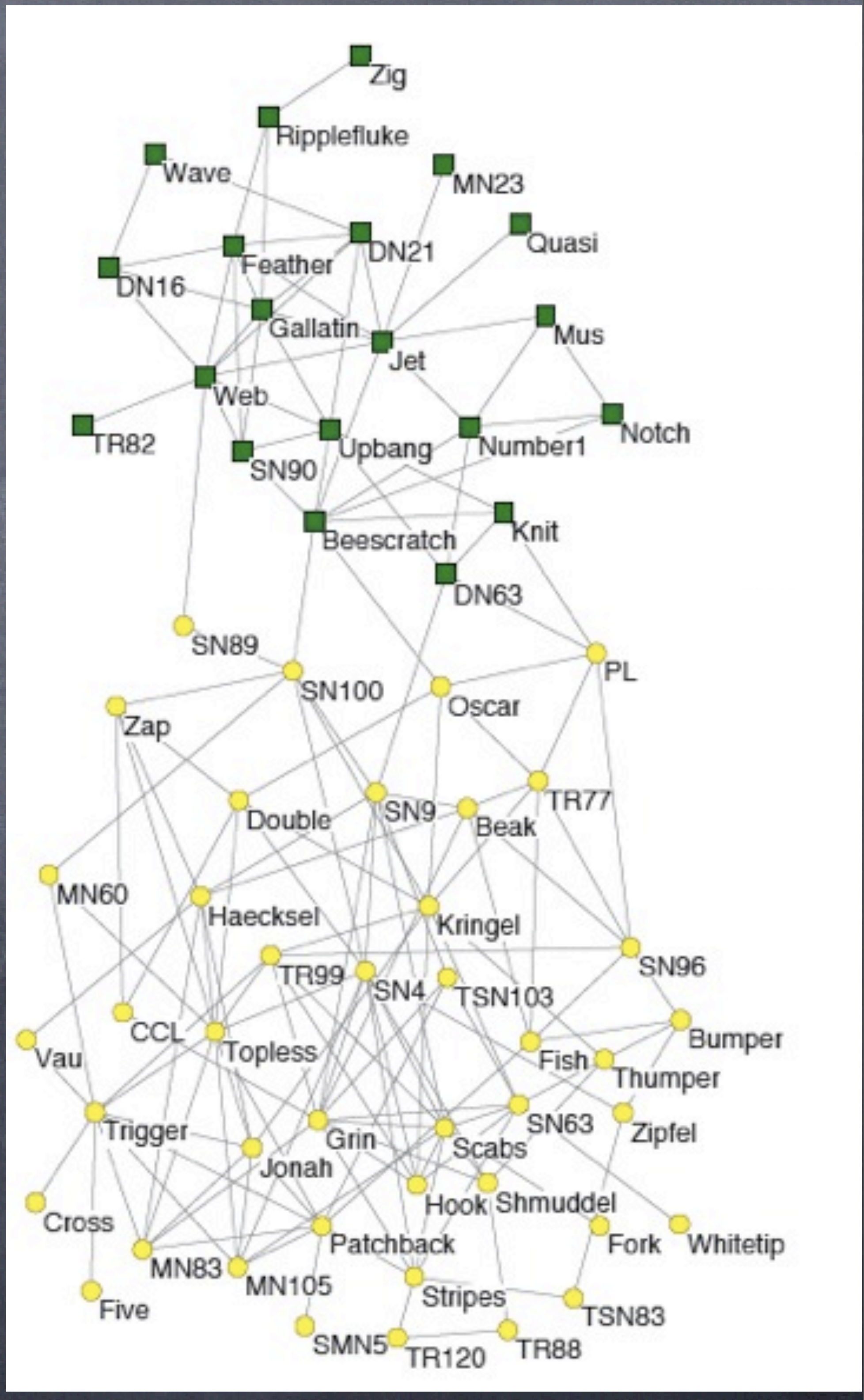
Detection of functional modules from network topology

Lenka Zdeborová
Florent Krzakala (ESPCI)
Cris Moore (Santa Fe Inst.)
Aurelien Decelle (LPTMS)

Popular example

- 62 bottlenose dolphins living in Doubtful Sound in New Zealand observed by Lusseau (2003).
- Edge if the pair is seen together more often than expected by chance.
- The group separated in two groups after one dolphin left the place.
- The structure of the two groups can be predicted from the network topology (Arenas, Fernandez, Gomez'2008)

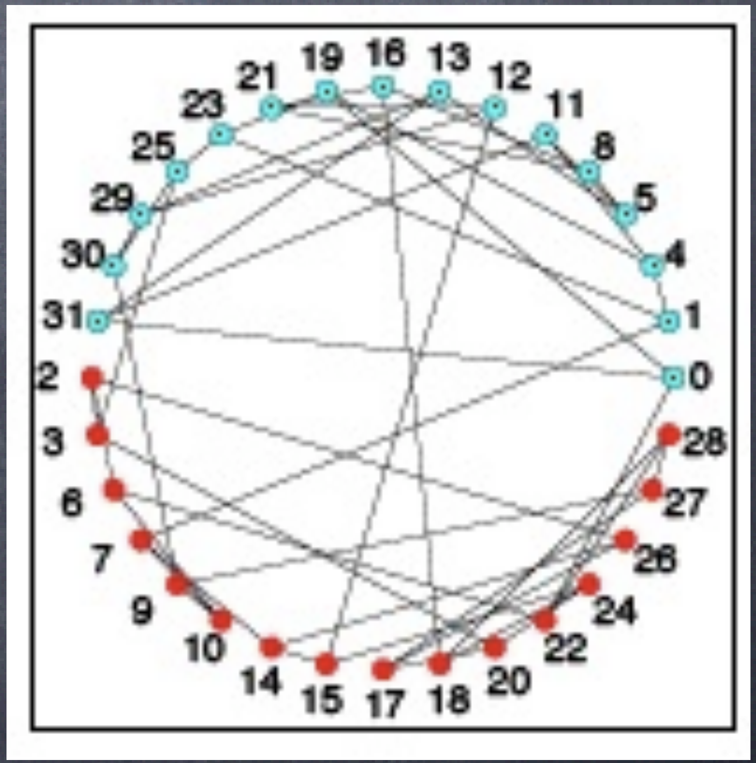


State of art

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

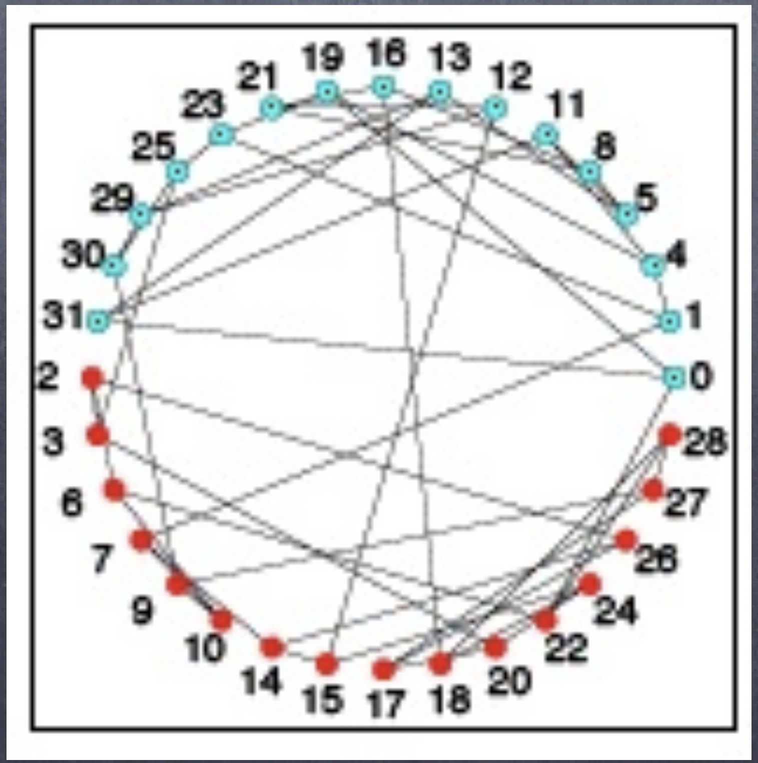
$$Q = \frac{1}{2M} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2M} \right) \delta_{q_i, q_j}$$

- Current methods are unable to tell that a random graph does not have any communities. E.g.: Ising model on random graphs of degree 3, in the best bisection only about 11.4% of edges between the two groups.



State of art

- Missing measures of significance, estimates of probability of error ...
- Focus on communities: Nodes of the same kind tend to be more inter-connected. Not useful in many cases, e.g. food-web, adjacency of words in text.



Need for more fundamental and formal approach!

Need for more fundamental and formal approach!

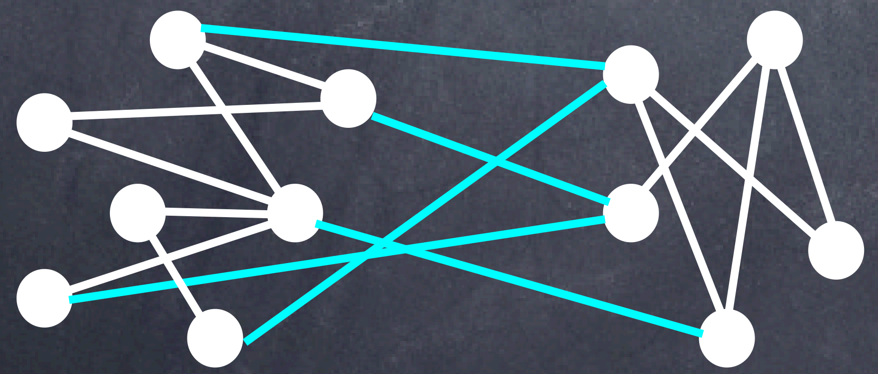
spin glass :-)

Block model

Generate a random network as follows:

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

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



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

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

Block model

Generate a random network as follows:

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

Question 1: Given $q, \{n_a\}, \{p_{ab}\}$ what is the best possible guess for the original group assignment?

Question 2: Given only the graph, what is the best guess for $q, \{n_a\}, \{p_{ab}\}$

Question 2: Given only the graph, what is the best guess for $\{n_a, p_{ab}\}$

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

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

Question 2: Given only the graph, what is the best guess for $\{n_a, p_{ab}\}$

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

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

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

Question 2: Given only the graph, what is the best guess for $\{n_a, p_{ab}\}$

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

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

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

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

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

Equilibrium statistical physics of

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

Equilibrium statistical physics of

$$-H(\{q_i\}) = \sum_{i=1}^{N} \log n_{q_i} + \sum_{ij} \left[A_{ij} \log p_{q_i q_j} + (1 - A_{ij}) \log (1 - p_{q_i q_j})\right]$$

$$= \sum_{i=1}^{N} \log n_{q_i} + \sum_{(ij)\in E} \log \frac{p_{q_i q_j}}{1 - p_{q_i q_j}} + \sum_{a,b=1}^{q} N_a N_b \log (1 - p_{ab})$$

Partition function maximized if and only if:

quenched energy = annealed energy

"Nishimori condition"

$$\frac{1}{N} \left\langle \sum_i \delta_{a,q_i} \right\rangle = n_a$$

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

Learning of parameters

(1) Compute the averages:

- ➡ With Monte Carlo (detailed balance)
- ➡ With belief propagation (= Bethe-Peierls = TAP equations = cavity method) faster

$$\psi_{q_i}^{i\to j} = \frac{1}{Z^{i\to j}} n_{q_i} e^{-h_{q_i}} \prod_{k\in\partial i\setminus j} \left[\sum_{q_k} c_{q_k q_i} \psi_{q_k}^{k\to i} \right]$$

$$h_{q_i} = \frac{1}{N} \sum_k \sum_{q_k} c_{q_k q_i} \psi_{q_k}^{k} \qquad p_{ab} = \frac{c_{ab}}{N}$$

Learning of parameters

(1) Compute the averages:

- ➡ With Monte Carlo (detailed balance)
- ➡ With belief propagation (= Bethe-Peierls = TAP equations = cavity method) faster

(2) Update parameters as

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

(3) Repeat till convergence.

Question 1: Given $\{n_a\}, \{p_{ab}\}$ what is the best possible guess for the original group assignment?

Question 1: Given $\{n_a\}, \{p_{ab}\}$ what is the best possible guess for the original group assignment?

Bayes optimal inference (in error correcting codes by Nishimori'93, Sourlas'94):

(1) Compute marginals (local magnetizations)

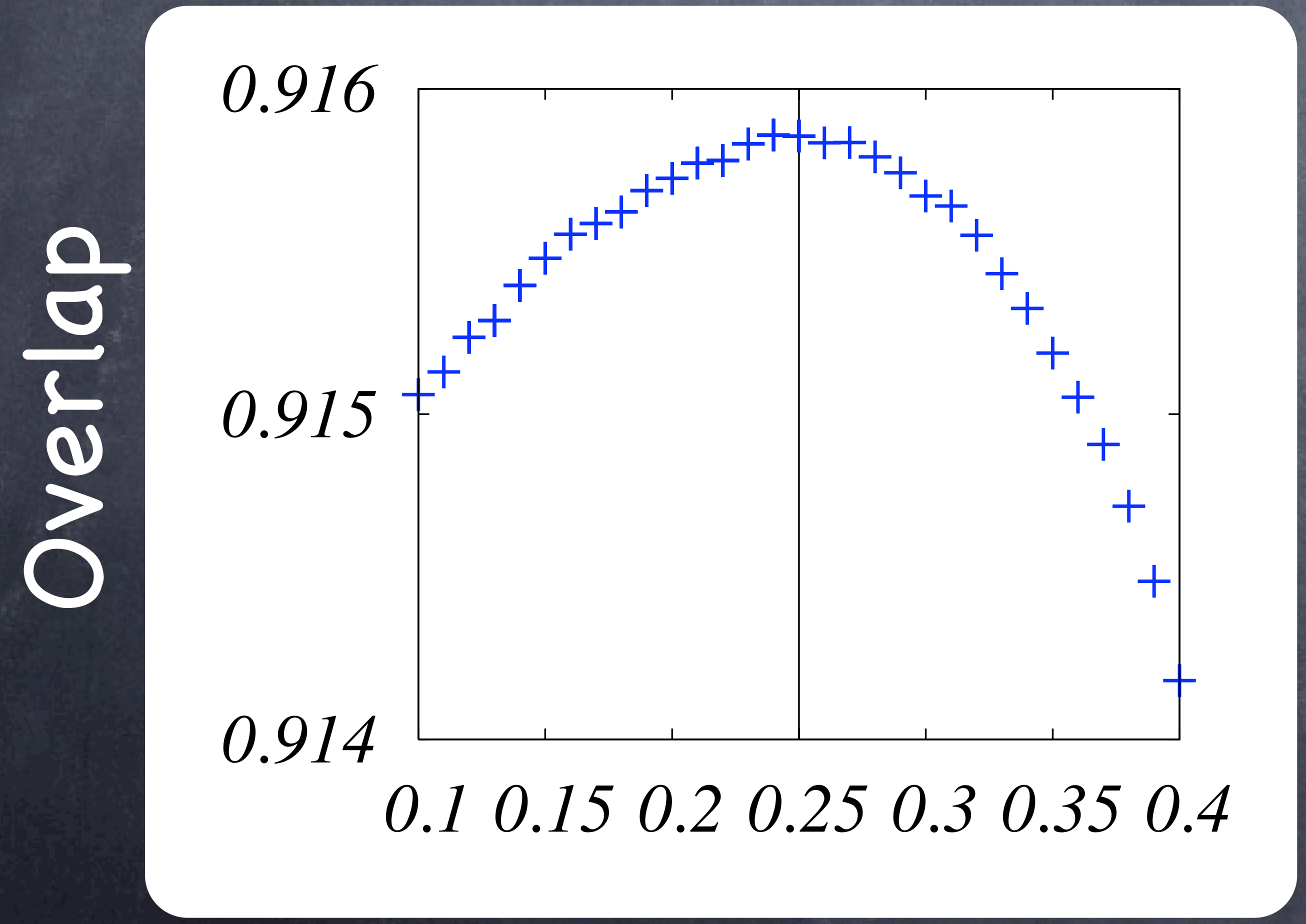
(2) For each node take the most probable value.

Question 1: Given $\{n_a\}, \{p_{ab}\}$ what is the best possible guess for the original group assignment?

Bayes optimal inference (in error correcting codes by Nishimori'93, Sourlas'94):

(1) Compute marginals (local magnetizations)

(2) For each node take the most probable value.



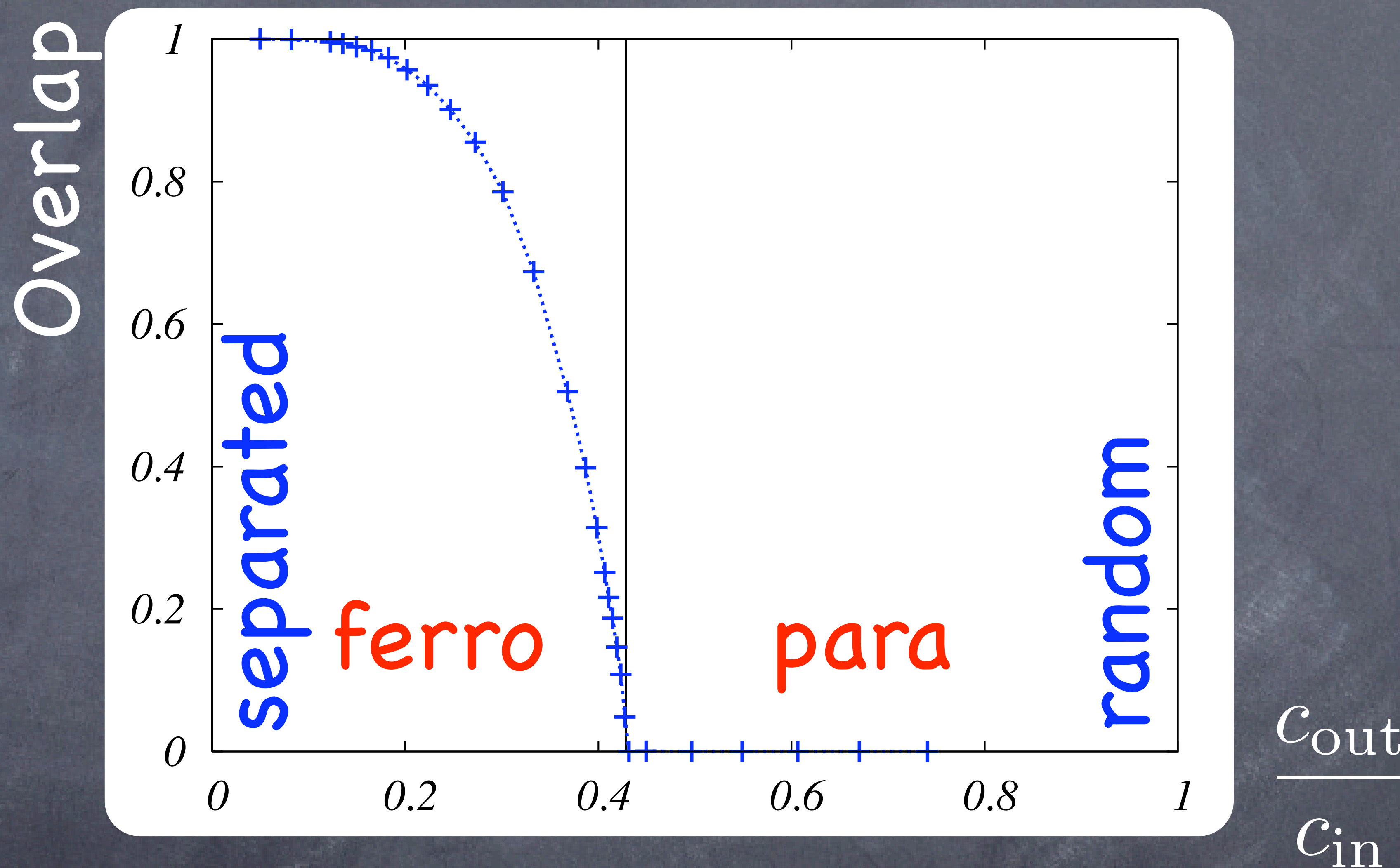
This overlap is maximized at the right value of $\{n_a\}, \{p_{ab}\}$

Example I

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

$q = 4,\ c = 16$

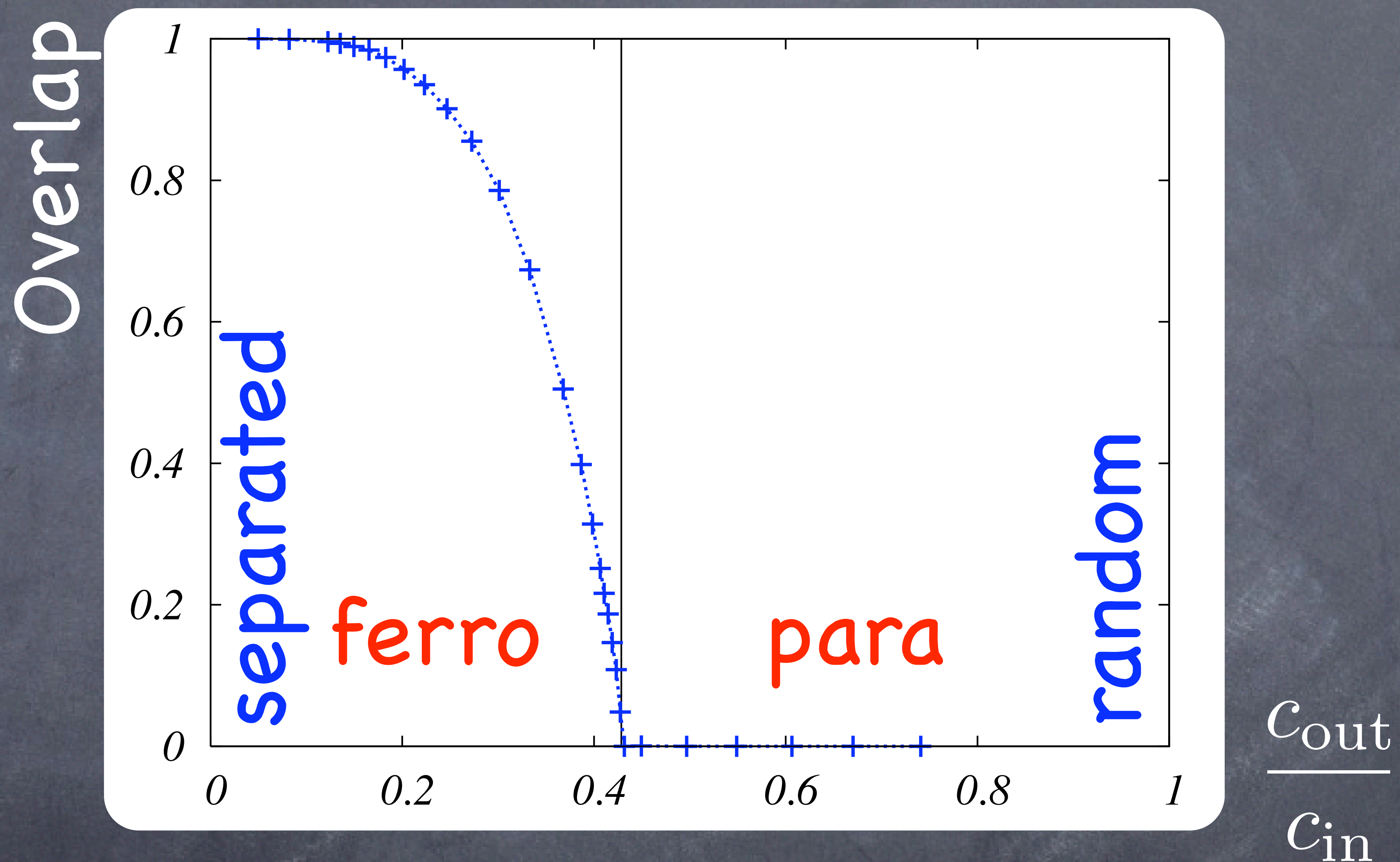
ferromagnet = communities



Example I

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

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



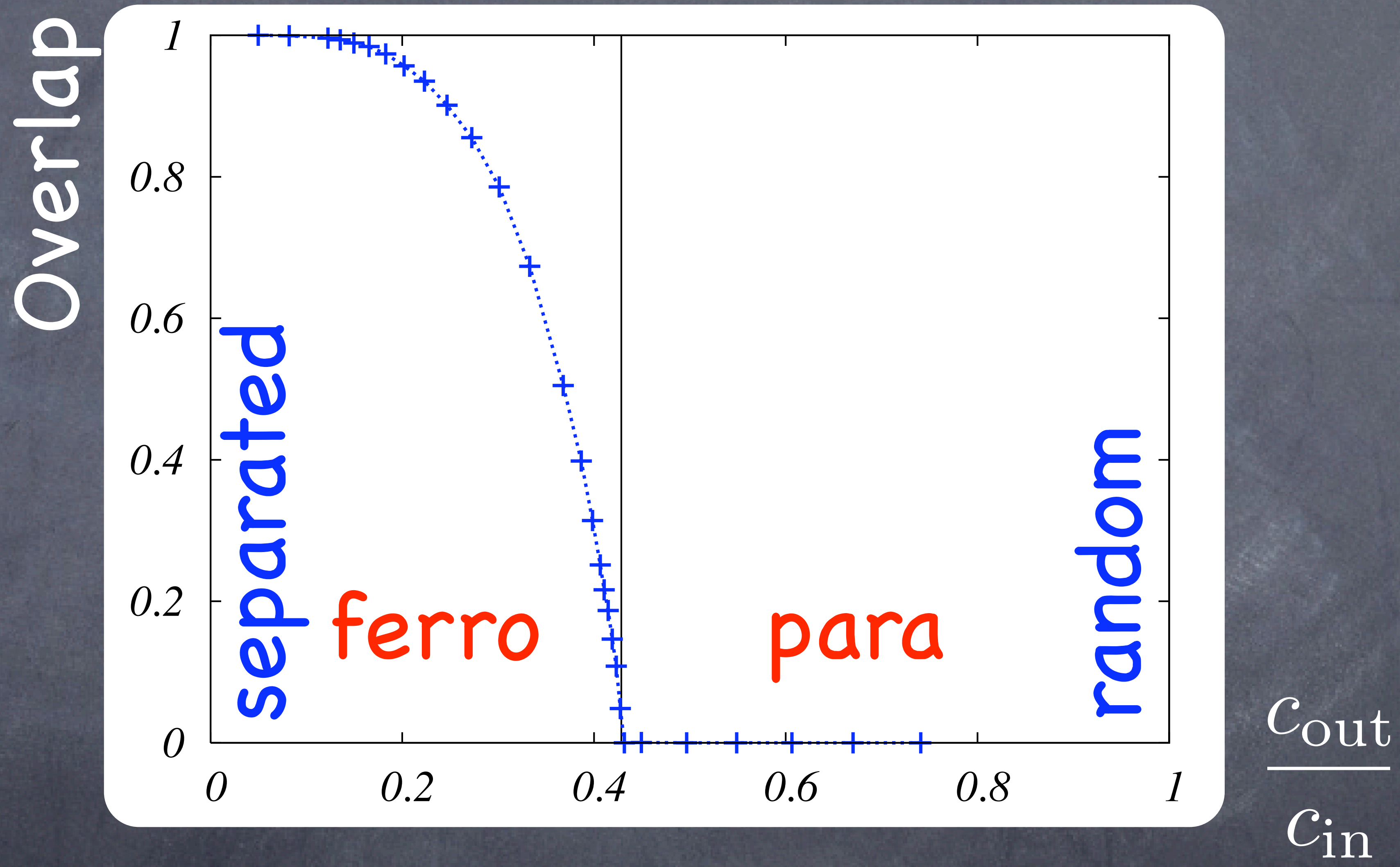
Paramagnetic phase: Random graph was created (Achlioptas, Coja-Oghlan'08). Zero overlap between an equilibrium configuration and the original one. Learning impossible.

Example I

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

$q = 4,\ c = 16$

ferromagnet = communities

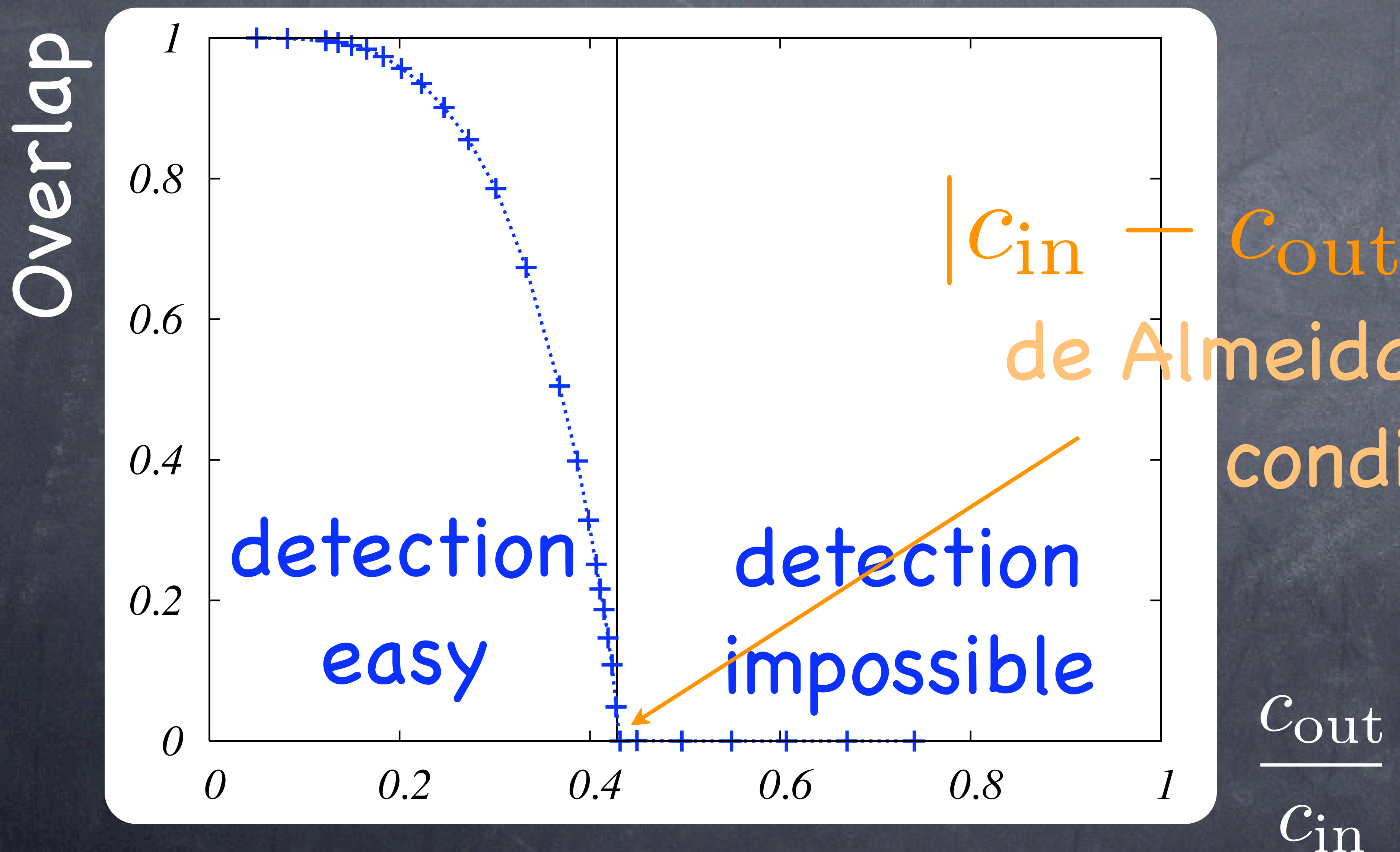


Ferromagnetic phase: Network contains information about modules, equilibration easy (Nishimori line no RSB, no glass).

Example I

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

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



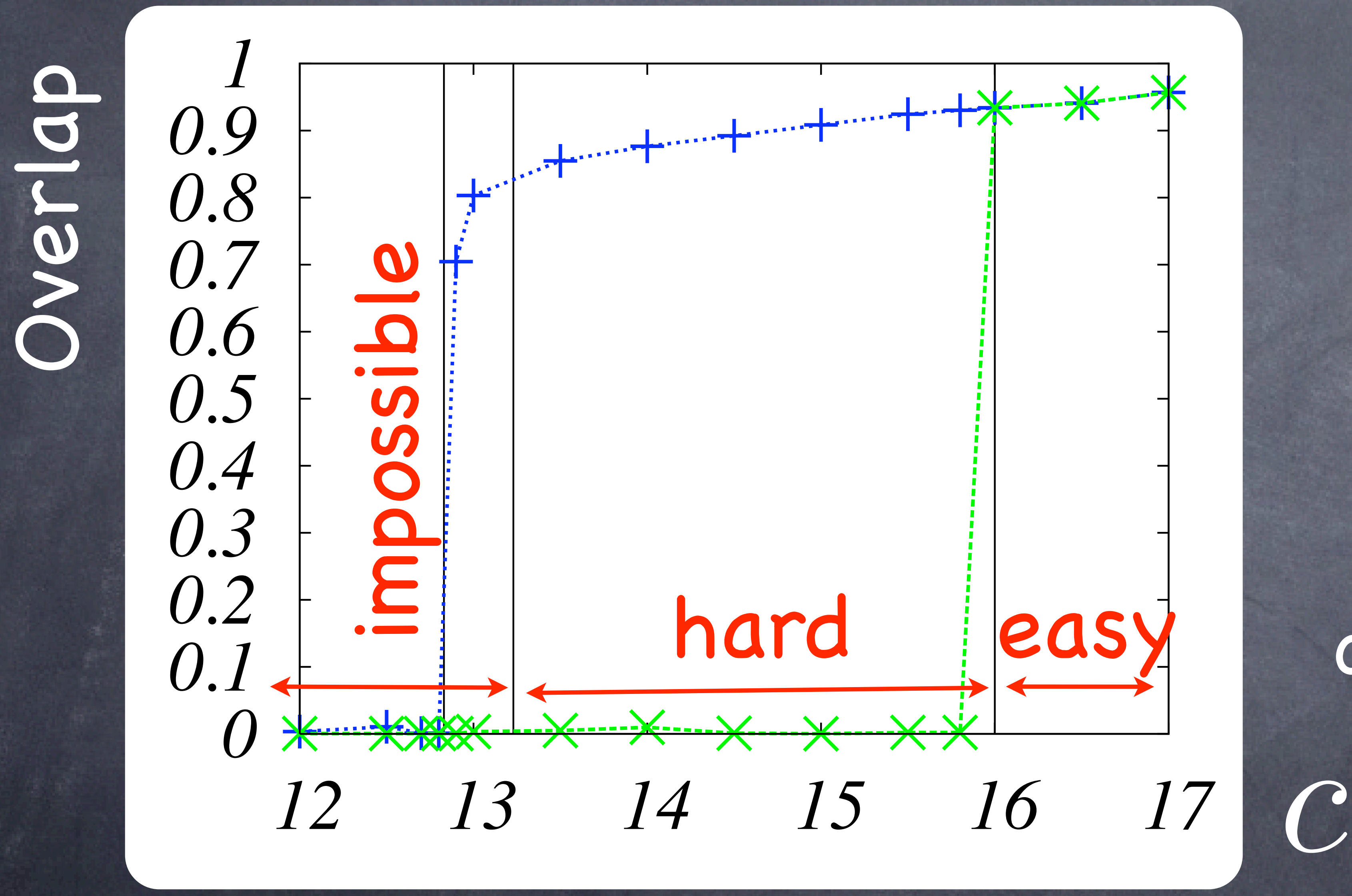
$$|c_{\text{in}} - c_{\text{out}}| \geq q\sqrt{c}$$

de Almeida-Thouless condition

Example II

anti-ferromagnet = coloring

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



Impossible - random graph created, paramagnetic phase

$c_{\mathrm{K}} = 13.23$

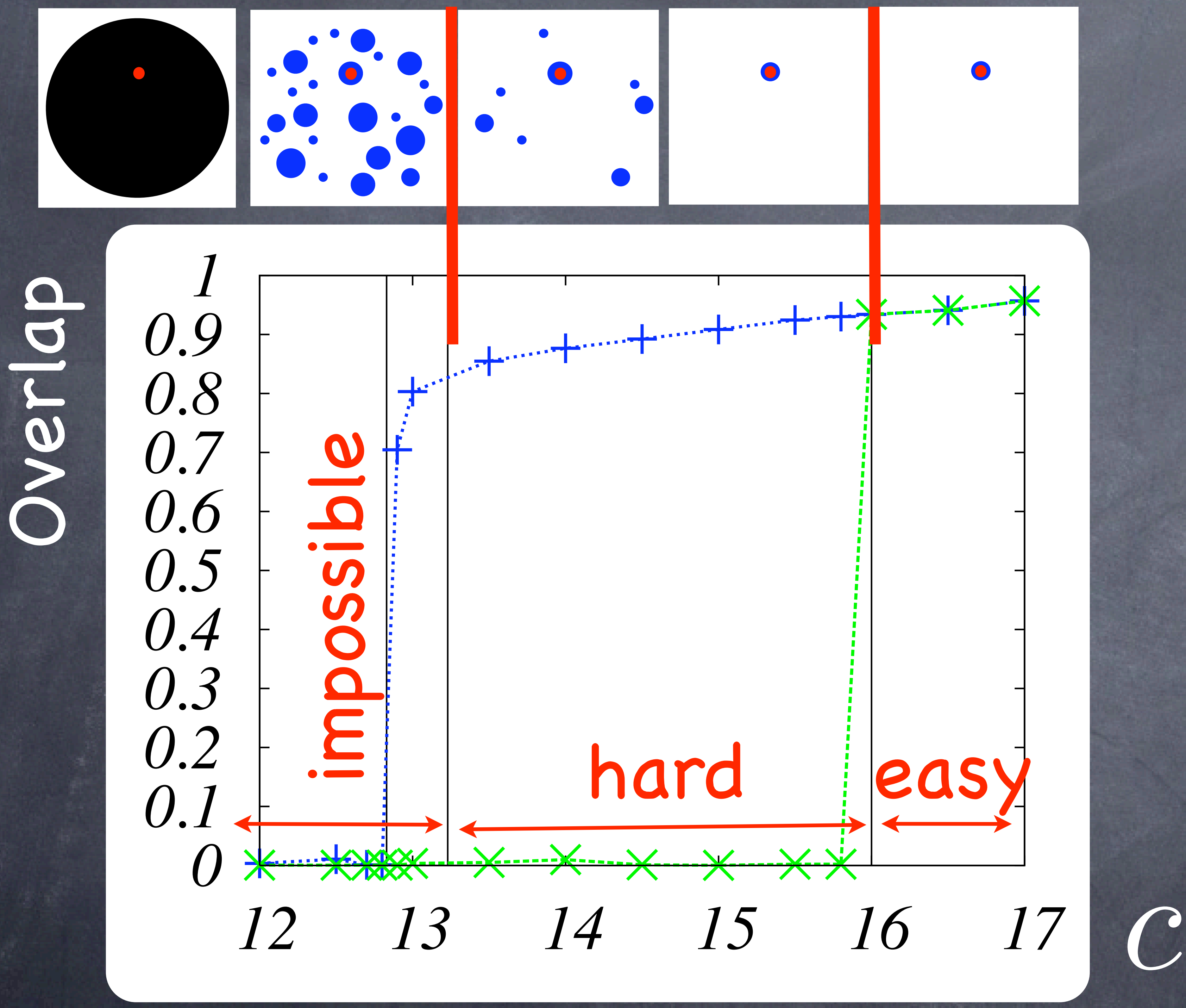
Easy - planted configuration attractive beyond AT condition

$c_{\mathrm{AT}} = 16$

Values of phase transitions the same as in random graph coloring (Zdeborova, Krzakala'07)

Example II

anti-ferromagnet = coloring



Hard - equilibrium solution correlated with the planted configuration, but hidden in an 1RSB phase

Values of phase transitions the same as in random graph coloring (Zdeborova, Krzakala'07)

Question 2: Given only the graph, what is the best guess for $q, \{n_a\}, \{p_{ab}\}$

Iterate with BP

$$\frac{1}{N}\left\langle \sum_i \delta_{a,q_i} \right\rangle = n_a$$

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

Output:

I) Asymptotically exact values of parameters $\{n_a\}, \{p_{ab}\}$

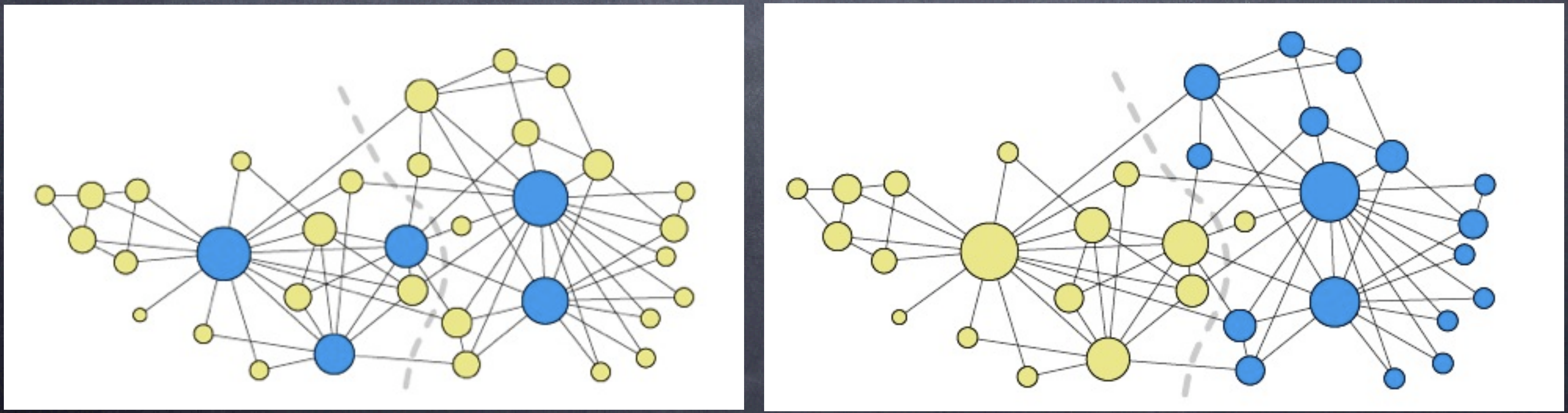
II) Learning impossible or hard.

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)

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



Conclusion

- Using basic properties of (planted) Potts and spin glass models gives us fundamental approach and new algorithms for module detection in networks.
- In progress: Using the degree corrected model for analysis of real networks.

arXiv:1012.????