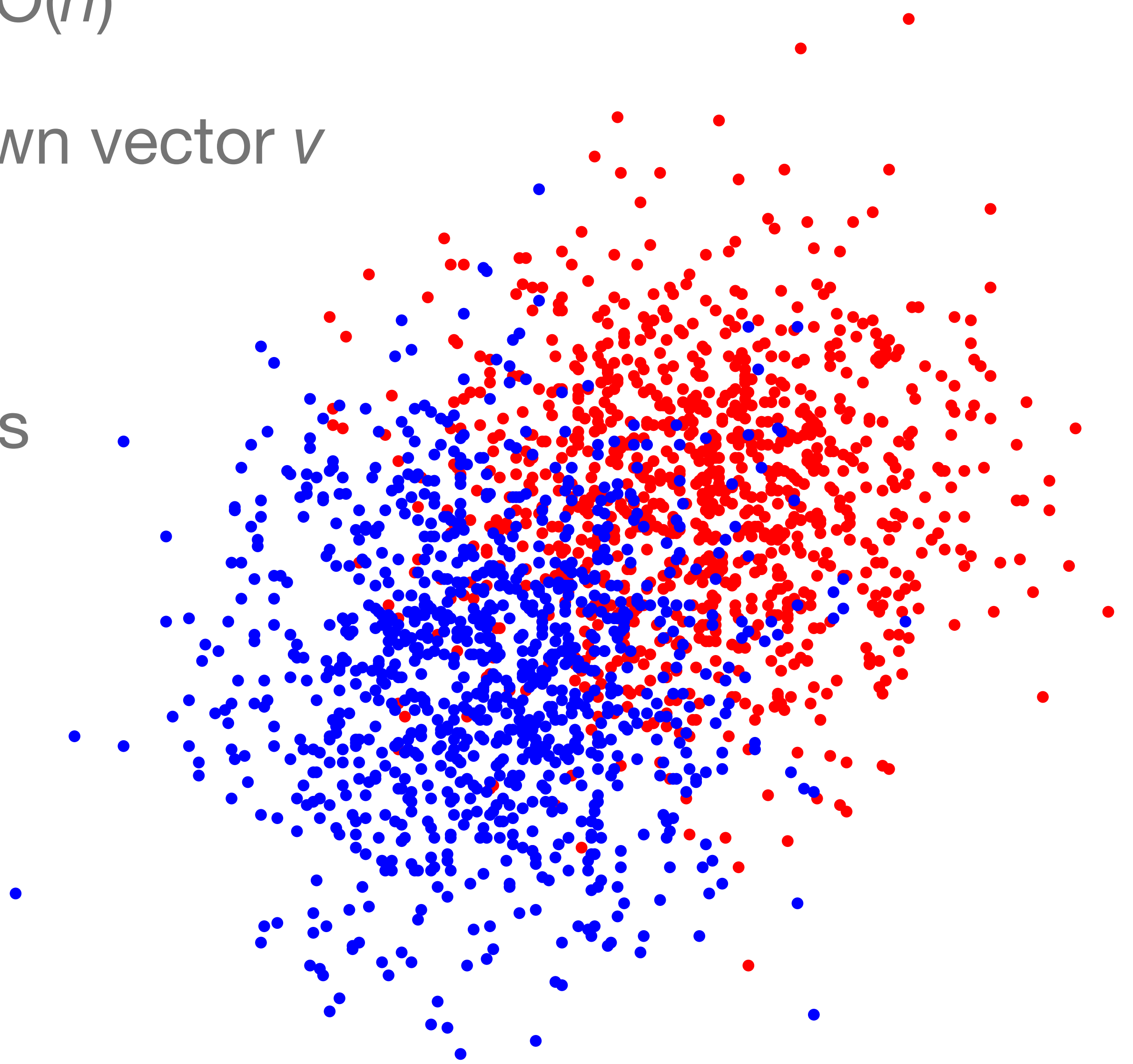
two clusters centered at $\pm v$ for some unknown vector v

each point has a Gaussian noise vector u_i

$|v|=O(1)$ but u_i has variance 1 along each axis

when can we...

- find v?
- label the points?
- confirm that there are two clusters?



Sparse, high-dimensional clustering

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

two clusters centered at $\pm v$ for some unknown vector v

each point has a Gaussian noise vector u_i

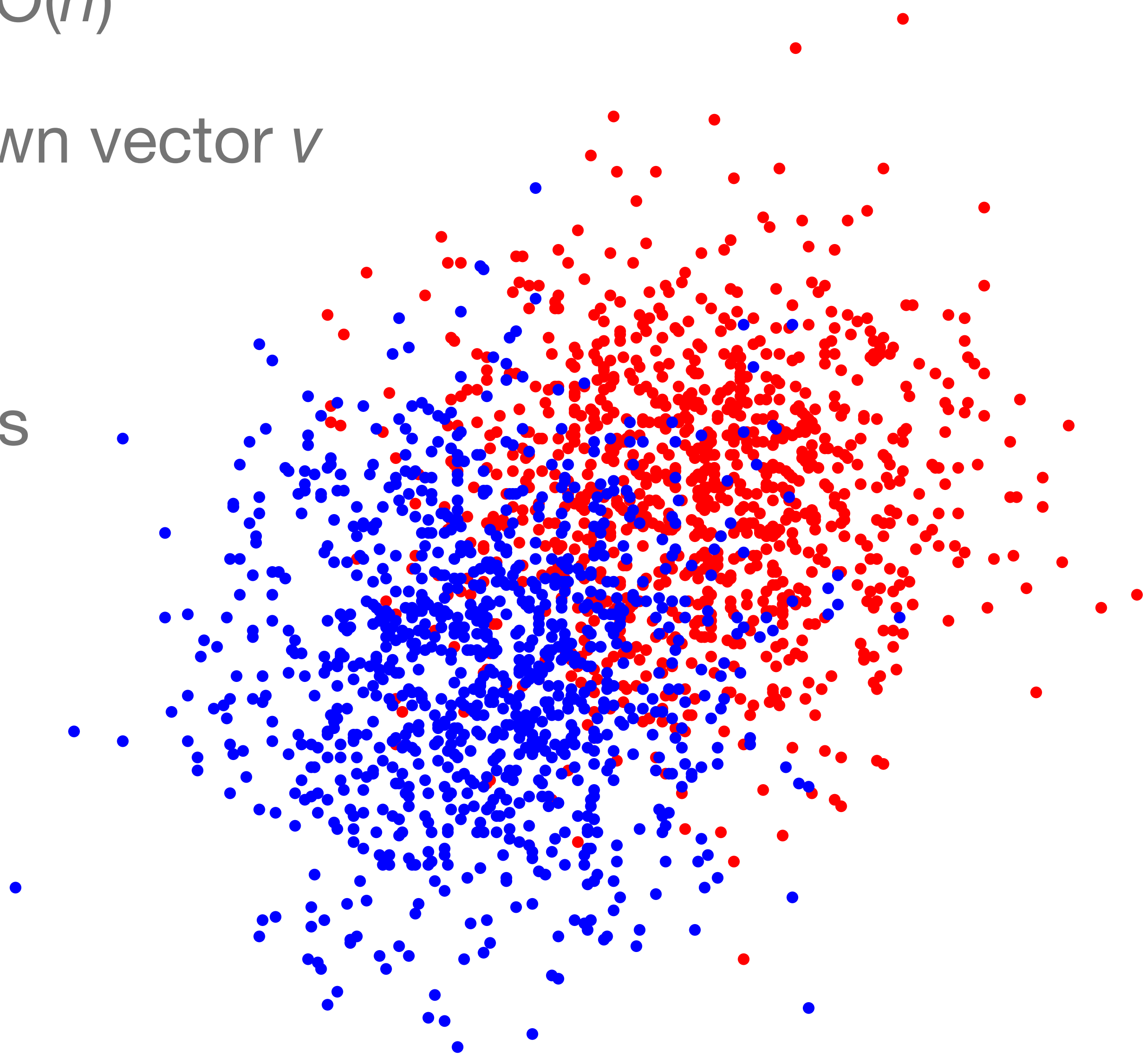
$|v|=O(1)$ but u_i has variance 1 along each axis

when can we...

find v?

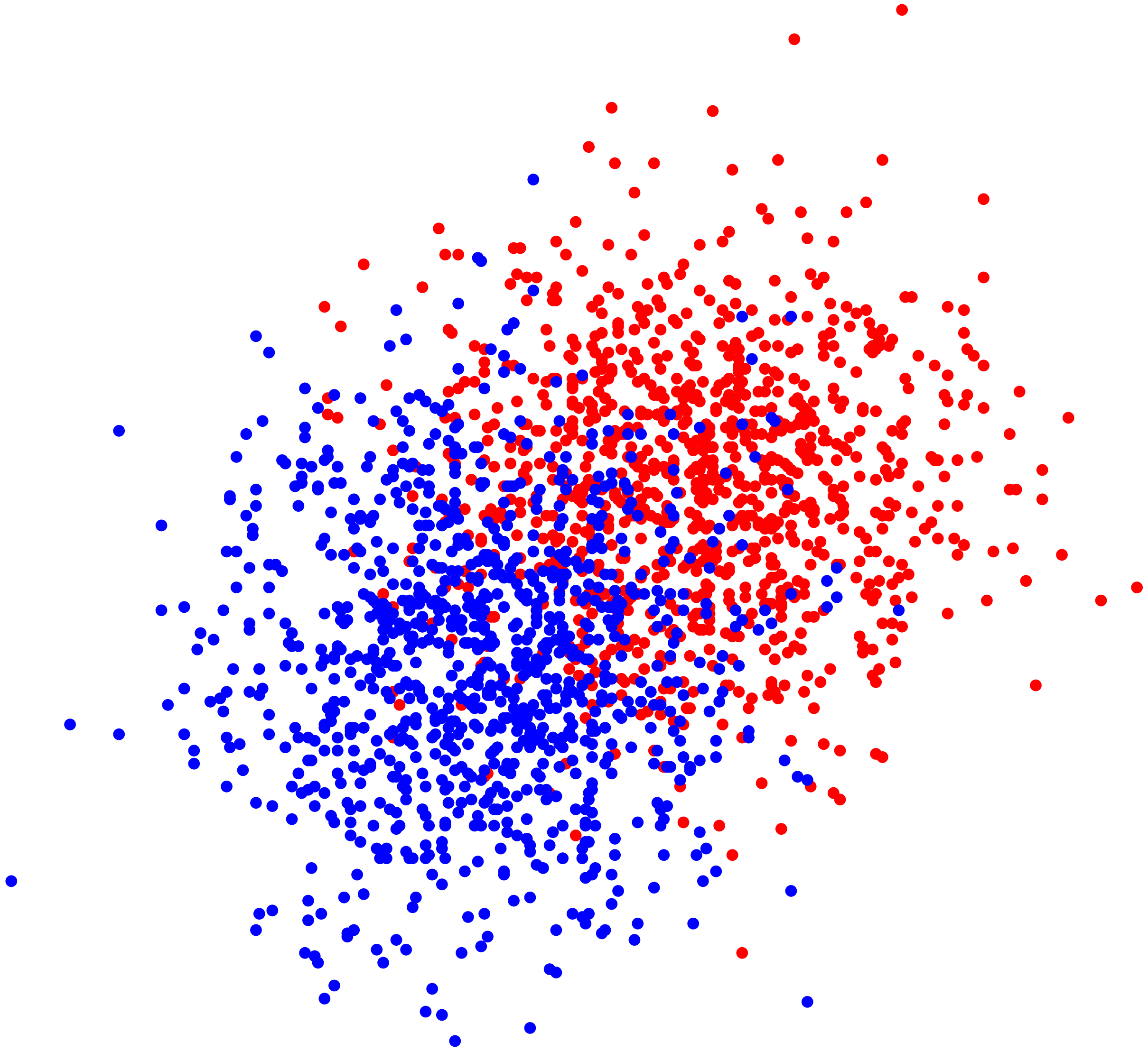
label the points?

confirm that there are two clusters?



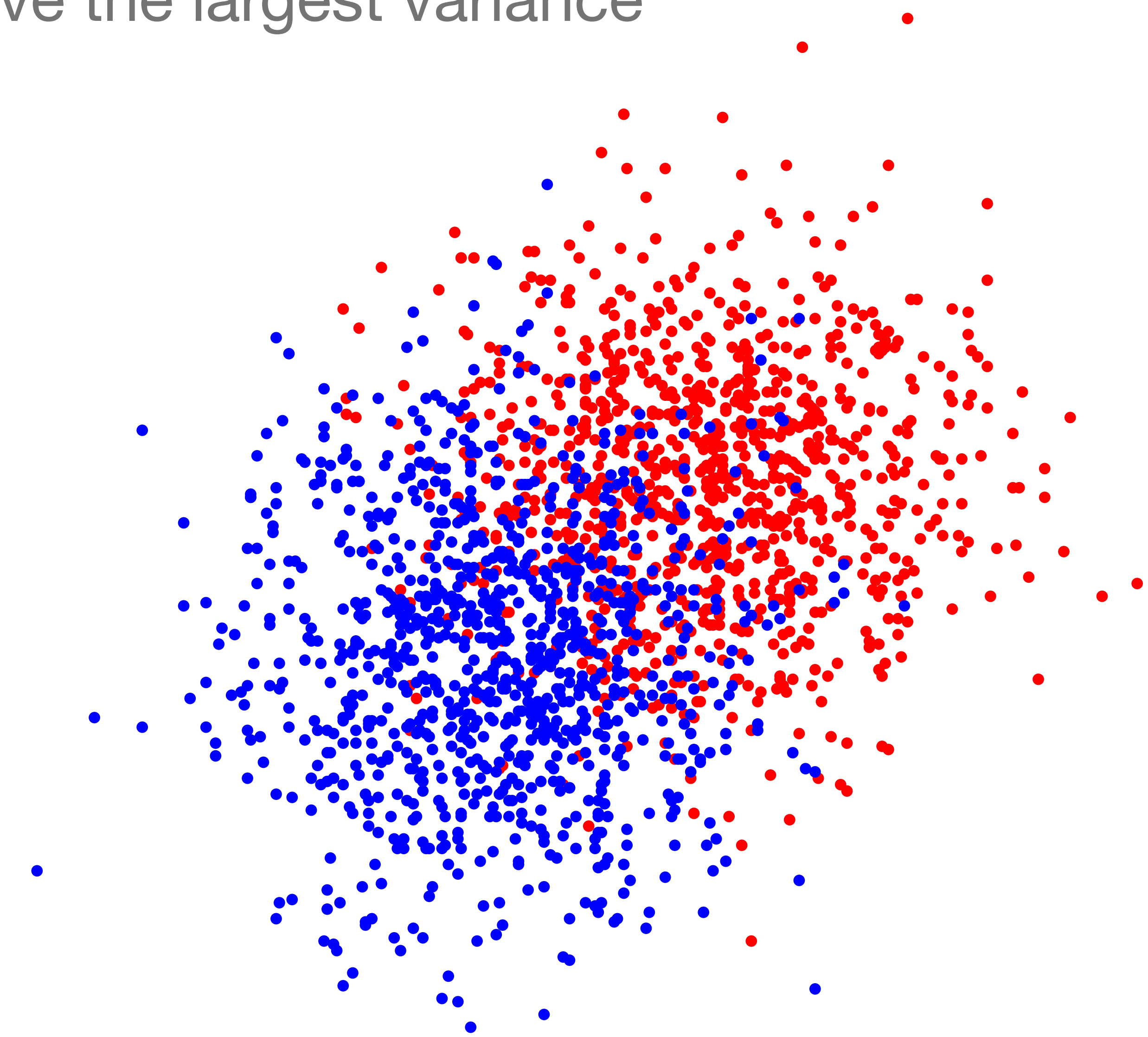
are there phase transitions as a function of $|v|$ vs. $|u|$ and m/n?

PCA (Principal Component Analysis)



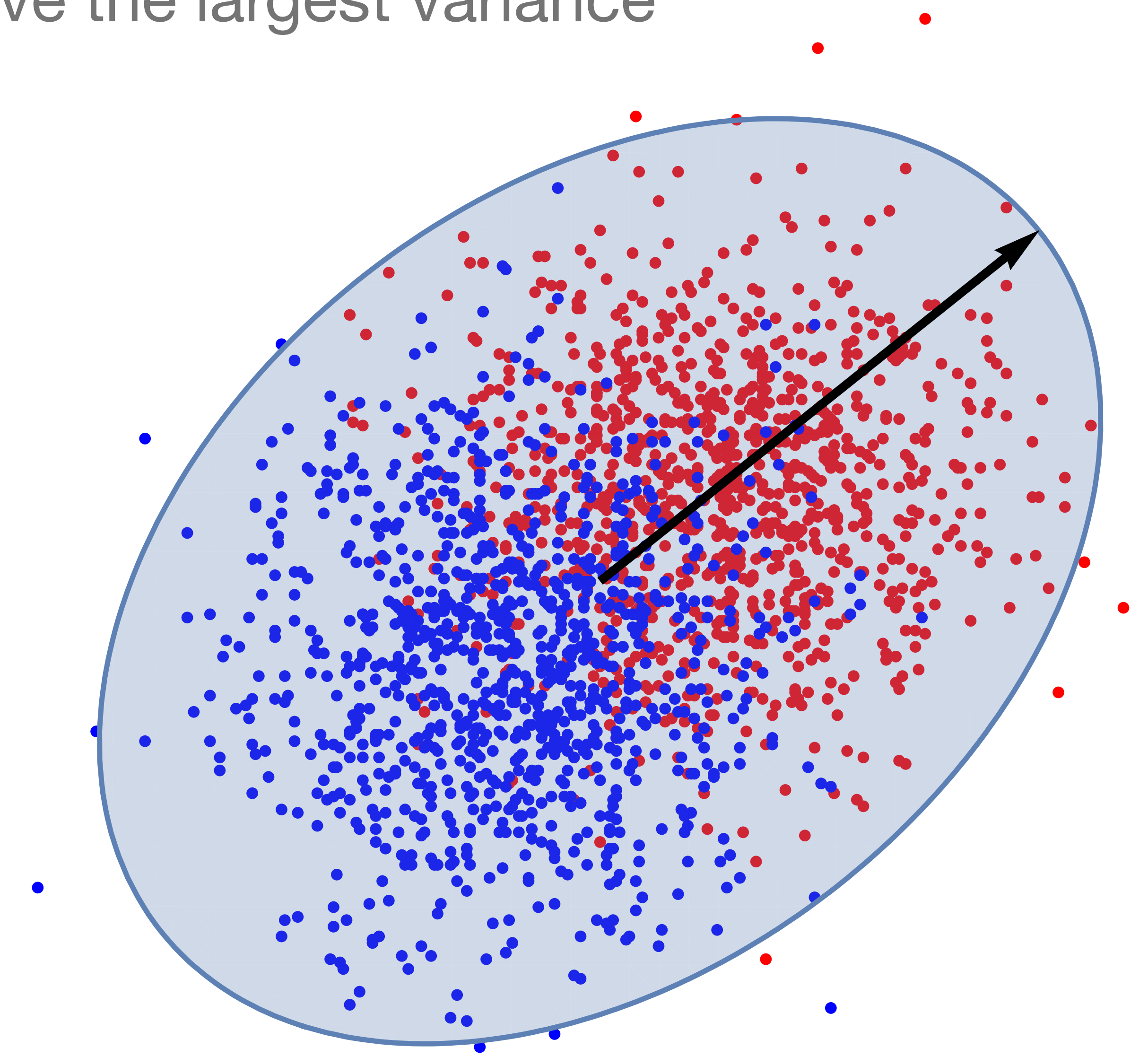
PCA (Principal Component Analysis)

find the direction along which the points have the largest variance



PCA (Principal Component Analysis)

find the direction along which the points have the largest variance

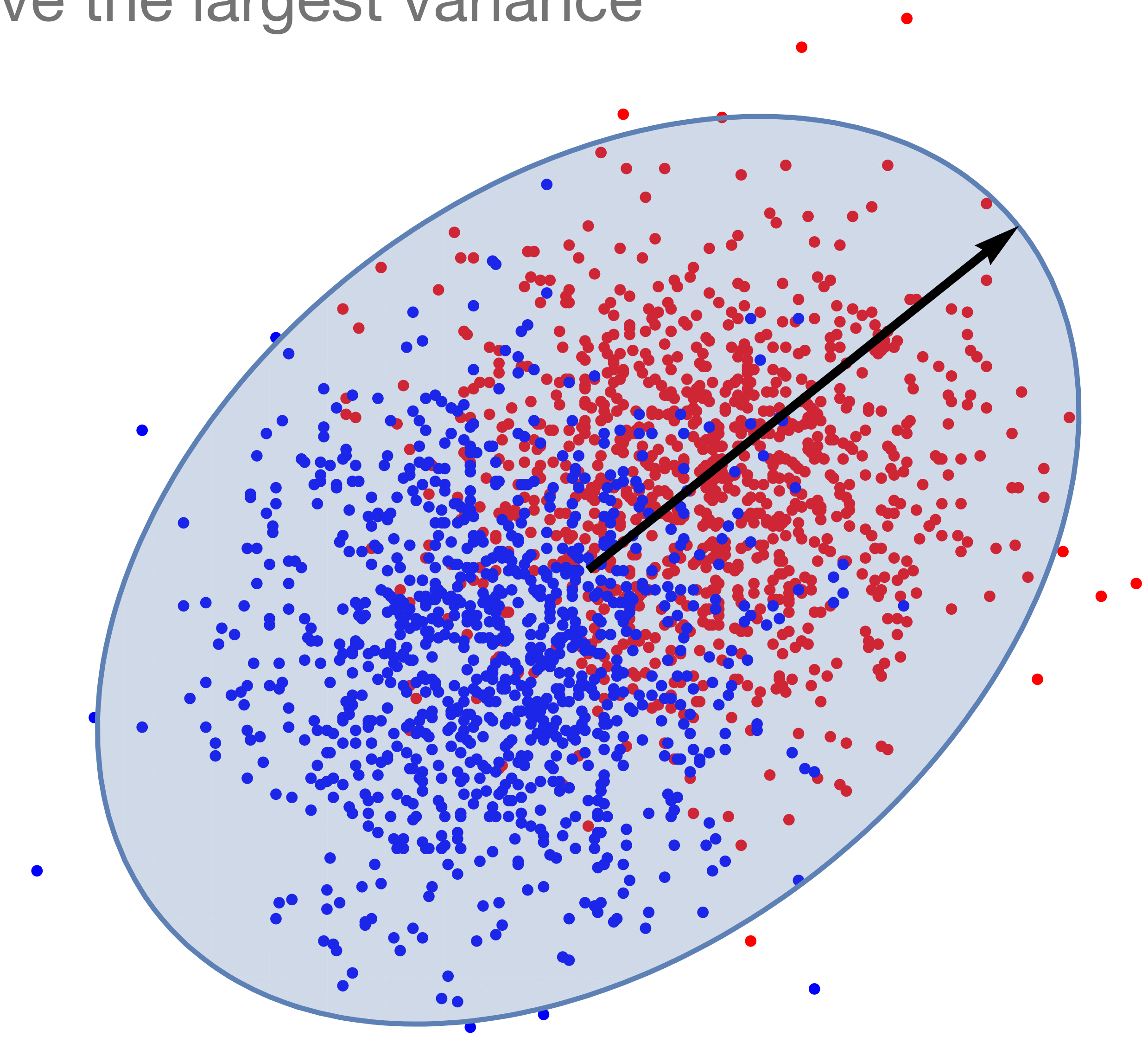


PCA (Principal Component Analysis)

find the direction along which the points have the largest variance

first eigenvector of the matrix

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



PCA (Principal Component Analysis)

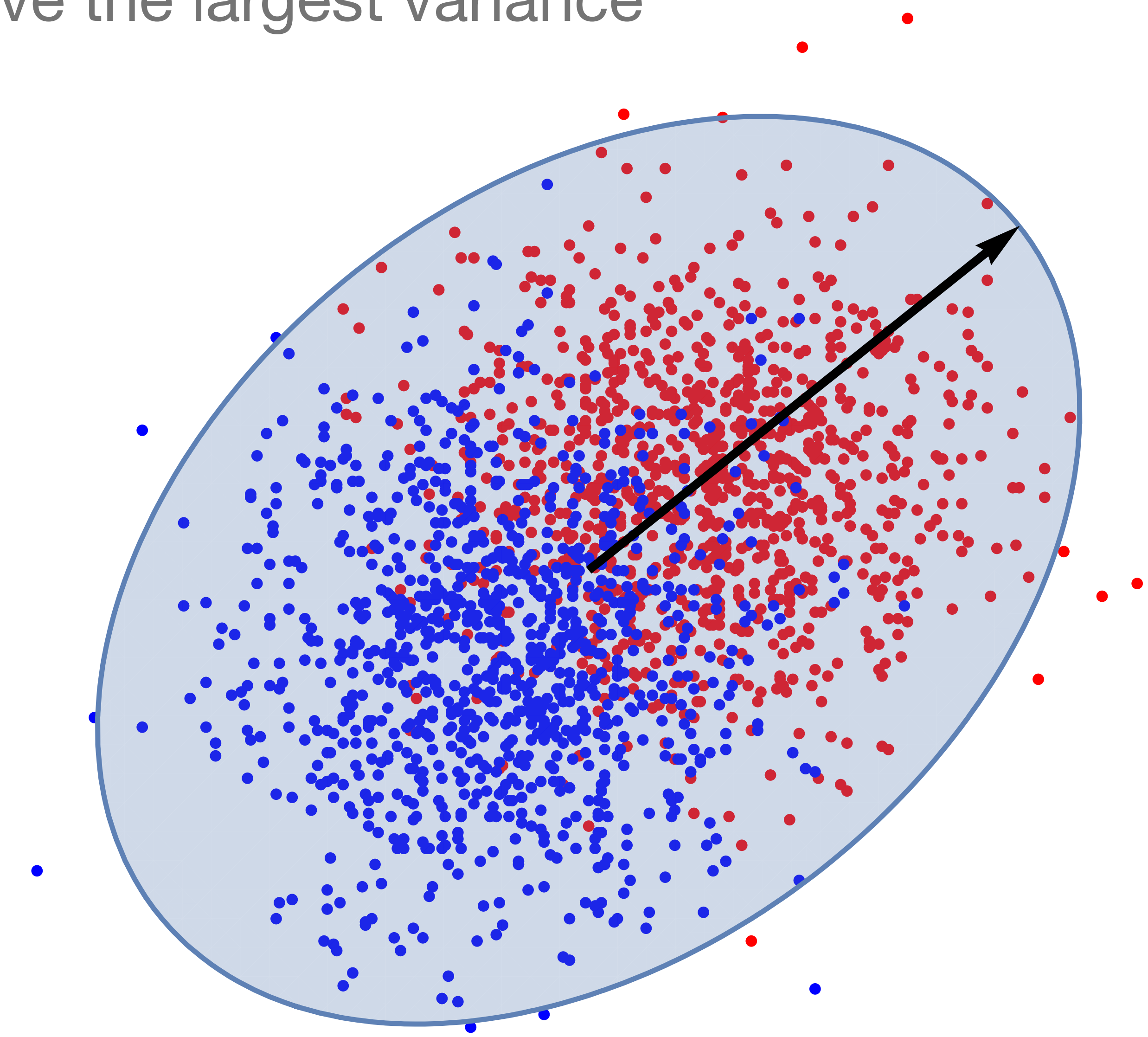
find the direction along which the points have the largest variance

first eigenvector of the matrix

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

this is a *Wishart random matrix*

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



PCA (Principal Component Analysis)

find the direction along which the points have the largest variance

first eigenvector of the matrix

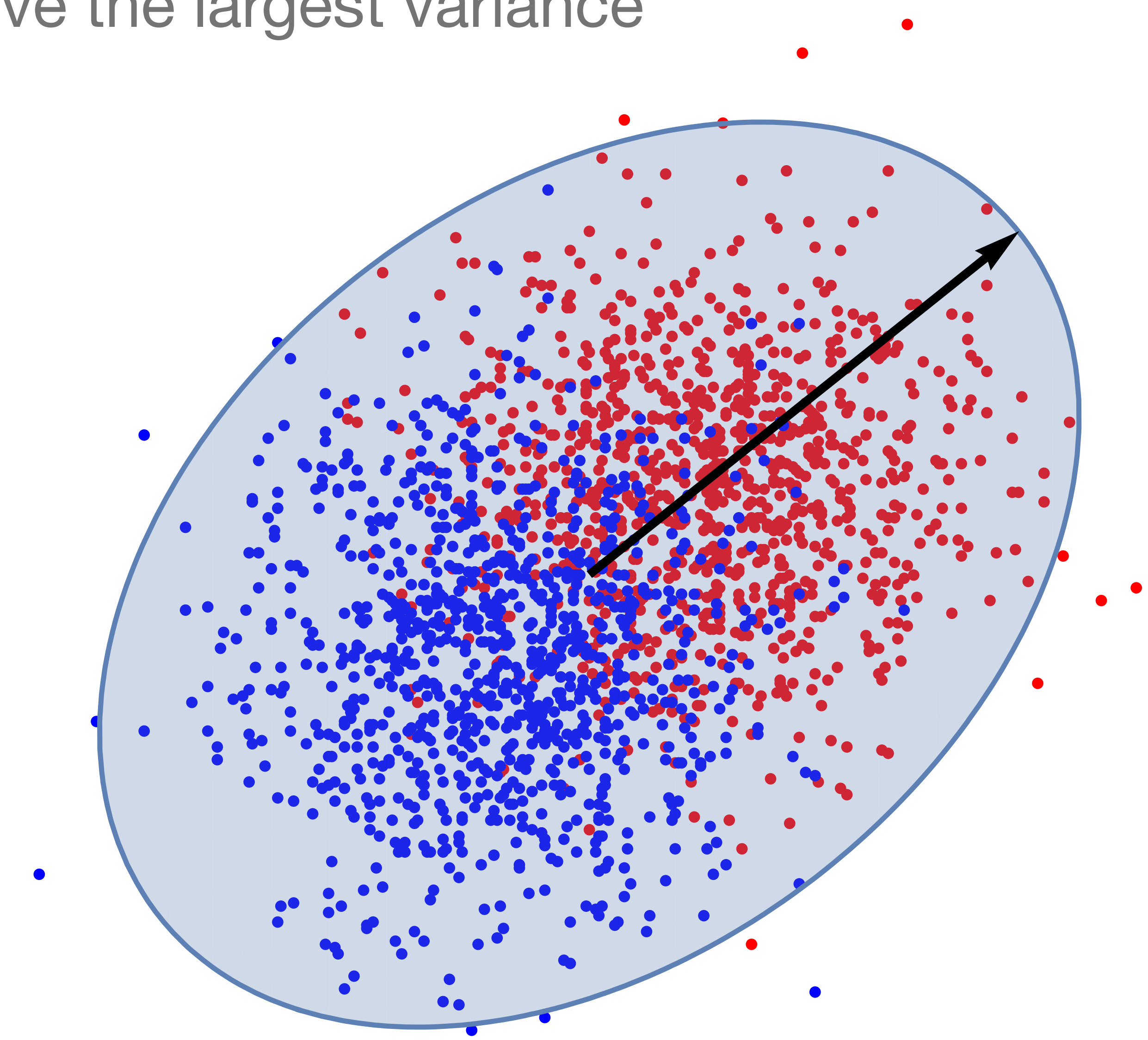
$$\frac{1}{m}\sum_{i=1}^{m} x_i \otimes x_i$$

this is a *Wishart random matrix*

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

plus a rank-1 perturbation

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



PCA (Principal Component Analysis)

find the direction along which the points have the largest variance

first eigenvector of the matrix

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

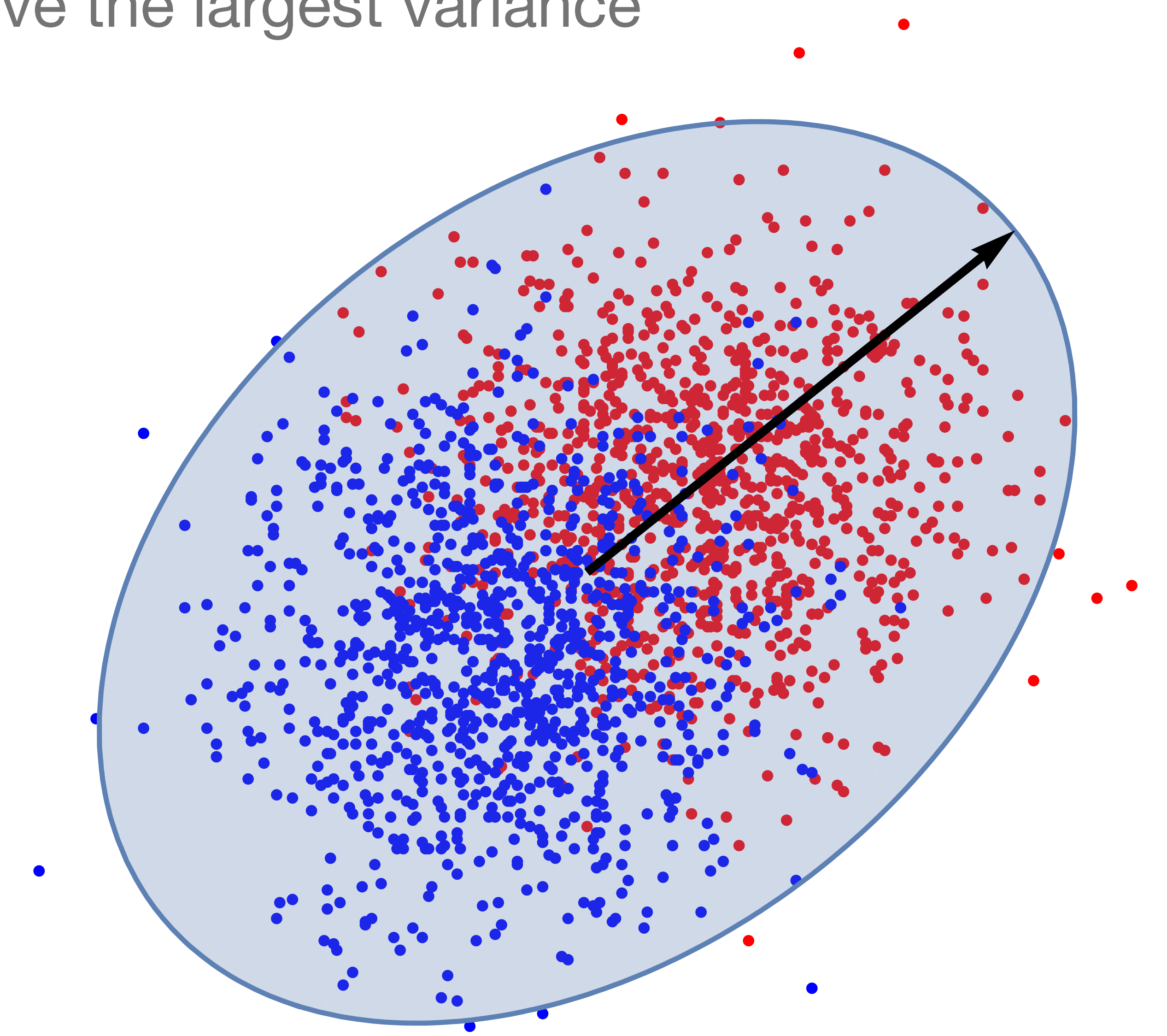
this is a *Wishart random matrix*

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

plus a rank-1 perturbation

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

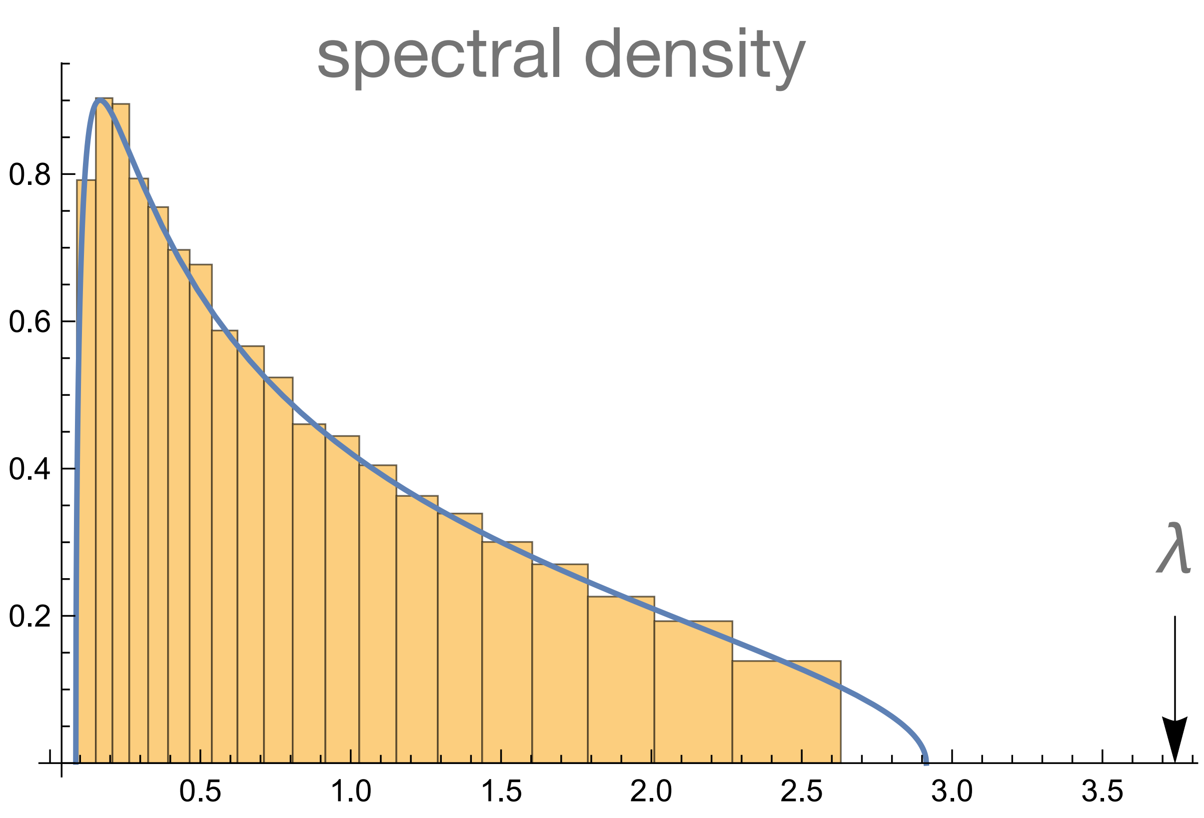
when does PCA work? and how accurately?



A phase transition

A phase transition

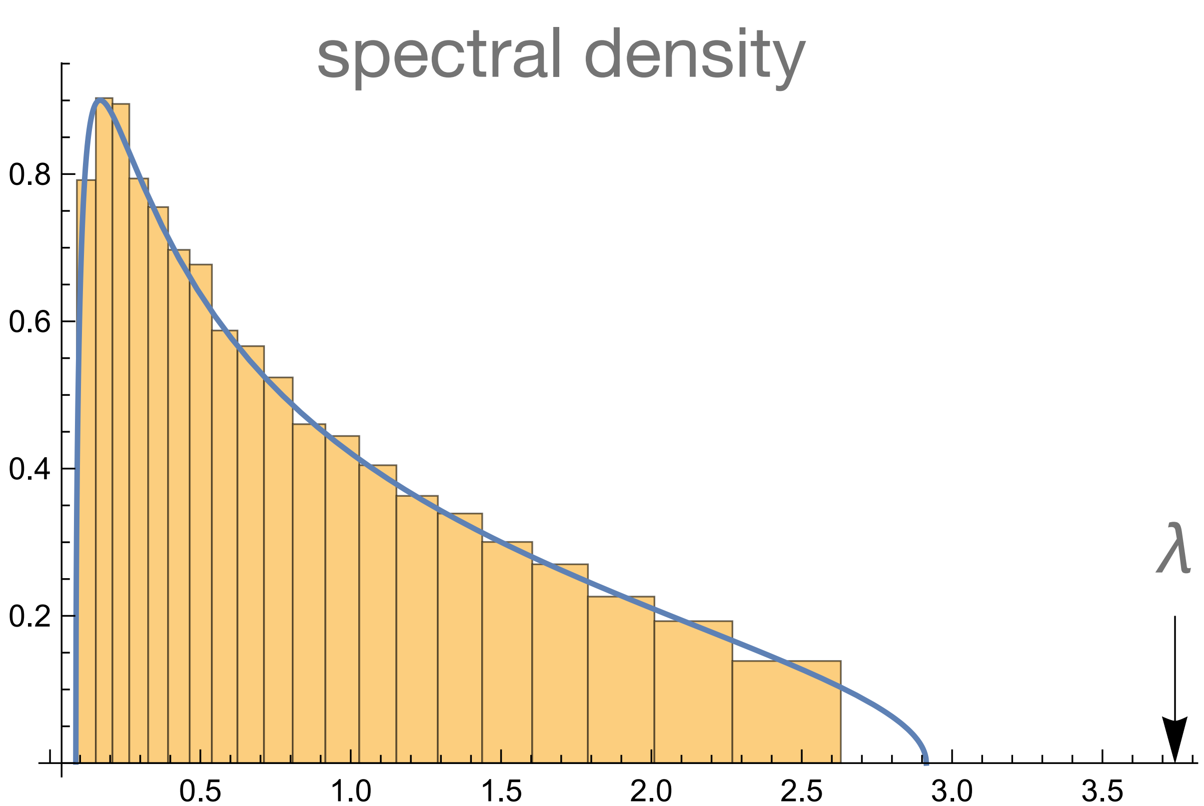
when does this perturbation rise above the "bulk" of random eigenvectors?



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?



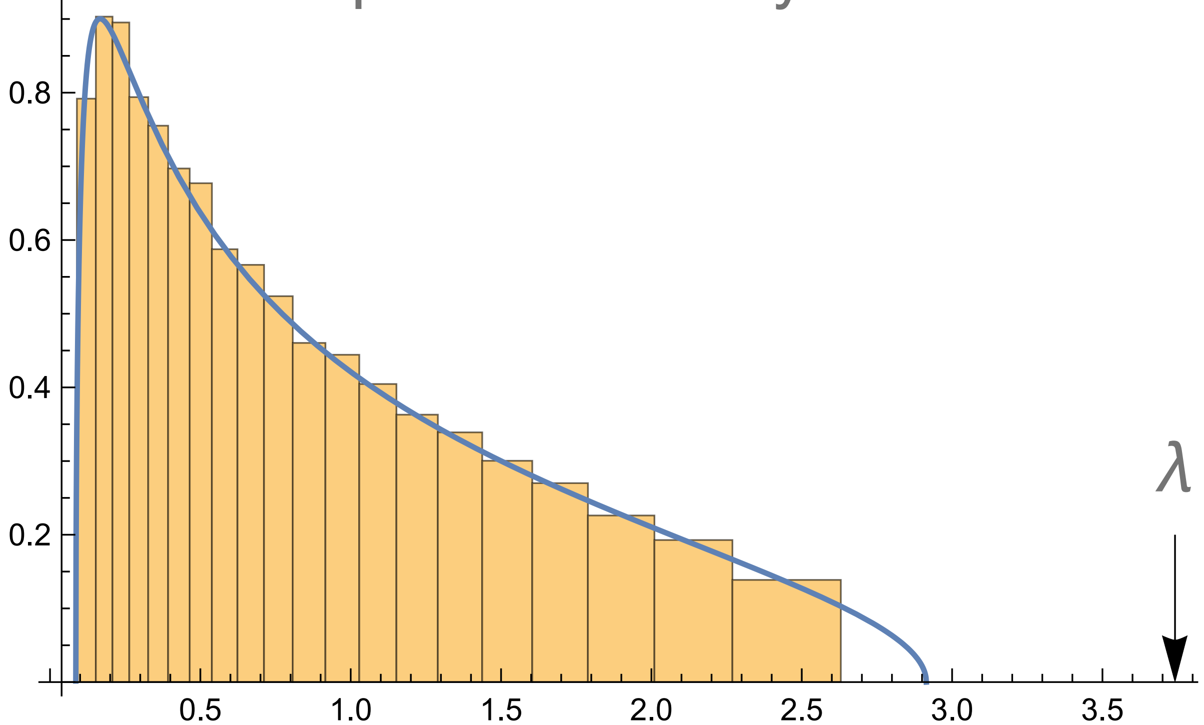
A phase transition

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

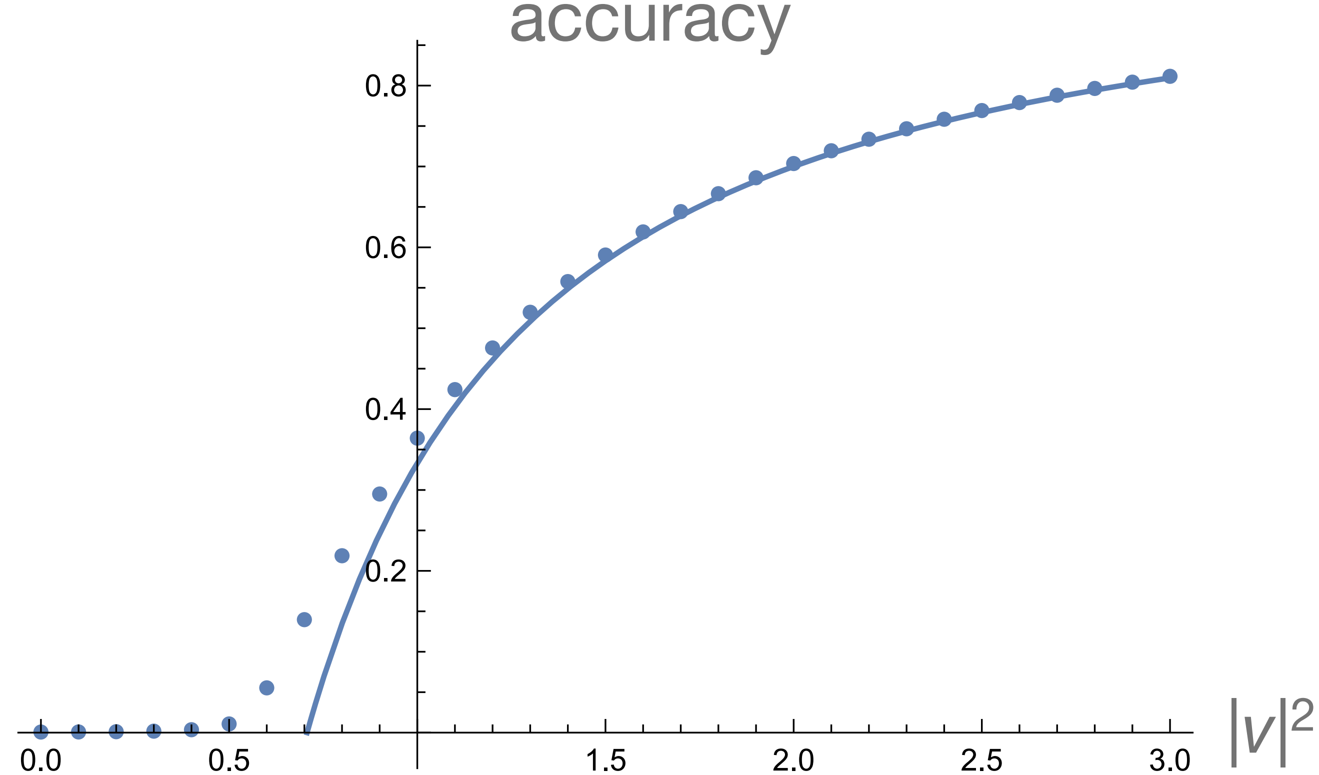
when it does, how accurately does the leading eigenvector point to v?

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

spectral density



accuracy



A phase transition

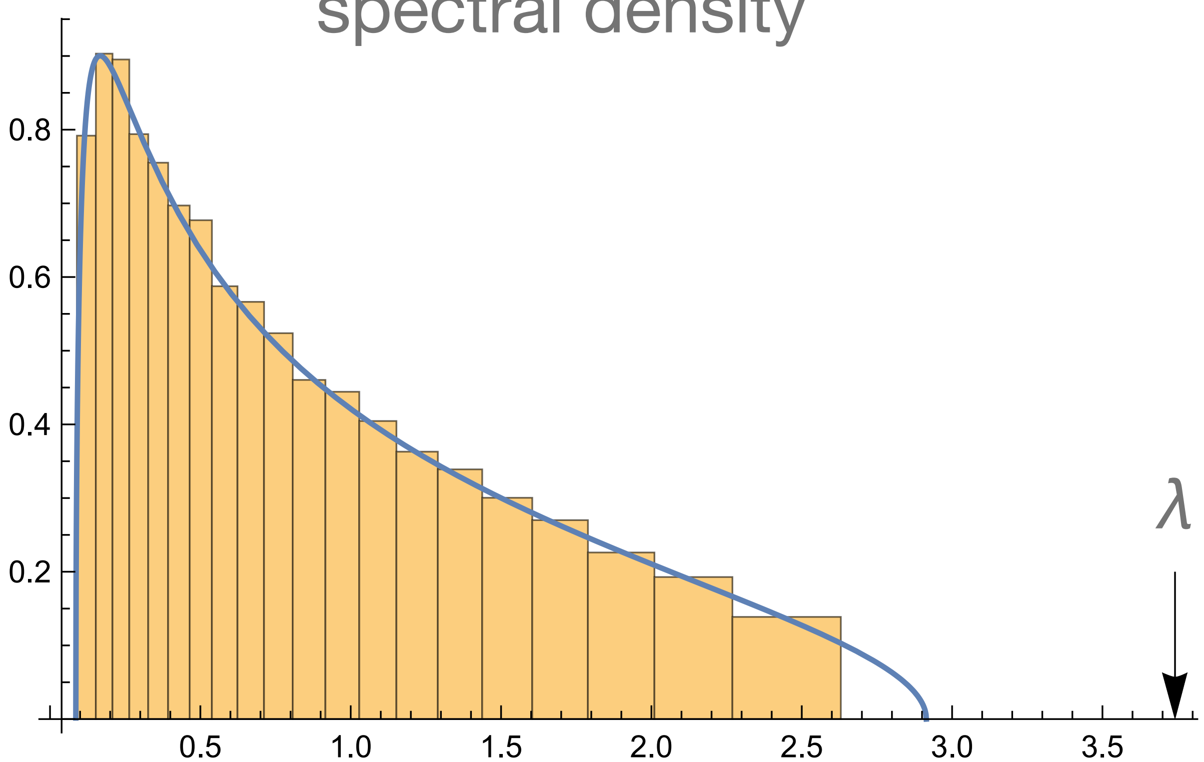
when does this perturbation rise above the “bulk” of random eigenvectors?

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

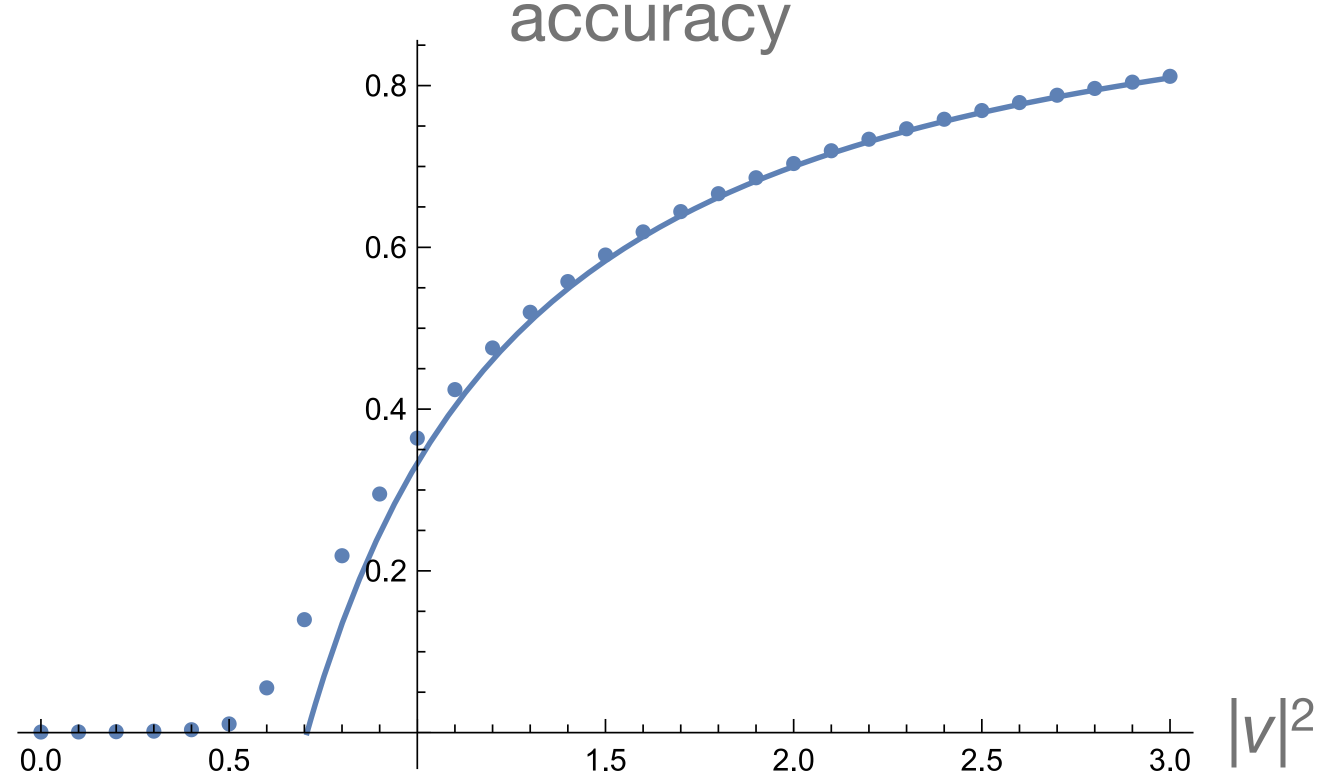
a phase transition at $m/n = 1/|v|^4$

but PCA isn’t optimal...

spectral density



accuracy



Morals: physics meets machine learning

Morals: physics meets machine learning

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

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

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

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

Morals: physics meets machine learning

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

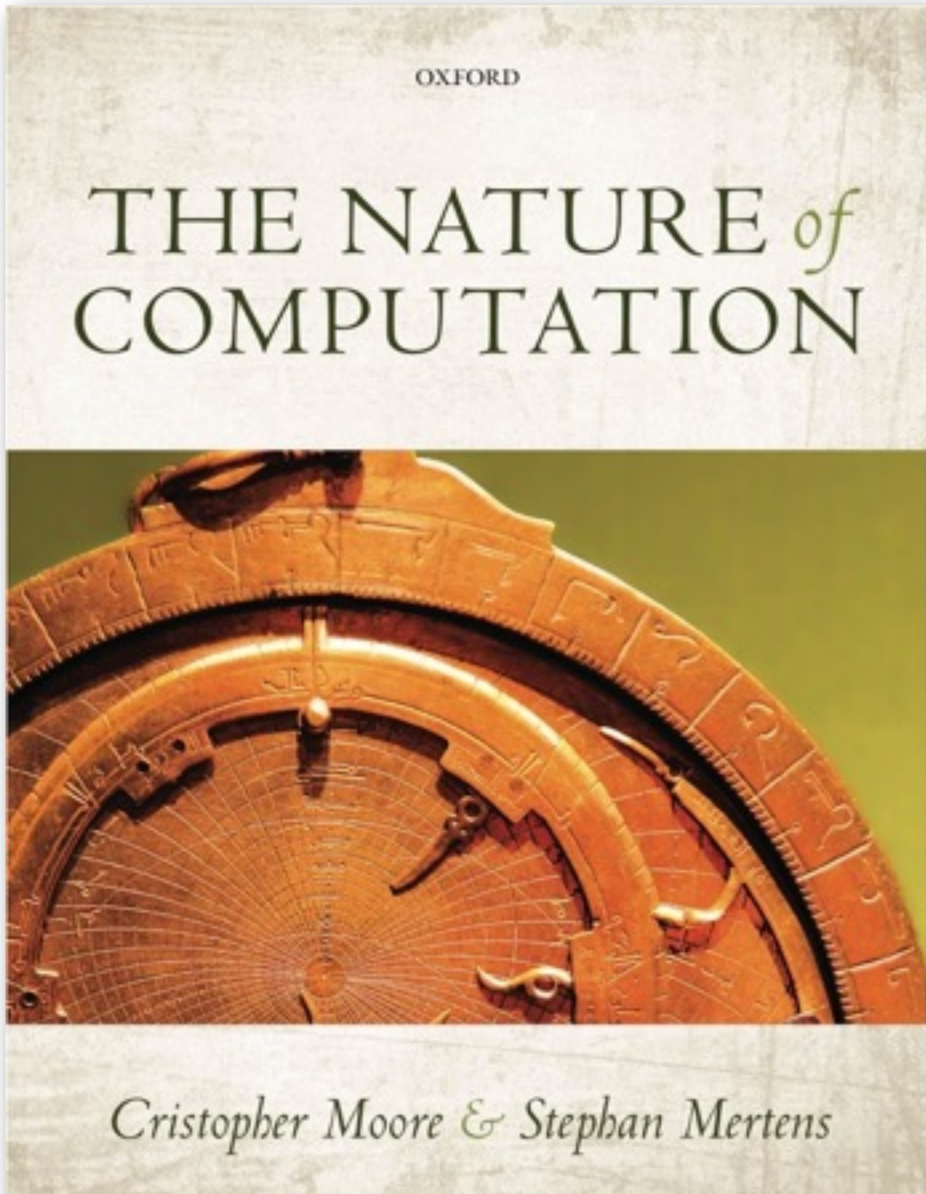
ideas from physics can help us find **optimal algorithms** that succeed all the way up to these transitions

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"

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