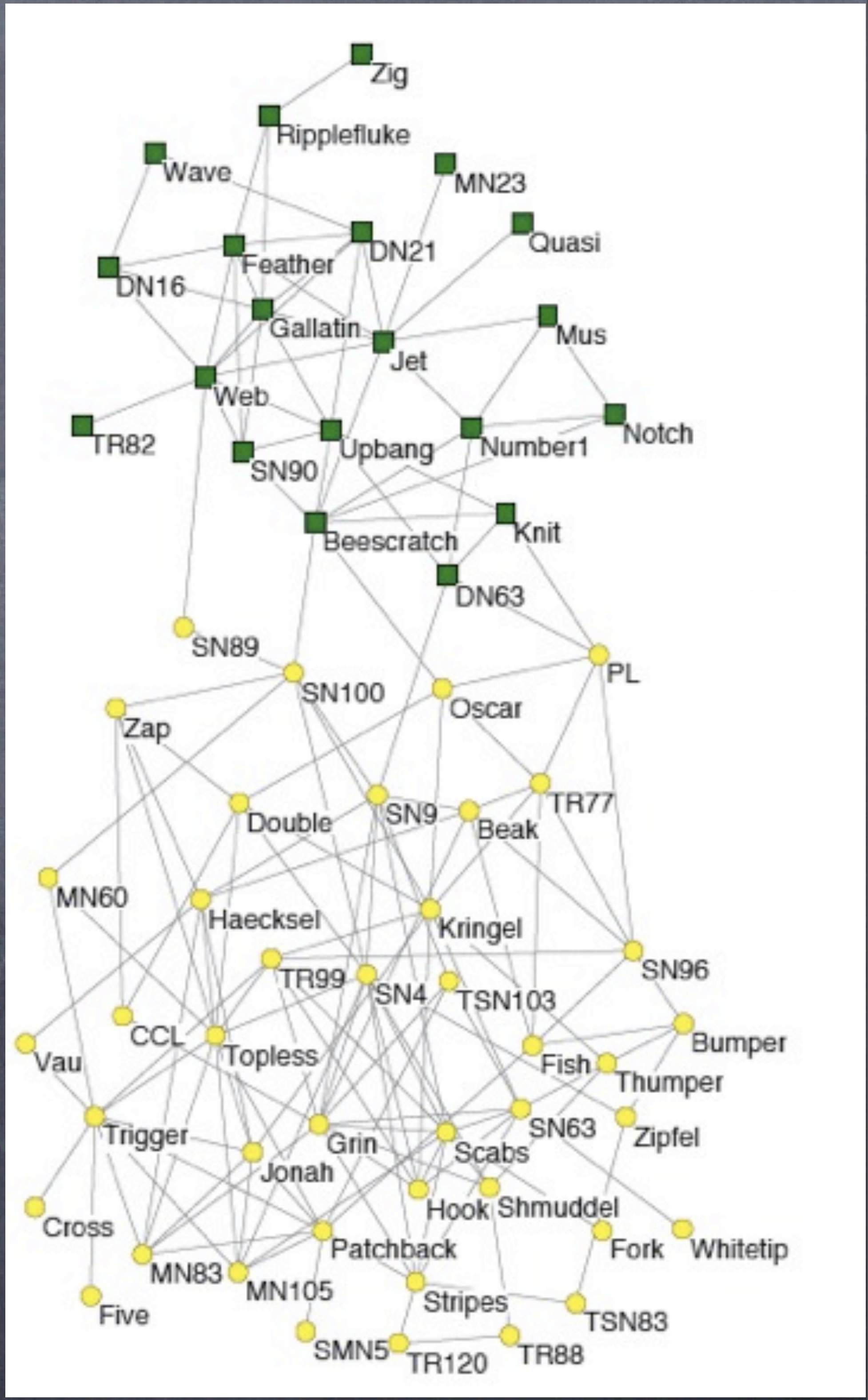
Easy, hard, and impossible inference

...and application to community detection

Florent Krzakala

Lenka Zdeborová (CEA Saclay)
Cris Moore (Santa Fe Inst.)
Aurelien Decelle (LPTMS Orsay) ➠ **talk later this afternoon**

Community structure...



... is observed in many systems:

- Online communities
- Word adjacency networks
- Food webs
- Metabolic networks
- Protein-protein interaction networks
- ...

<u>The problem:</u>

Predict the community structure from the topology of the network

(our) Motivations

- New algorithm for community detection (Bayesian inference using Belief Propagation)

- "Phase transitions" in inference/inverse problems ? (Hard, Easy, and Impossible as in 3-SAT?)

- Community detection is connected to many problems in inference, statistical physics and computer science:
  - Planted models, compressed sensing
  - Finite temperature decoding
  - Reconstruction on trees with noisy channels
  - Random optimization (coloring, partitioning...)
  - Spin glass and Nishimori symmetry
  - Glass transition vs first-order...

State of art

- Hundreds of papers on the topic *(Newman, Girvan'04, ...)*
- Maximize modularity function

$$Q = \frac{1}{2M} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2M} \right) \delta_{q_i, q_j}$$

State of art

- Hundreds of papers on the topic *(Newman, Girvan'04, ...)*
- Maximize modularity function

$$Q = \frac{1}{2M}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2M}\right)\delta_{q_i,q_j}$$

- Problem: this method (and virtually any method in the literature) is unable to tell that a random graph does not have any communities.

State of art

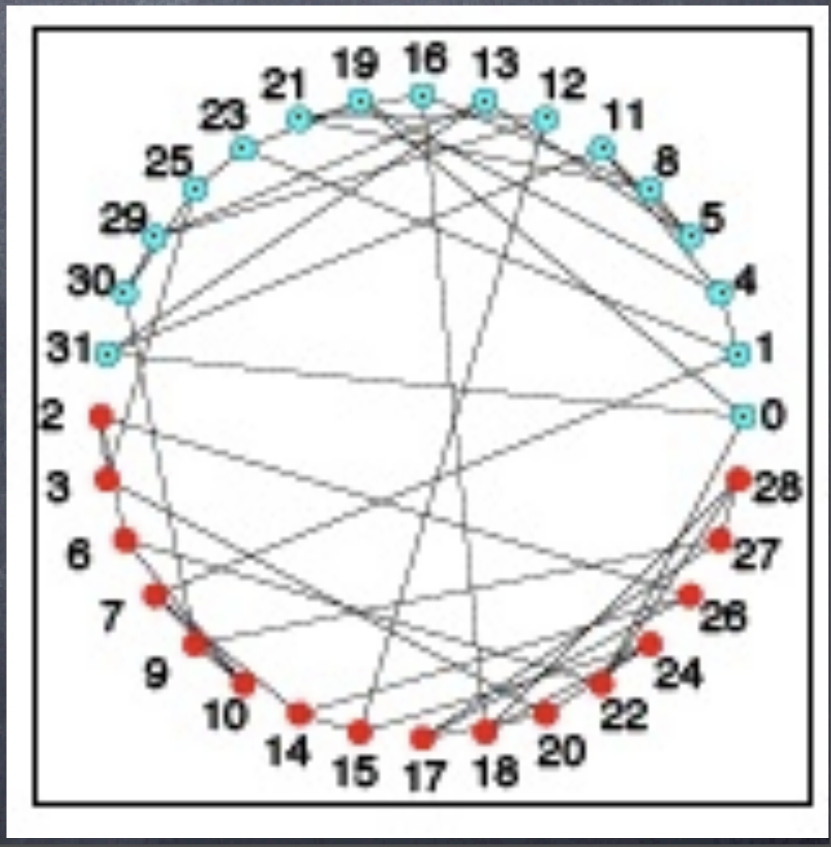
- Hundreds of papers on the topic *(Newman, Girvan'04, ...)*
- Maximize modularity function

$$Q = \frac{1}{2M} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2M} \right) \delta_{q_i, q_j}$$

- Problem: this method (and virtually any method in the literature) is unable to tell that a random graph does not have any communities.

Example:

Ising model on a 3-regular random graphs
Best bisection looks like a good clustering
(only 11% of edges between the 2 groups)



State of art

- Problem: this method (and virtually any method in the literature) is unable to tell that a random graph does not have any communities.

State of art

- Problem: this method (and virtually any method in the literature) is unable to tell that a random graph does not have any communities.
- Missing measures of significance, estimates of probability of error ...

State of art

- Problem: this method (and virtually any method in the literature) is unable to tell that a random graph does not have any communities.
- Missing measures of significance, estimates of probability of error ...
- Maximizing inter-connections? But nodes of the same kind are not always inter-connected (e.g. food-web, adjacency of words in text...), and can also be directed.

State of art

- Problem: this method (and virtually any method in the literature) is unable to tell that a random graph does not have any communities.
- Missing measures of significance, estimates of probability of error ...
- Maximizing inter-connections? But nodes of the same kind are not always inter-connected (e.g. food-web, adjacency of words in text...), and can also be directed.
- Equal group sizes? There is no reason for this a priori...

State of art

- Problem: this method (and virtually any method in the literature) is unable to tell that a random graph does not have any communities.
- Missing measures of significance, estimates of probability of error ...
- Maximizing inter-connections? But nodes of the same kind are not always inter-connected (e.g. food-web, adjacency of words in text...), and can also be directed.
- Equal group sizes? There is no reason for this a priori...

Need for a more fundamental, and principled approach:
Let's switch to Bayesian inference, and synthetic data

The Block model

Generate a random network as follows:

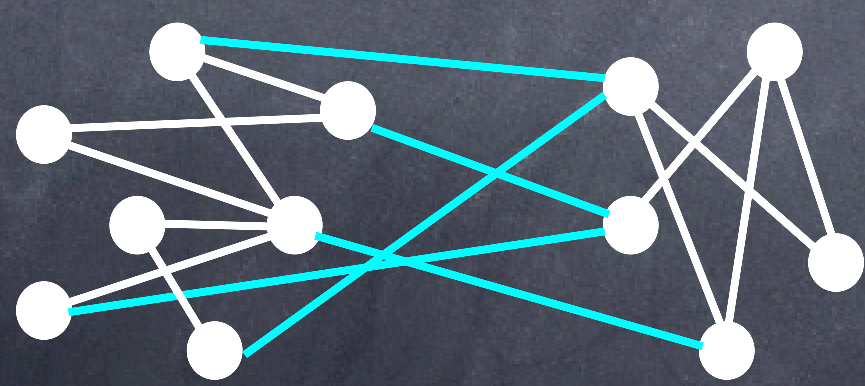
- q groups, N nodes
- n_a proportion of nodes in group $a = 1, \dots, q$
- $p_{ab} = \frac{c_{ab}}{N}$ probability that an edge present between node from group a and another from group b

The Block model

Generate a random network as follows:

- q groups, N nodes
- n_a proportion of nodes in group $a = 1, \ldots, q$
- $p_{ab} = \dfrac{c_{ab}}{N}$ probability that an edge present between node from group a and another from group b

$n_1 = 7/12 \qquad n_2 = 5/12$



$p_{11} = p_{22} = 0.39$

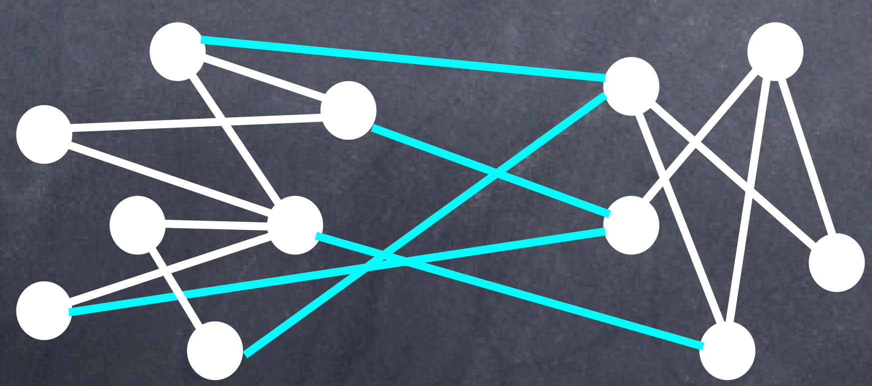
$p_{12} = p_{21} = 0.14$

The Block model

Generate a random network as follows:

- q groups, N nodes
- n_a proportion of nodes in group $a = 1, \ldots, q$
- $p_{ab} = \dfrac{c_{ab}}{N}$ probability that an edge present between node from group a and another from group b

$n_1 = 7/12 \qquad n_2 = 5/12$



$p_{11} = p_{22} = 0.39$
$p_{12} = p_{21} = 0.14$

I am giving you the network, can you infer the values of q, n_a and p_{ab}? Can you detect the original assignment?

A Bayesian Approach to community detection

$$P(\{n_a, p_{ab}\}|G) = \frac{P(\{n_a, p_{ab}\})}{P(G)} P(G|\{n_a, p_{ab}\})$$

$$= \frac{P(\{n_a, p_{ab}\})}{P(G)} \sum_{\{q_i\}} P(G, \{q_i\}|\{n_a, p_{ab}\})$$

A Bayesian Approach to community detection

$$P(\{n_a, p_{ab}\}|G) = \frac{P(\{n_a, p_{ab}\})}{P(G)} P(G|\{n_a, p_{ab}\})$$

$$= \frac{P(\{n_a, p_{ab}\})}{P(G)} \sum_{\{q_i\}} P(G, \{q_i\}|\{n_a, p_{ab}\})$$

$$P(G, \{q_i\}|\{n_a, p_{ab}\}) = \prod_{i=1}^{N} n_{q_i} \prod_{ij} p_{q_i q_j}^{A_{ij}} (1 - p_{q_i q_j})^{1-A_{ij}}$$

A Bayesian Approach to community detection

$$
\begin{aligned}
P(\{n_a, p_{ab}\}|G) &= \frac{P(\{n_a, p_{ab}\})}{P(G)} P(G|\{n_a, p_{ab}\}) \\
&= \frac{P(\{n_a, p_{ab}\})}{P(G)} \sum_{\{q_i\}} P(G, \{q_i\}|\{n_a, p_{ab}\})
\end{aligned}
$$

$$
P(G, \{q_i\}|\{n_a, p_{ab}\}) = \prod_{i=1}^{N} n_{q_i} \prod_{ij} p_{q_i q_j}^{A_{ij}} (1 - p_{q_i q_j})^{1-A_{ij}}
$$

$$
Z(\{n_a, p_{ab}\}) \equiv \sum_{\{q_i\}} P(G, \{q_i\}|\{n_a, p_{ab}\})
$$

Maximize Z to learn $\{n_a, p_{ab}\}$

A Bayesian Approach to community detection

$$Z(\{n_a, p_{ab}\}) \equiv \sum_{\{q_i\}} P(G, \{q_i\} | \{n_a, p_{ab}\})$$

Maximize Z to learn $\{n_a, p_{ab}\}$

$$P(G, \{q_i\} | \{n_a, p_{ab}\}) = \prod_{i=1}^{N} n_{q_i} \prod_{ij} p_{q_i q_j}^{A_{ij}} (1 - p_{q_i q_j})^{1-A_{ij}}$$

A Bayesian Approach to community detection

$$Z(\{n_a, p_{ab}\}) \equiv \sum_{\{q_i\}} P(G, \{q_i\} | \{n_a, p_{ab}\})$$

Maximize Z to learn $\{n_a, p_{ab}\}$

$$P(G, \{q_i\} | \{n_a, p_{ab}\}) = \prod_{i=1}^{N} n_{q_i} \prod_{ij} p_{q_i q_j}^{A_{ij}} (1 - p_{q_i q_j})^{1 - A_{ij}}$$

Equilibrium statistical physics of the Hamiltonian:

$$\begin{aligned} -H(\{q_i\}) &= \sum_{i=1}^{N} \log n_{q_i} + \sum_{ij} \left[A_{ij} \log p_{q_i q_j} + (1 - A_{ij}) \log (1 - p_{q_i q_j}) \right] \\ &= \sum_{i=1}^{N} \log n_{q_i} + \sum_{(ij) \in E} \log \frac{p_{q_i q_j}}{1 - p_{q_i q_j}} + \sum_{a,b=1}^{q} N_a N_b \log (1 - p_{ab}) \end{aligned}$$

A Bayesian Approach to community detection

Once the parameters $\{n_a, p_{ab}\}$ have been inferred:

- A configuration sampled from the Boltzmann measure has the correct group sizes and number of connections between groups

- The configuration overlapping the most with the original assignment is obtained by computing marginals (local magnetizations) and taking the most probable value.

(as in finite temperature decoding Nishimori'93, Sourlas'94)

$$P(G,\{q_i\}|\{n_a,p_{ab}\}) = \prod_{i=1}^{N} n_{q_i} \prod_{ij} p_{q_i q_j}^{A_{ij}} (1-p_{q_i q_j})^{1-A_{ij}}$$

A Bayesian Approach to community detection

(1) Compute averages:

- ➡ With Monte Carlo (detailed balance) slow....
- ➡ With Belief Propagation faster and **exact** for large networks generated by the block model

(2) Update parameters to perform a steepest ascent

$$n_a = \frac{1}{N}\left\langle \sum_i \delta_{a,q_i} \right\rangle . \qquad p_{ab} n_a n_b = \frac{1}{N^2}\left\langle \sum_{(ij)\in E} \delta_{a,q_i}\delta_{b,q_i} \right\rangle .$$

(3) Repeat until convergence.

(4) Assign the most probable value: $q_i = \mathrm{argmax}_{q_i} P_i(q_i)$

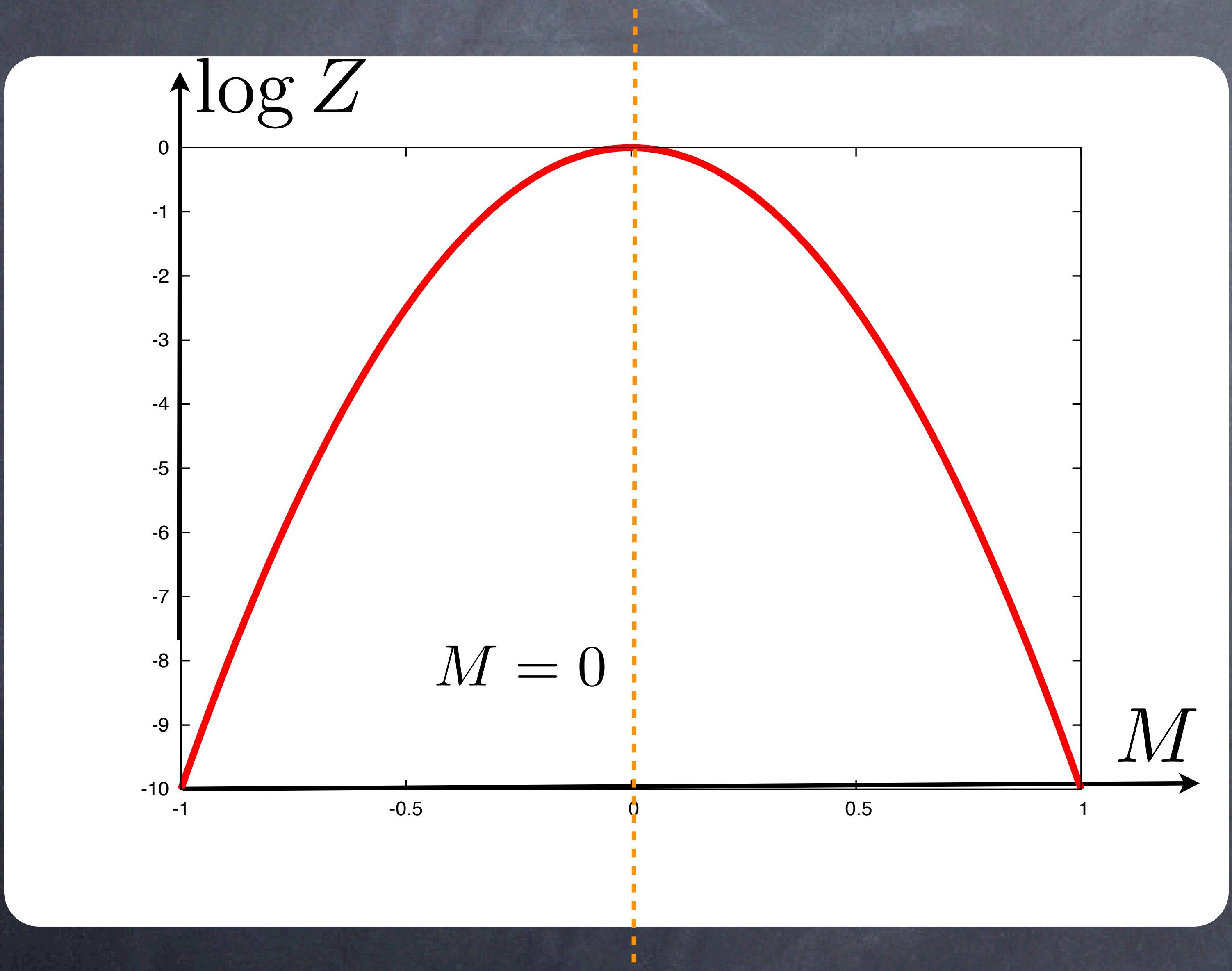
Phase transition in Community detection for the Block model

Consider the a priori difficult cases where each community has the average degree

⇓

3 different cases may arise depending on the parameters used to generate the network

(1) The paramagnetic case: Impossible inference



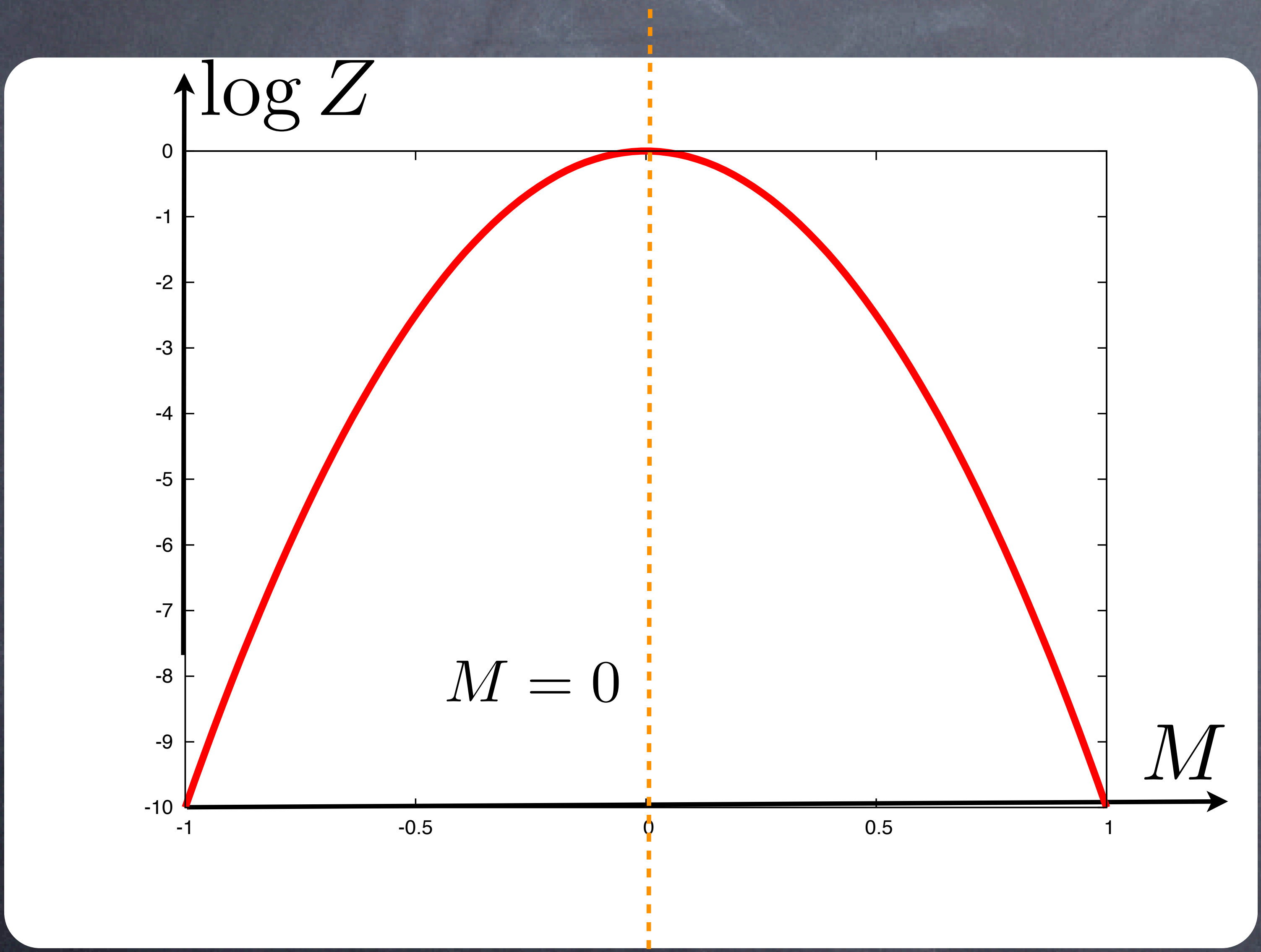
M is the (normalized) overlap with the original assignment

Assume we know the correct parameters $\{n_a, p_{ab}\}$

The maximum partition sum is obtained for trivial "paramagnetic" marginals

$$P_i(q) = n_q \quad \forall i$$

(1) The paramagnetic case: Impossible inference



M is the (normalized) overlap with the original assignment

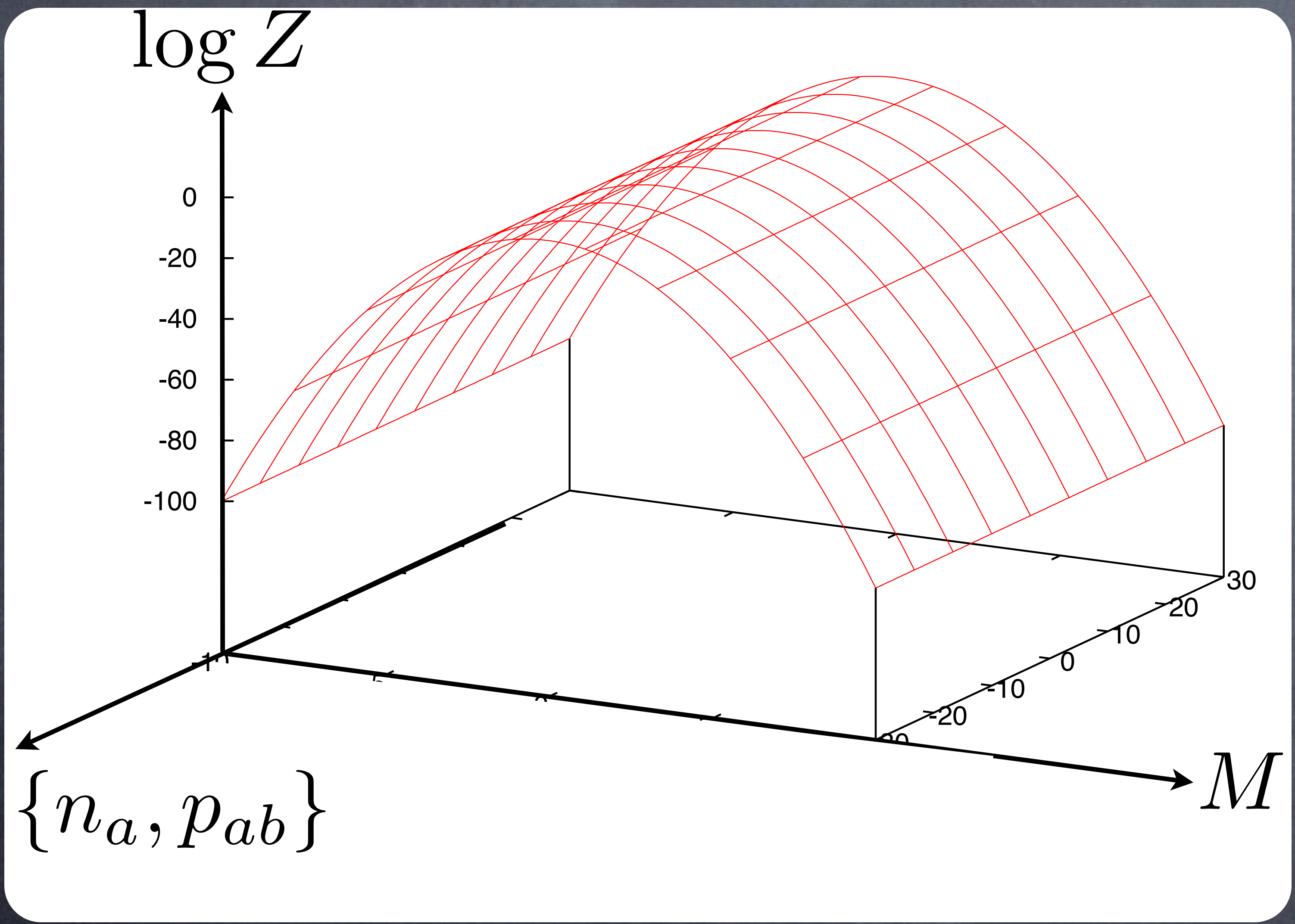
Assume we know the correct parameters $\{n_a, p_{ab}\}$

The maximum partition sum is obtained for trivial "paramagnetic" marginals

$$P_i(q) = n_q \quad \forall i$$

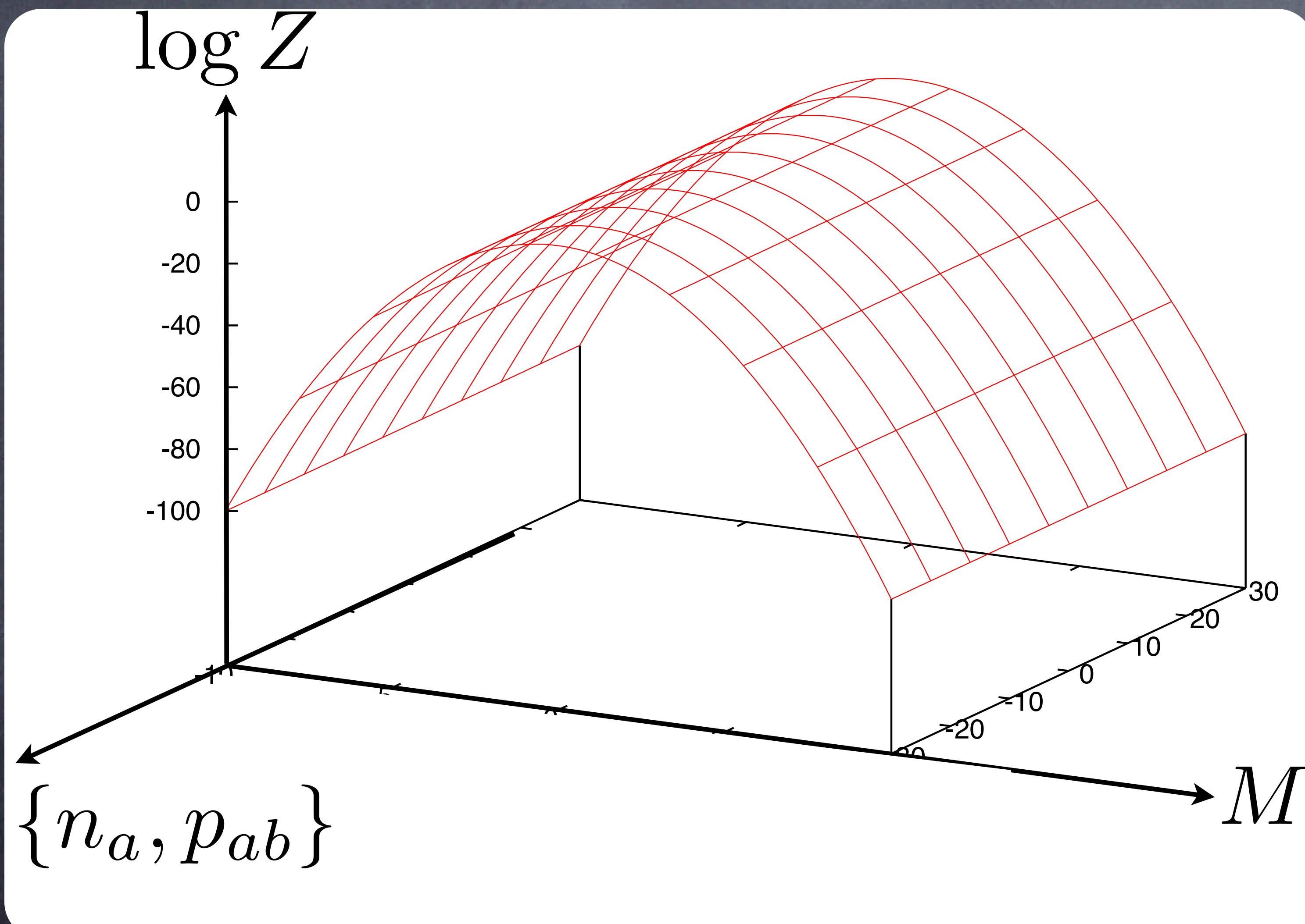
The original assignment can **not** be detected

(1) The paramagnetic case: Impossible inference



Log Z is flat in the "parameters" direction

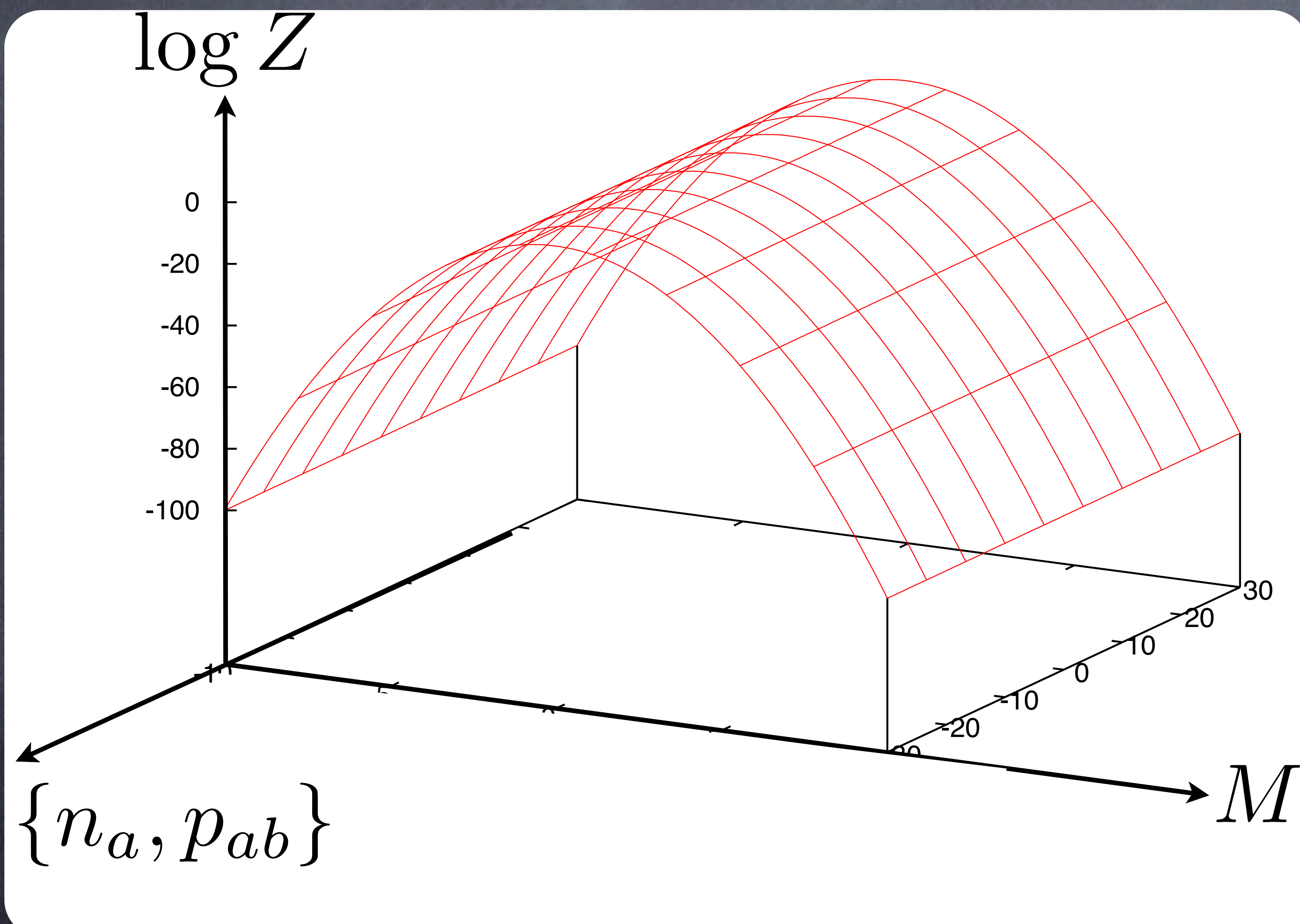
(1) The paramagnetic case: Impossible inference



Log Z is flat in the "parameters" direction

Inference of parameters is impossible

(1) The paramagnetic case: Impossible inference



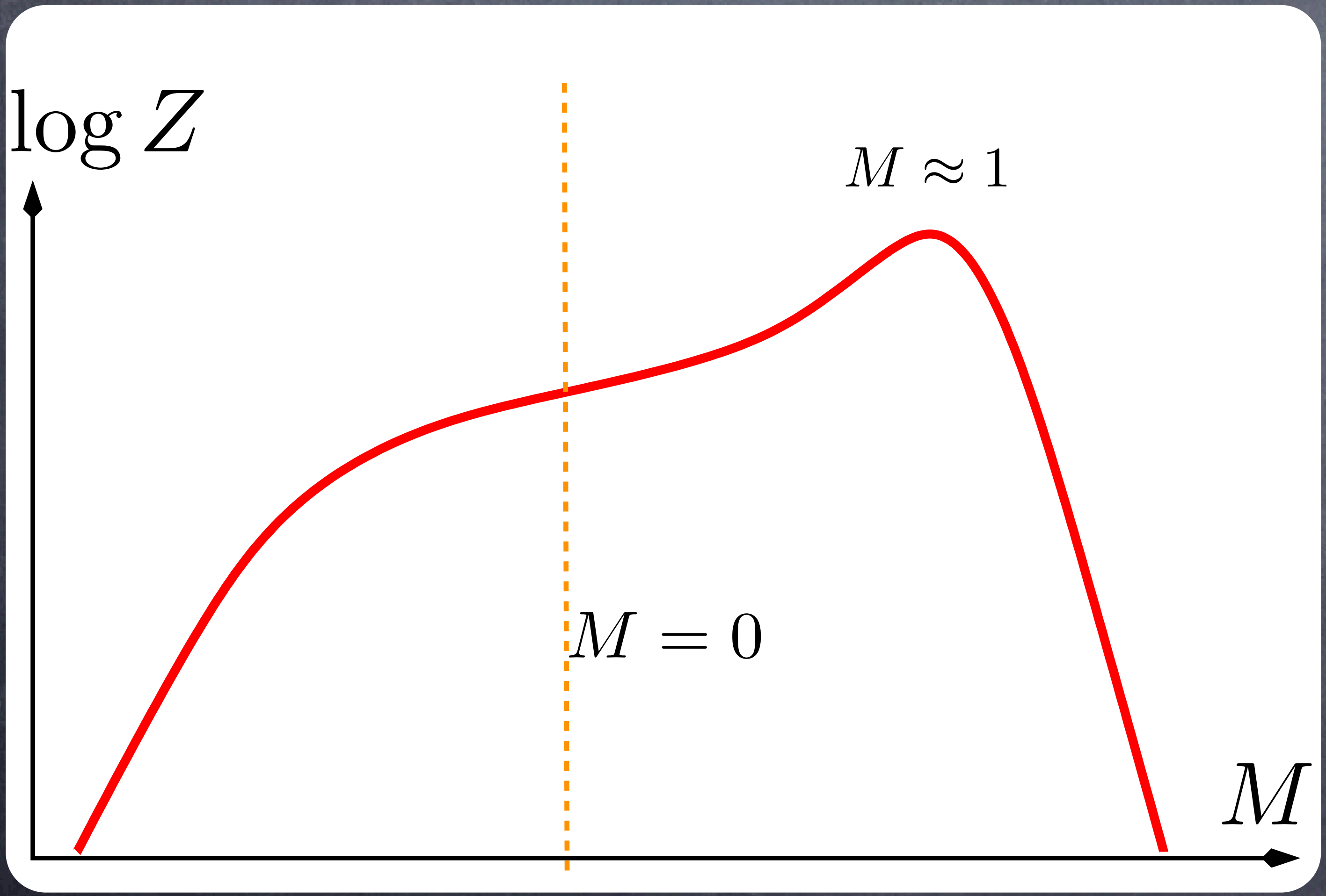
Log Z is flat in the "parameters" direction

Inference of parameters is impossible

In fact, what have been created is simply a random graph!

Can be proved by generalizing a theorem on quiet planting (Achlioptas, Coja-Oghlan'08).

(2) The ordered case: Easy inference



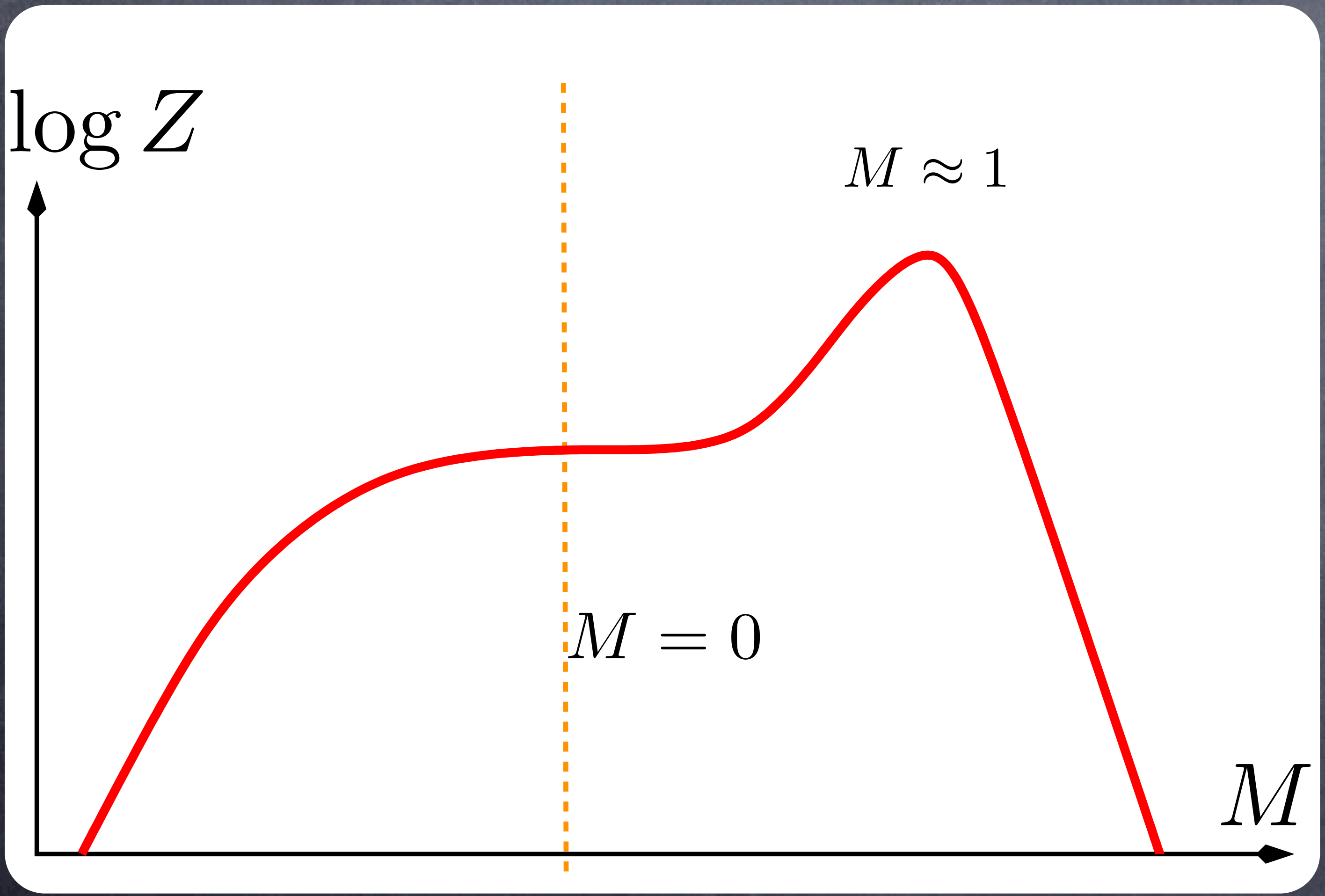
Assume we know the correct parameters $\{n_a, p_{ab}\}$

The maximum partition sum is now obtained for "ordered" non trivial marginals

The original assignment can now be detected

(2) The ordered case: Easy inference

Look for the critical case (spinodal point)



Assume we know the correct parameters $\{n_a, p_{ab}\}$

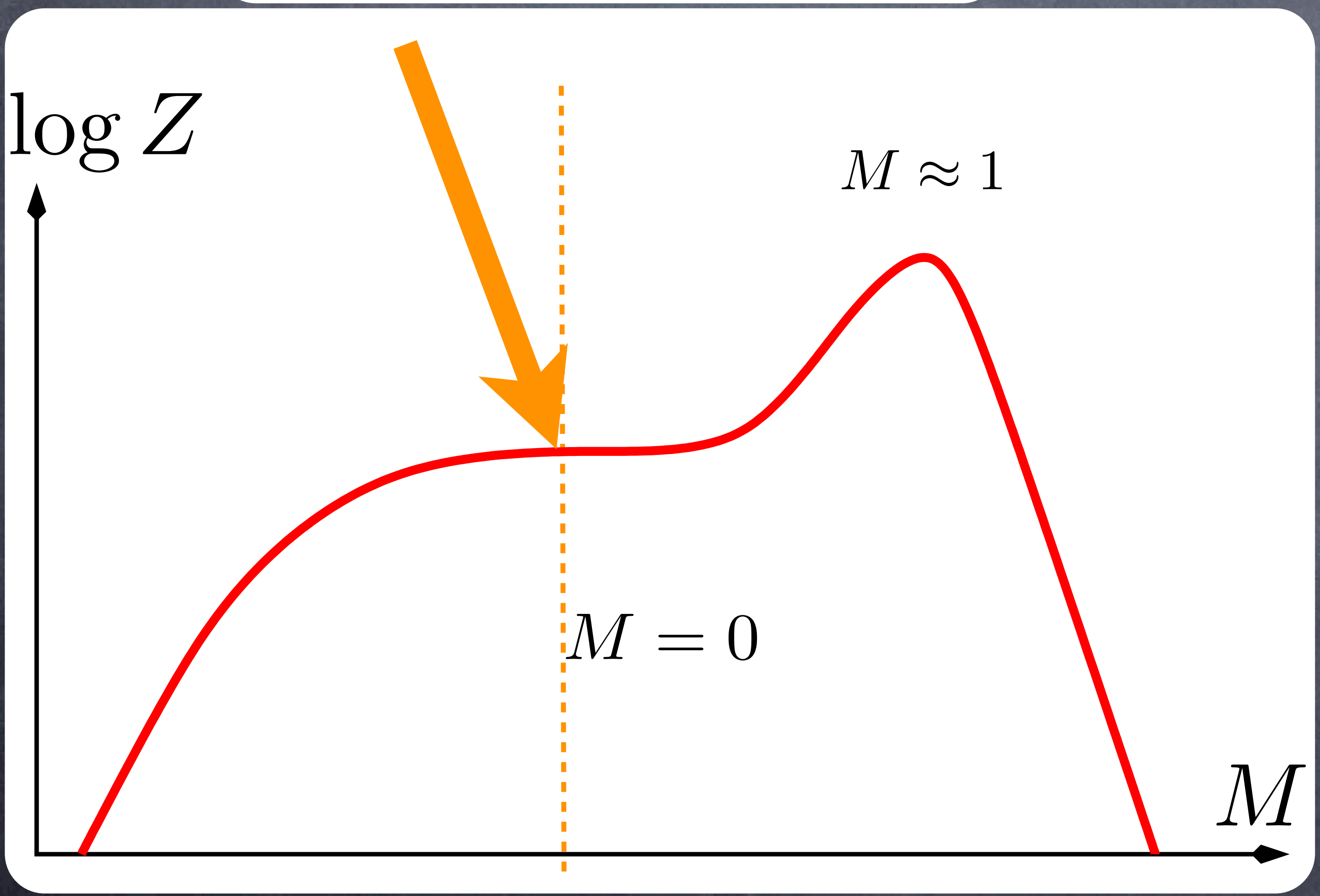
The maximum partition sum is now obtained for "ordered" non trivial marginals

The original assignment can now be detected

(2) The ordered case: Easy inference

Look for the critical case (spinodal point)

$$\left.\frac{d^2 \log Z(m)}{dm^2}\right|_{m=0} = 0$$



Assume we know the correct parameters $\{n_a, p_{ab}\}$

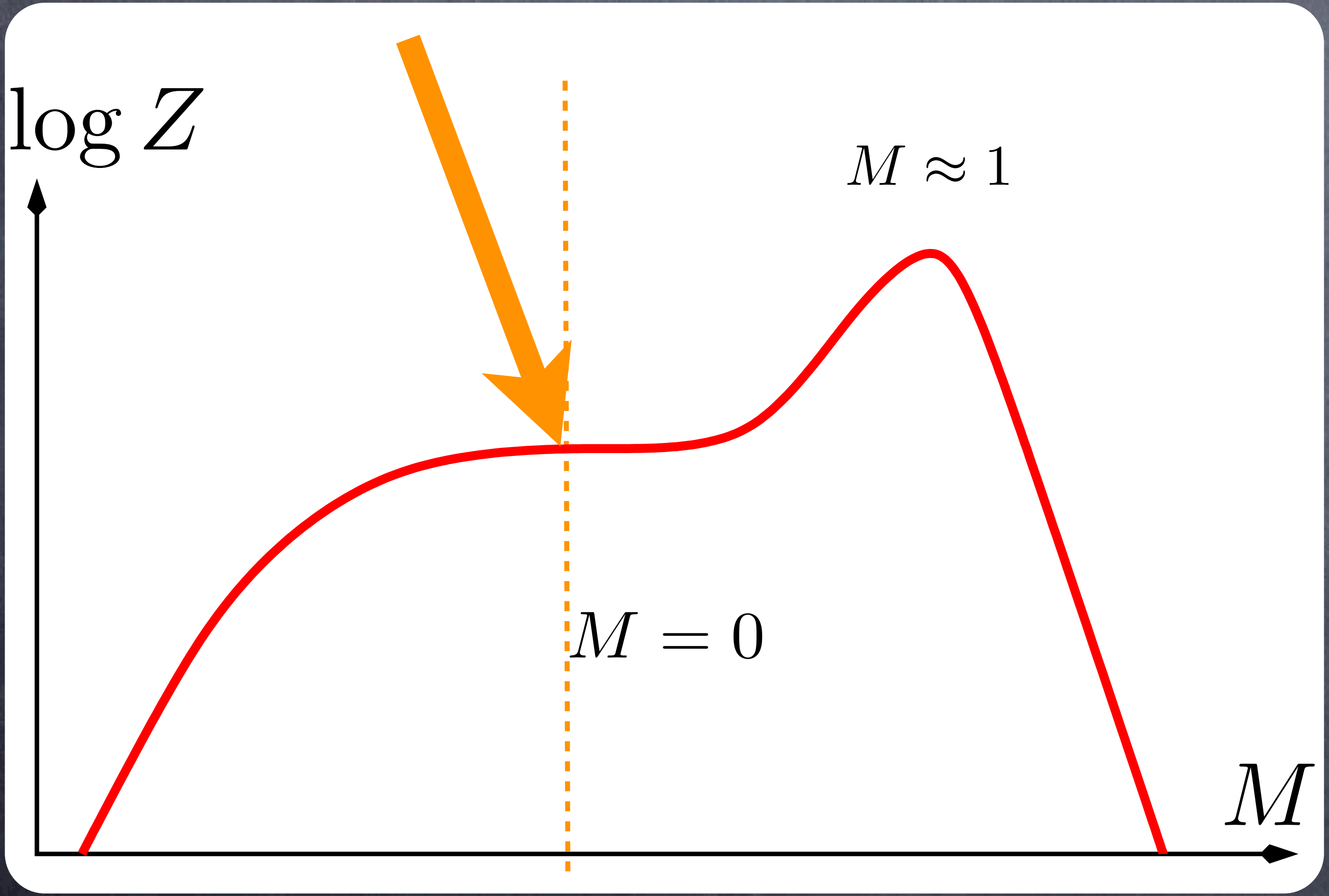
The maximum partition sum is now obtained for "ordered" non trivial marginals

The original assignment can now be detected

(2) The ordered case: Easy inference

Look for the critical case (spinodal point)

$$\left.\frac{d^2 \log Z(m)}{dm^2}\right|_{m=0} = 0$$



Assume we know the correct parameters $\{n_a, p_{ab}\}$

The maximum partition sum is now obtained for "ordered" non trivial marginals

The original assignment can now be detected

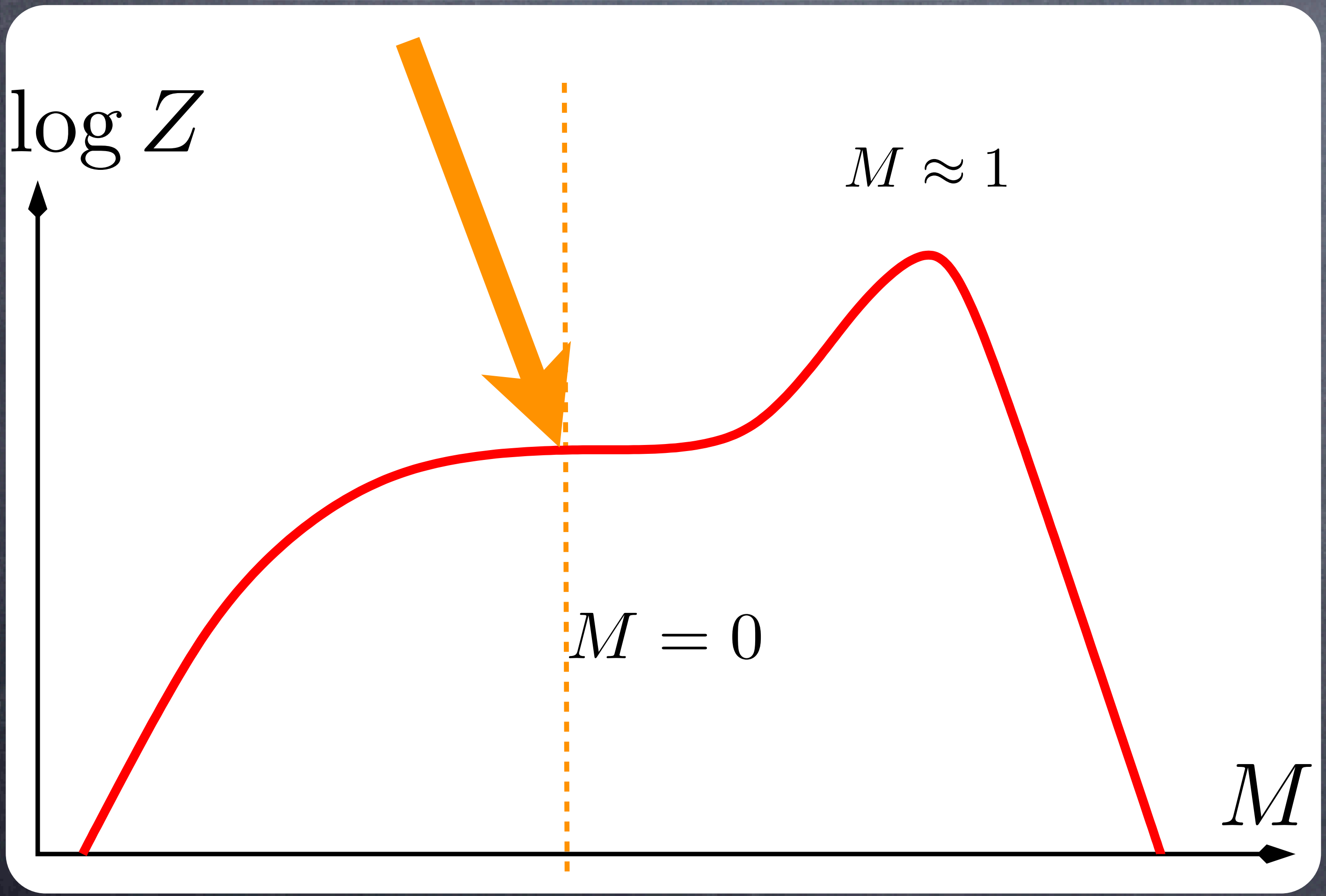
Physics: spinodal, or "de Almeida-Thouless" condition

Computer Science: "Kesten-Stigum" condition on census reconstruction

(2) The ordered case: Easy inference

Look for the critical case (spinodal point)

$$\left.\frac{d^2 \log Z(m)}{dm^2}\right|_{m=0} = 0 \quad \Rightarrow \quad |c_{\rm in} - c_{\rm out}| \geq q\sqrt{c}$$



Assume we know the correct parameters $\{n_a, p_{ab}\}$

The maximum partition sum is now obtained for "ordered" non trivial marginals

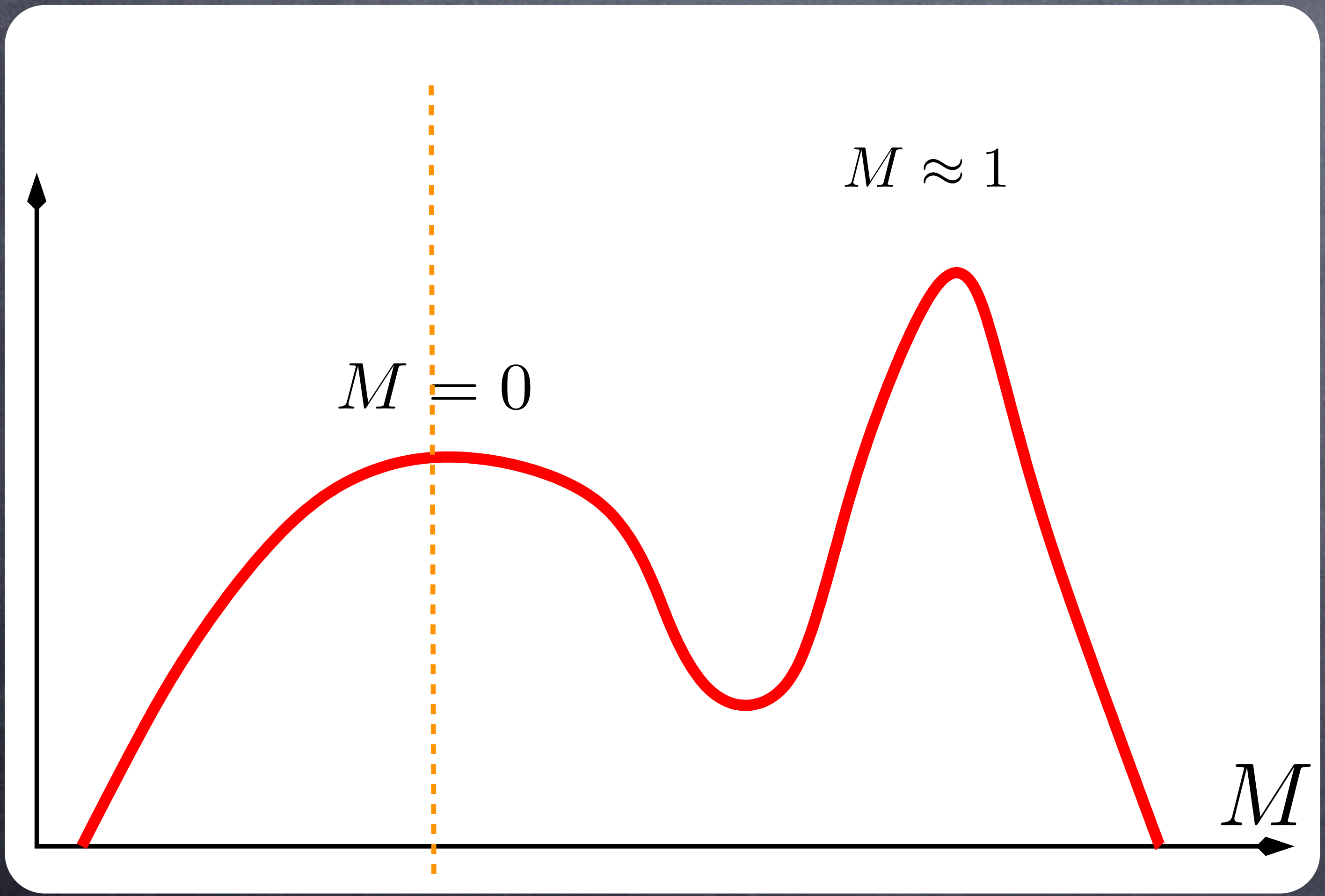
The original assignment can now be detected

Physics: spinodal, or "de Almeida-Thouless" condition

Computer Science: "Kesten-Stigum" condition on census reconstruction

(3) The "first-order" case: Hard inference

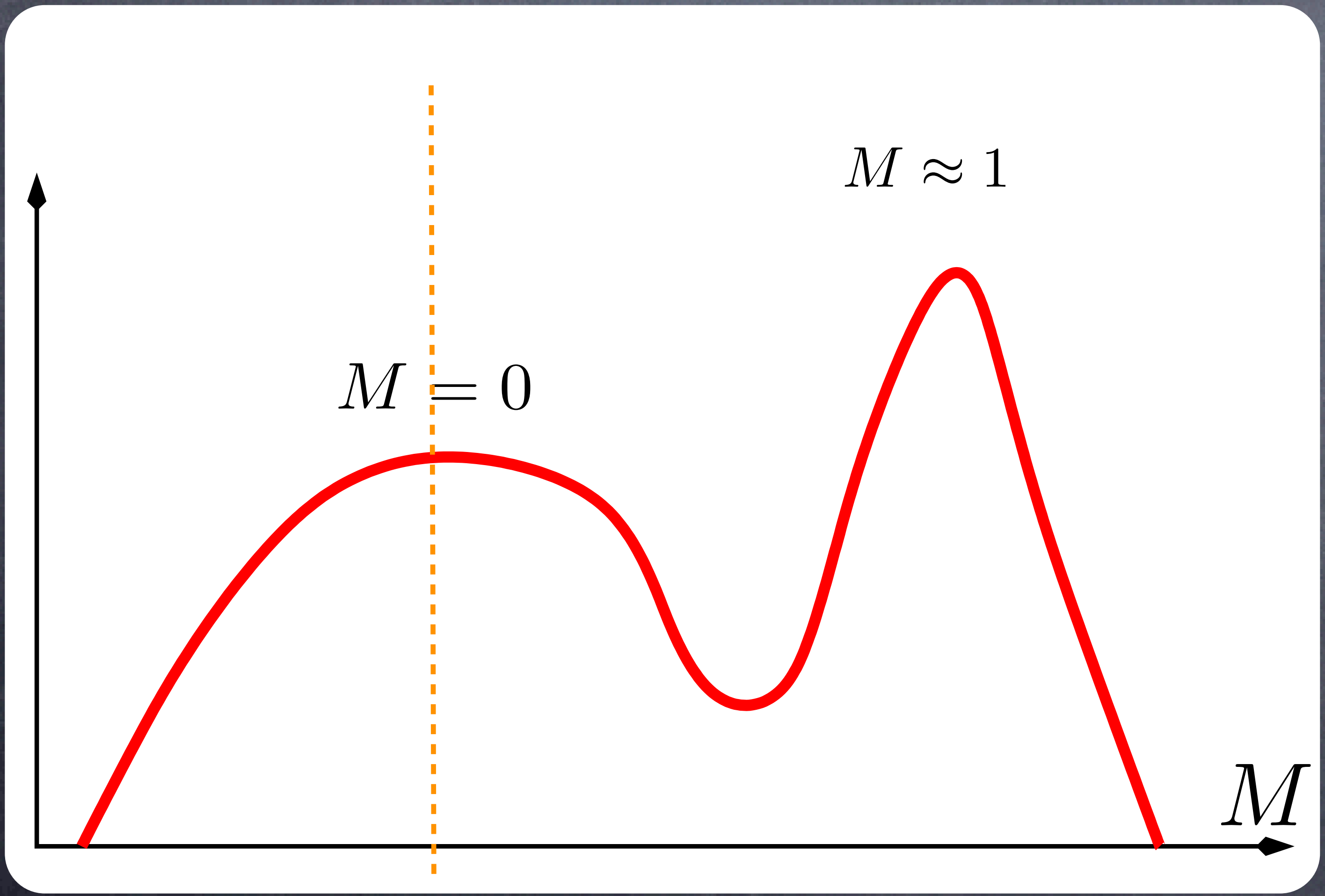
Assume we know the correct parameters $\{n_a, p_{ab}\}$



The maximum partition sum is obtained for "ordered" non trivial marginals...

... but finding this maximum is practically impossible!

(3) The "first-order" case: Hard inference



Assume we know the correct parameters $\{n_a, p_{ab}\}$

The maximum partition sum is obtained for "ordered" non trivial marginals...

... but finding this maximum is practically impossible!

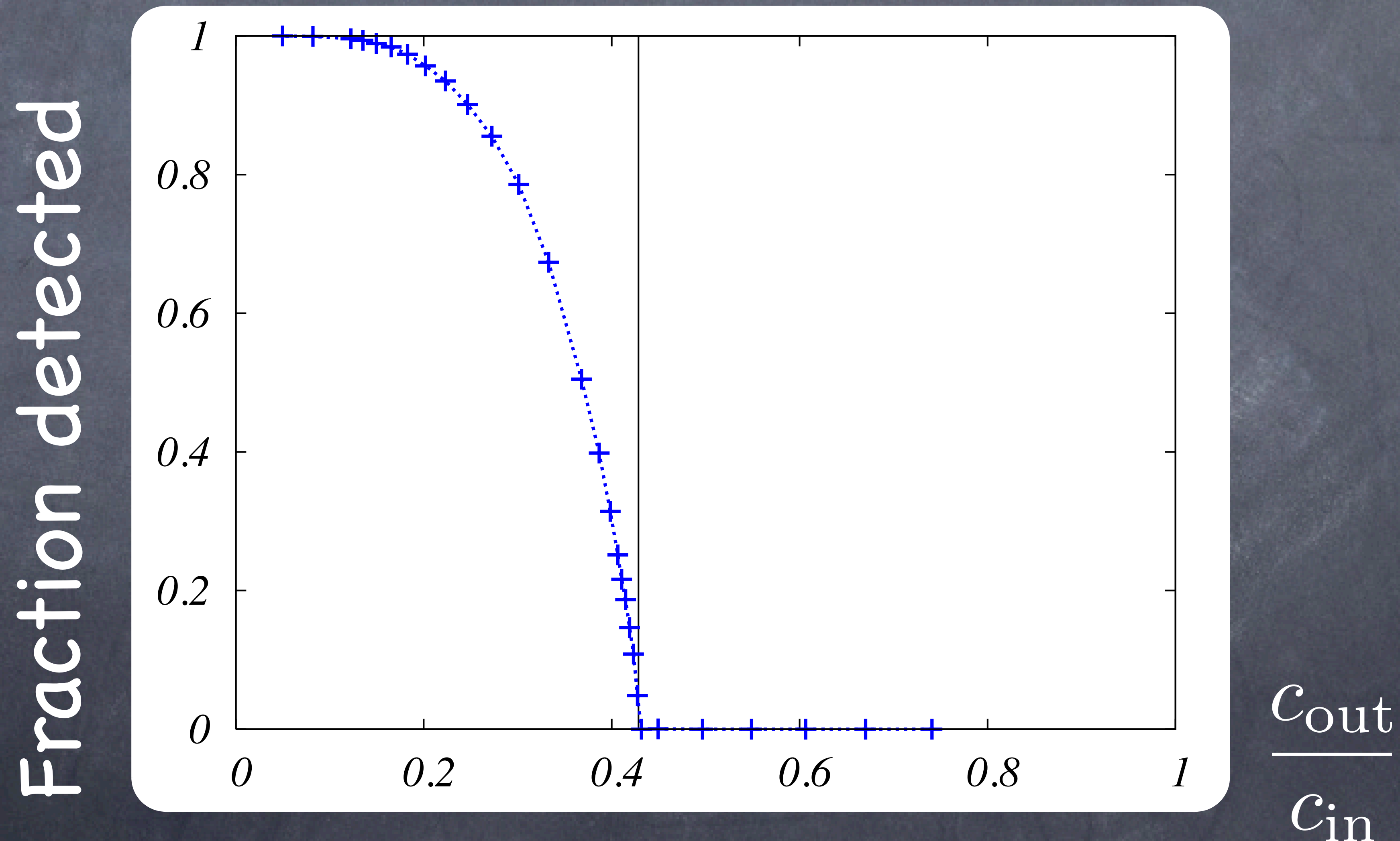
The original community **can** be detected

but one needs an **exponential** computational time

Example I: with assortative communities

$$q = 4,\ c = 16$$

$$n_a = \frac{1}{q},\ c_{aa} = c_{\text{in}},\ c_{a\neq b} = c_{\text{out}},\ cq = c_{\text{in}} + (q-1)c_{\text{out}}$$



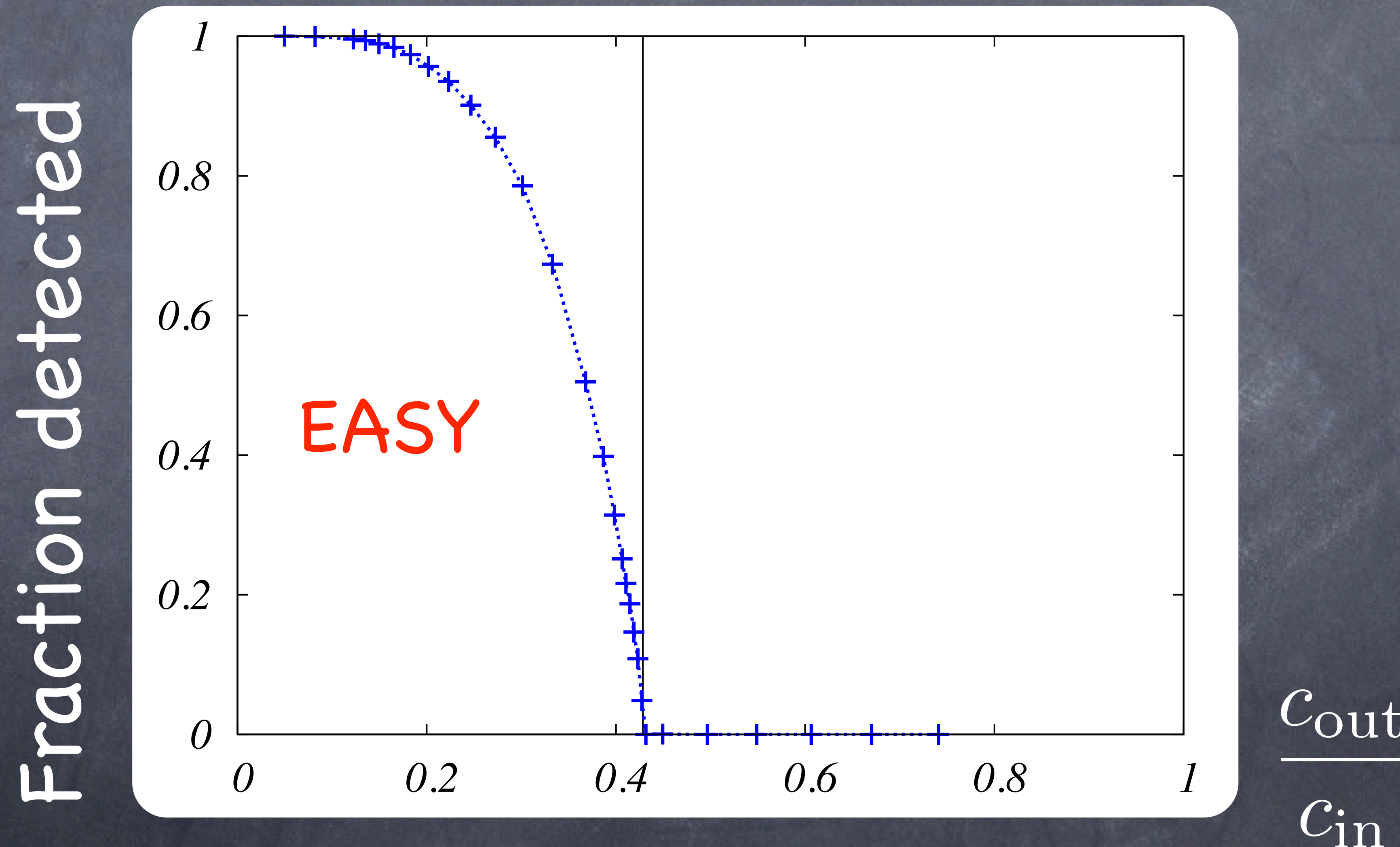
Planted Partitioning problem

Potts ferromagnet

Example I: with assortative communities

$$q = 4,\ c = 16$$

$$n_a = \frac{1}{q},\ c_{aa} = c_{\rm in},\ c_{a\neq b} = c_{\rm out},\ cq = c_{\rm in} + (q-1)c_{\rm out}$$

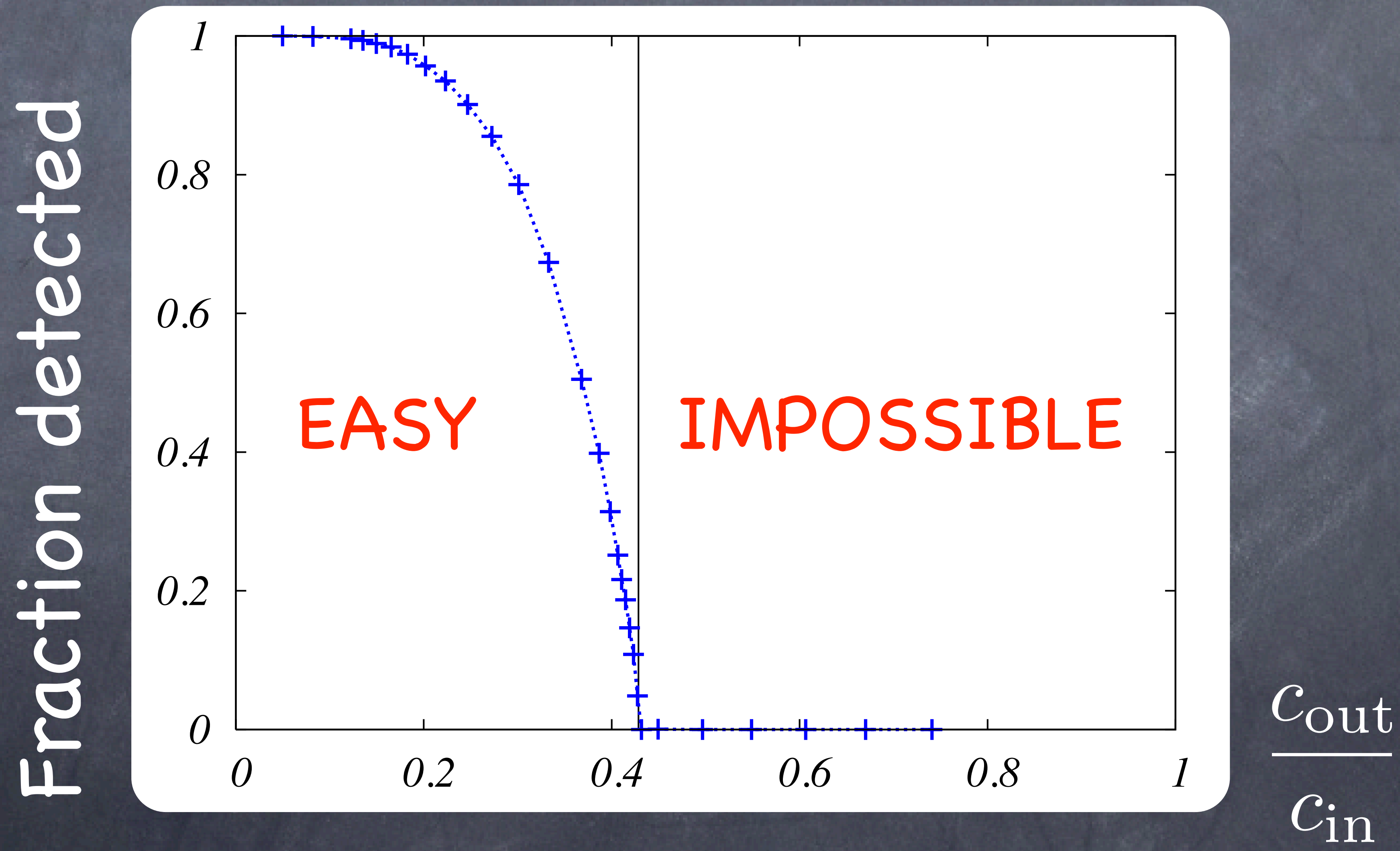


Planted Partitioning problem
Potts ferromagnet

Example I: with assortative communities

$$q = 4,\ c = 16$$

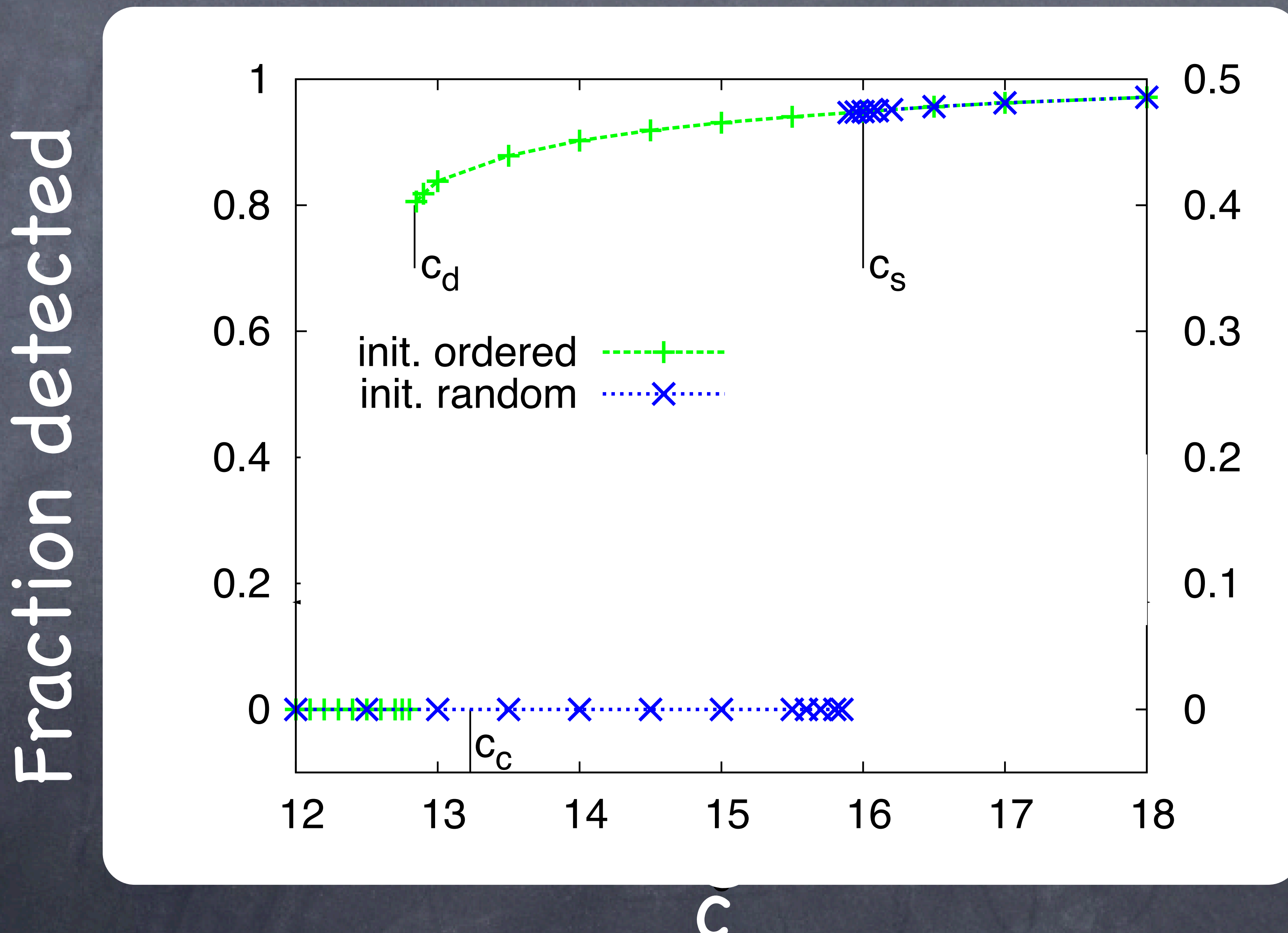
$$n_a = \frac{1}{q},\ c_{aa} = c_{\mathrm{in}},\ c_{a\neq b} = c_{\mathrm{out}},\ cq = c_{\mathrm{in}} + (q-1)c_{\mathrm{out}}$$



Planted Partitioning problem
Potts ferromagnet

Example II with "disassortative" communities

$$q = 5,\ n_a = \frac{1}{q},\ c_{aa} = 0,\ c_{a\neq b} = \frac{cq}{q-1},$$

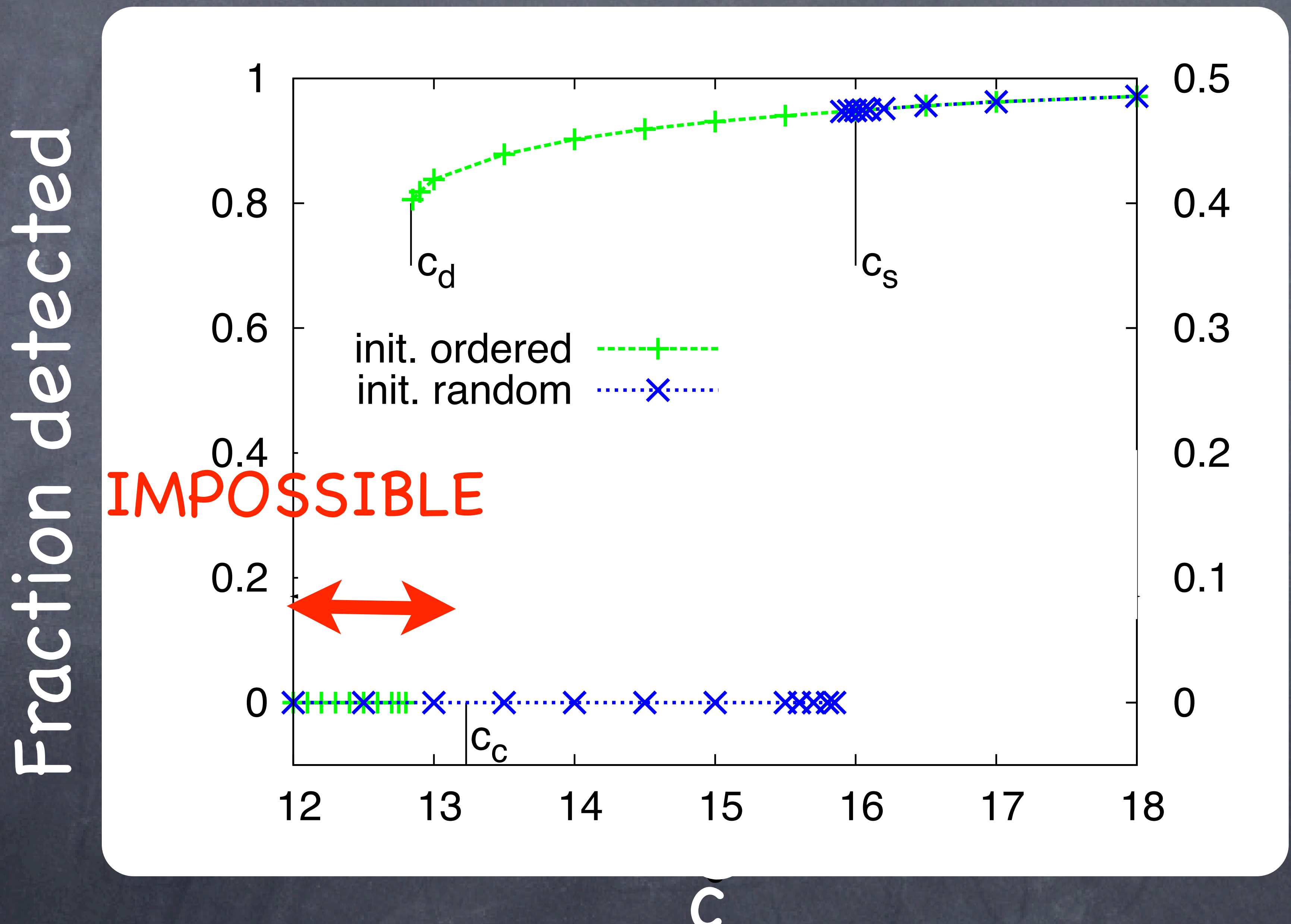


Planted Random graph coloring (Zdeborova, Krzakala'07)

Potts antiferromagnet

Example II with "disassortative" communities

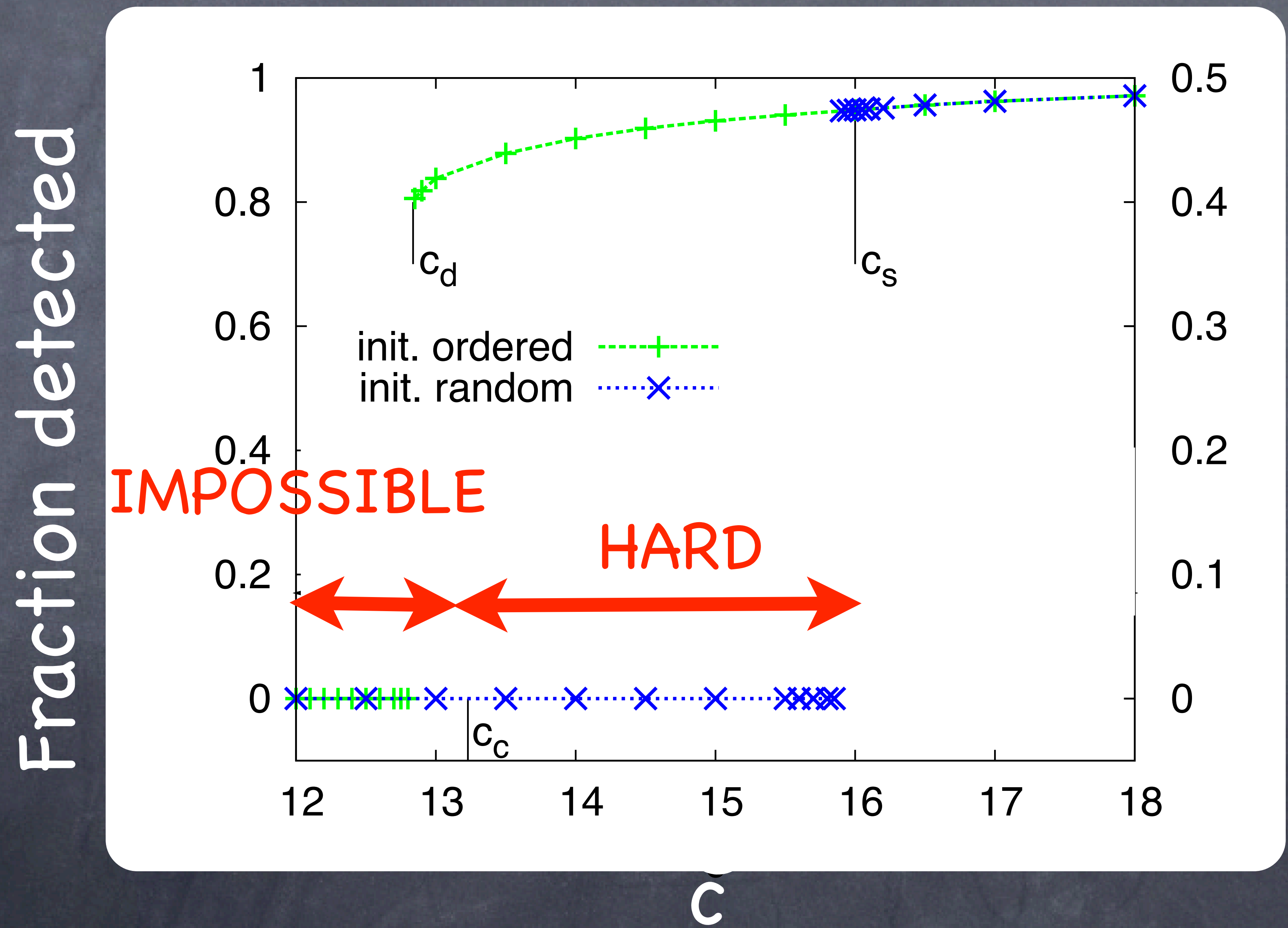
$$q = 5,\ n_a = \frac{1}{q},\ c_{aa} = 0,\ c_{a \neq b} = \frac{cq}{q-1},$$



Planted Random graph coloring (Zdeborova, Krzakala'07)
Potts antiferromagnet

Example II with "disassortative" communities

$$q=5,\ n_a=\frac{1}{q},\ c_{aa}=0,\ c_{a\neq b}=\frac{cq}{q-1},$$

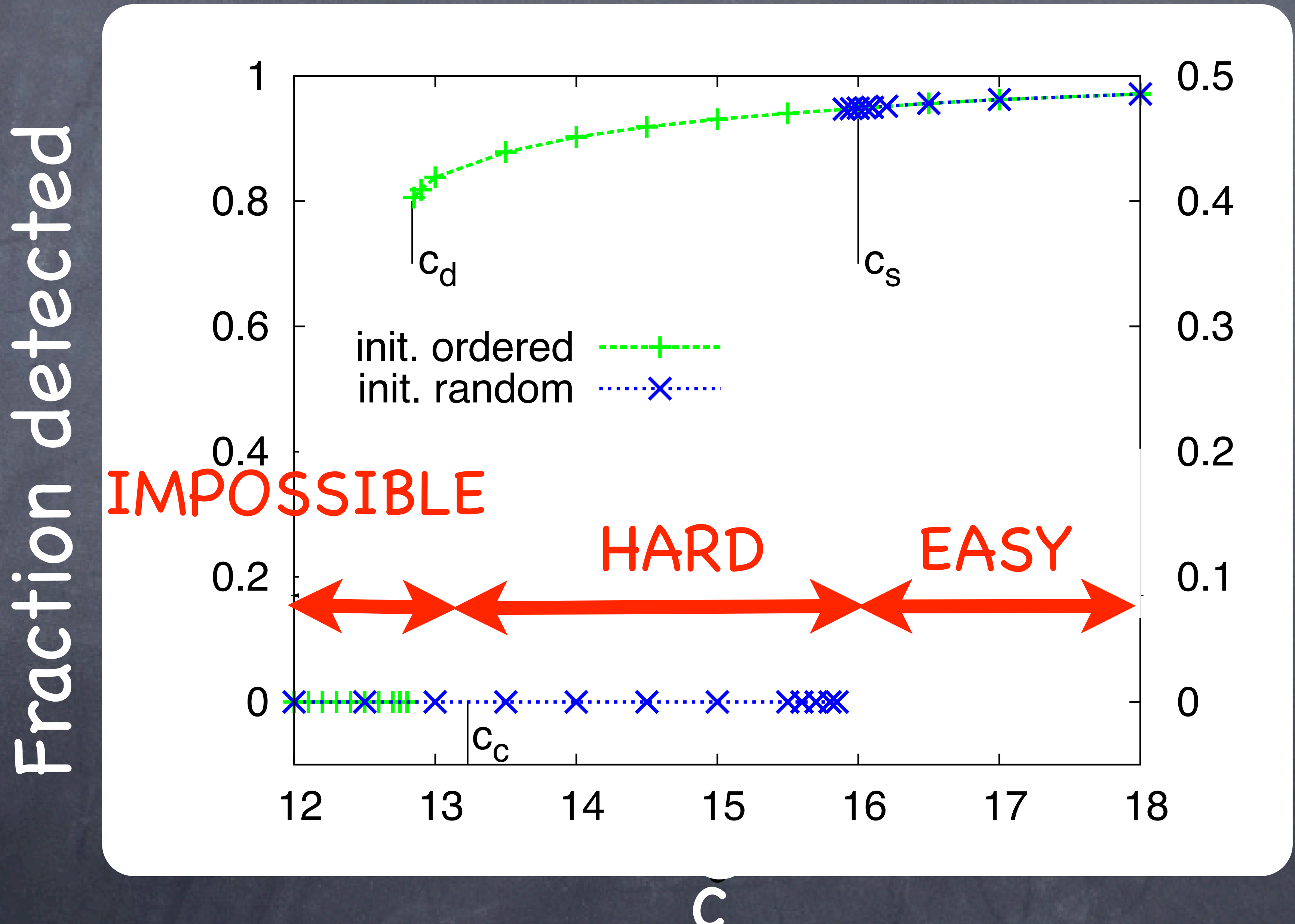


Planted Random graph coloring (Zdeborova, Krzakala'07)

Potts antiferromagnet

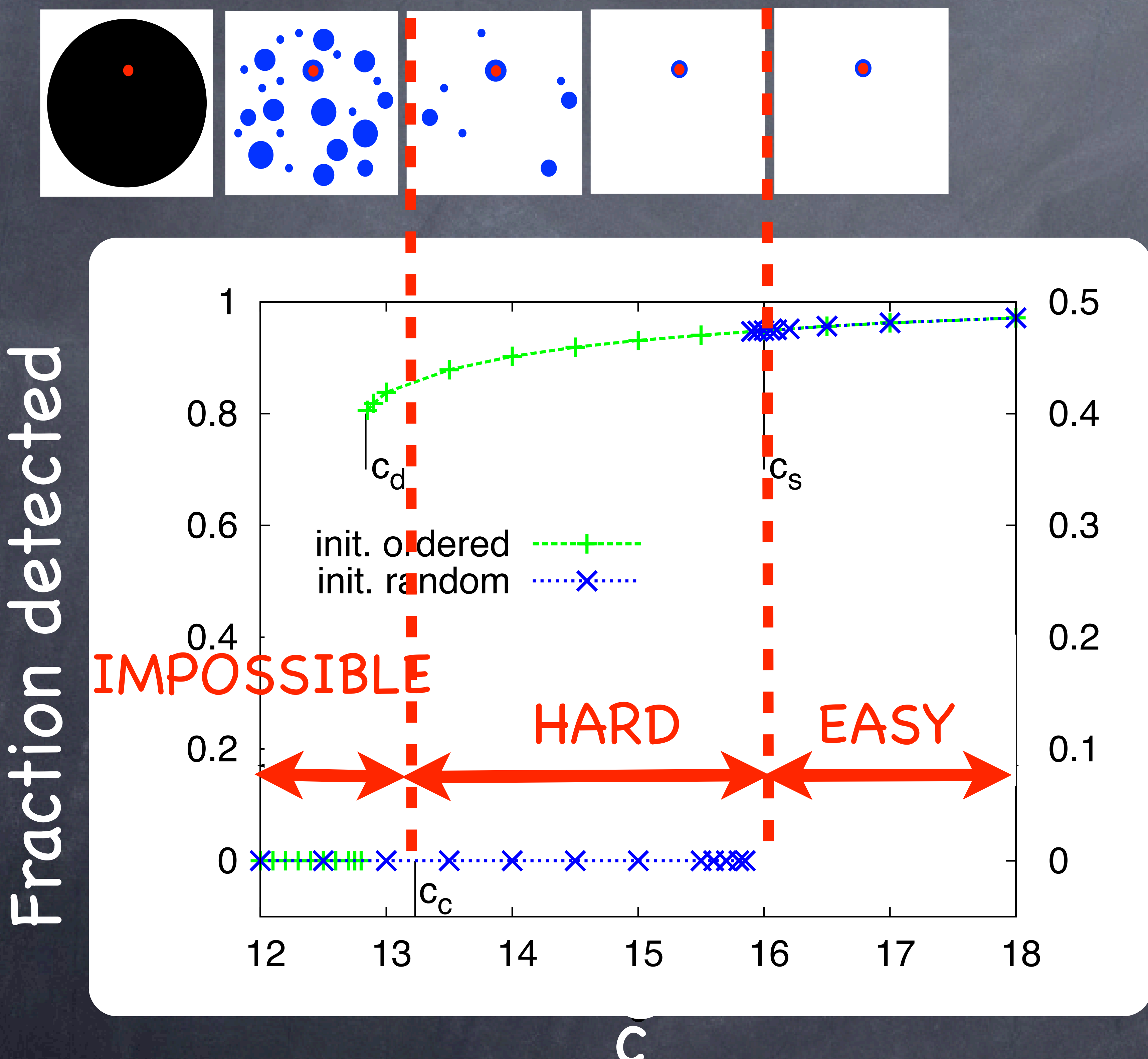
Example II with "disassortative" communities

$$q = 5,\ n_a = \frac{1}{q},\ c_{aa} = 0,\ c_{a \neq b} = \frac{cq}{q-1},$$



Planted Random graph coloring (Zdeborova, Krzakala'07)
Potts antiferromagnet

The Relation with Potts Spin Glasses



Impossible⇨Possible

=

Kauzmann transition

Hard⇨Easy

=

Almeida-Thouless

Planted Random graph coloring (Zdeborova, Krzakala'07)

Potts antiferromagnet

Inference in community detection

- Phase transitions from easy, hard and impossible inference
- BP allows for a fast and exact solution and is an optimal algorithm for the block model...
- ...and can be generalized to **any** local generative model.
- BP is also a very efficient tool for **real-world networks** (cf. Aurelien Decelle's Talk) and for **directed and weighted graphs.**

arxiv:1102.1182

Bonus

How to learn the number of groups?

Degree corrected block model

- Our block model generates Poisson degree distribution - it does not want to believe that nodes with very different degrees may be in the same group.
- Degree corrected (Karrer, Newman'10)

$$p_{q_i,q_j} = d_i d_j \omega_{q_i,q_j}$$