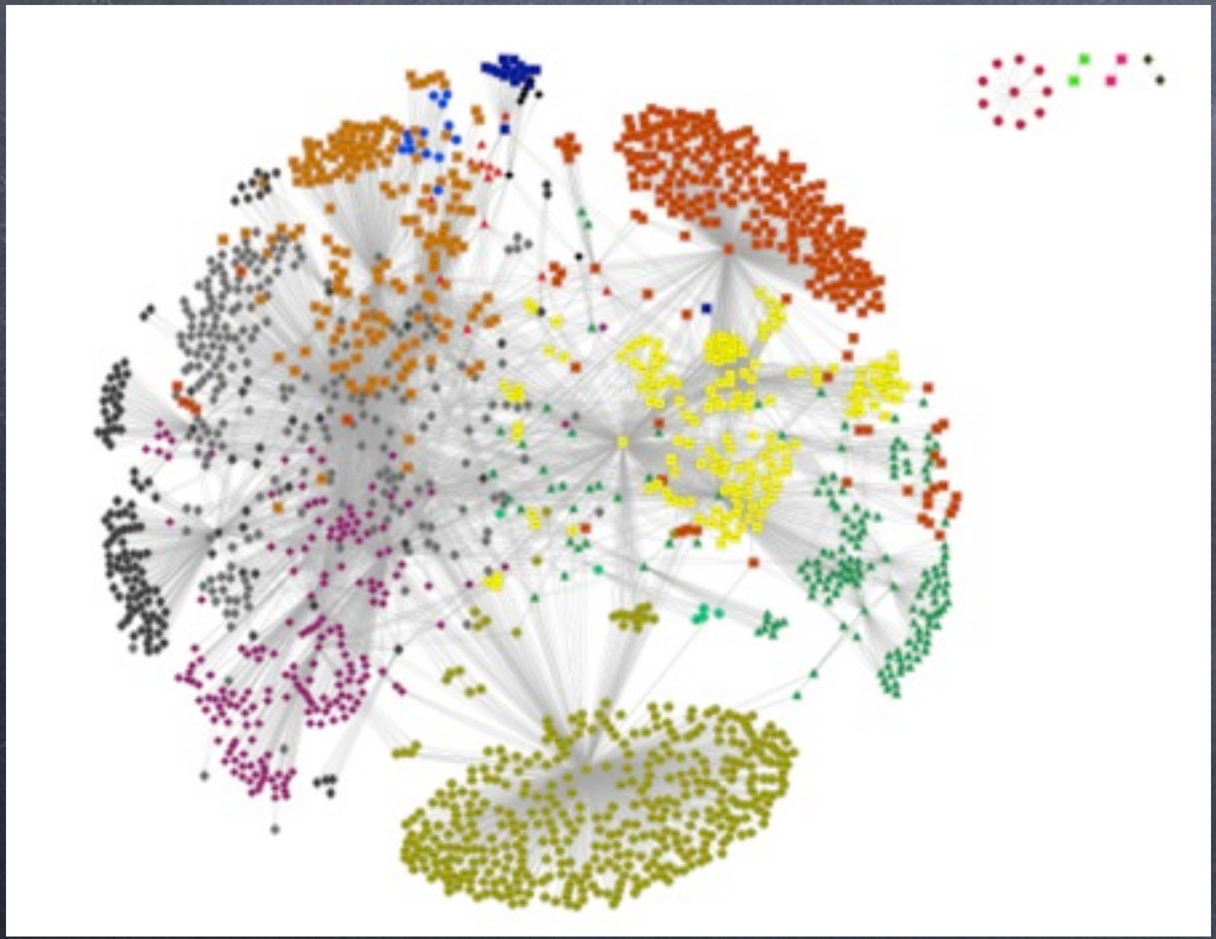
Clustering in sparse networks: Thresholds and optimal algorithms

Lenka Zdeborová
(IPhT, CNRS & CEA, Saclay, France)

in collaboration with
F. Krzakala, C. Moore, A. Decelle,
E. Mossel, A. Sly, J. Neeman, P. Zhang,
A. Saade, M. Lelarge.

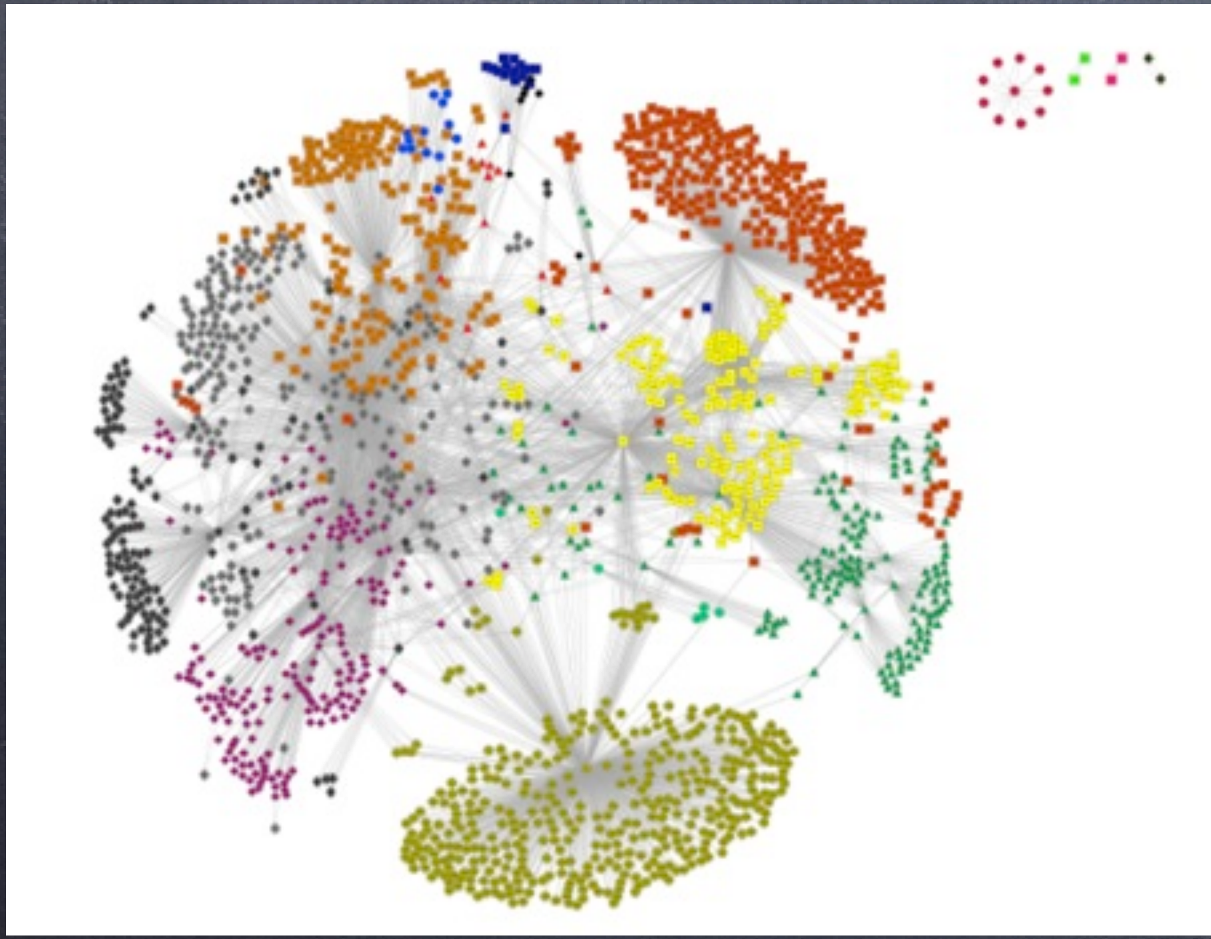
Networks

- Definition: Network = a graph = nodes/points, some pairs of nodes are connected by edges.
- Examples: Facebook friendships. People buying DVDs on Amazon. Authors citing other authors.



Clustering networks

- The problem: Given the graph, divide nodes into groups such that nodes in one group have similar pattern of connections to nodes in other groups. E.g. nodes in one groups are much more connected among themselves then to the other nodes.
- One of the basic methods in data processing. Dimensionality reduction, coarse-grained representation of the data.



- Organization in groups/ modules based on similarity = one of the basic structures data can have.

Algorithms for clustering networks

- Hundreds of different methods - for a review e.g.
  S. Fortunato, Physics Reports, 2010.

- Spectral clustering - the state of art clustering method in machine learning. Associate a matrix to the network (adjacency, random walk, Laplacian, etc.), compute its top eigenvalues. The corresponding eigenvectors encode the clusters in a "visible" way.

- Modularity maximization (Newman'06) - most popular among people studying complex networks.

$$Q = \sum_{(ij)} \left[A_{ij} - \frac{d_i d_j}{2M} \right] \delta_{s_i, s_j}$$

Optimal algorithms?

- What is a good criteria for optimality?
- Best performance of real data is tricky to evaluate.

Optimal algorithms?

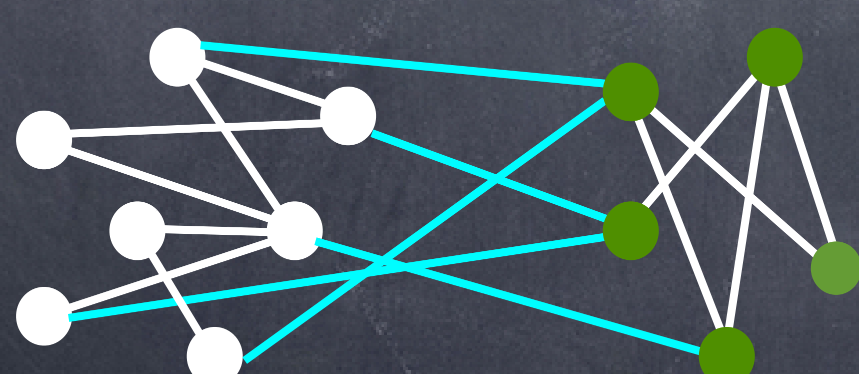
- What is a good criteria for optimality?
- Best performance of real data is tricky to evaluate.
- Alternatively: Consider a simple and natural model to generate clustered networks. Cluster the generated networks and compare to the "true clustering".
- Optimal algorithm is the such that maximizes the number of correctly labeled nodes.

Stochastic 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

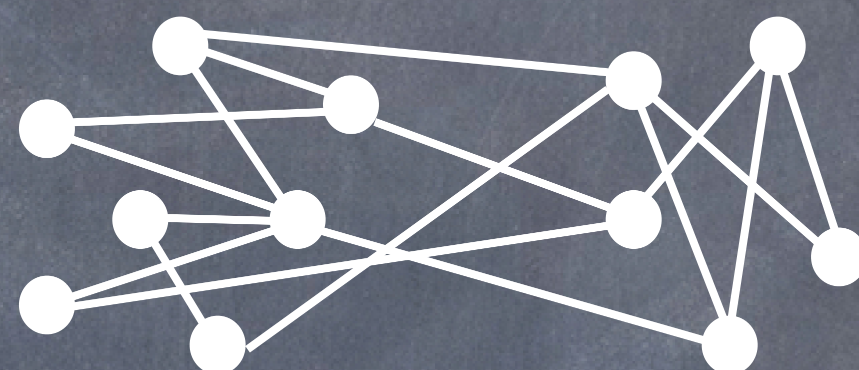
$n_1 = 7/12 \qquad n_2 = 5/12$



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

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

Goal in clustering



Given the graph, find back the assignment to groups.

In which limit do we work?

Large graphs: $N \to \infty$

Fixed number of large groups: $q = \Theta(1)$

(crucial for our method and results to be valid)

$$n_a = \frac{N_a}{N} = \Theta(1)$$

Sparse graphs: $p_{ab}N = c_{ab} = \Theta(1)$

(because this is the algorithmically challenging case, our results hold but are not very interesting for the denser case)

Bayes-optimal algorithm

(when parameters of the model known)

- Signal - assignment of nodes into groups $s_i \in \{1, \ldots, q\}$ with n_a fraction of nodes in each group.

- Measurement - the adjacency matrix A, with p_{ab} being the 'affinity' of group a to group b.

$$P(\{s_i\}) = \prod_{i=1}^{N} n_{s_i} \qquad P(A_{ij}|\{s_i\}) = \prod_{i \neq j} p_{s_i s_j}^{A_{ij}} (1 - p_{s_i s_j})^{1 - A_{ij}}$$

$$\boxed{P(\{s_i\}|A_{ij}) = \frac{1}{Z} P(\{s_i\}) P(A_{ij}|\{s_i\})}$$

Includes ALL available information about the signal

Bayes-optimal algorithm

Posterior probability distribution

$$P(\{s_i\}|A_{ij}) = \frac{1}{Z} P(\{s_i\}) P(A_{ij}|\{s_i\})$$

Marginal probabilities

$$\mu(s_i) \equiv \sum_{\{s_j\}_{j\neq i}} P(\{s_j\}_{j\neq i}, s_i|A_{ij})$$

Maximize number of correctly assigned nodes

$$s_i^* = \mathrm{argmax}_{s_i} \mu(s_i)$$

Bayes-optimal algorithm

Posterior probability distribution

$$P(\{s_i\}|A_{ij}) = \frac{1}{Z} P(\{s_i\}) P(A_{ij}|\{s_i\})$$

Marginal probabilities

$$\mu(s_i) \equiv \sum_{\{s_j\}_{j\neq i}} P(\{s_j\}_{j\neq i}, s_i|A_{ij})$$

Maximize number of correctly assigned nodes

$$s_i^* = \operatorname{argmax}_{s_i} \mu(s_i)$$

Trouble: Takes time exponential in N to evaluate.

Why not maximum-likelihood?

ML: $\max_{\{s_i\}} P(\{s_i\}|A_{ij})$ ✗

- In sparse graphs only O(1) edges per node = measurements per signal component.
- Curse of dimensionality: max likelihood over-fits and finds "good" costs even in a random graph.
- Random 3-regular graph has bi-partitions of nodes with only 11% of edges in between the two parts.

marginalize:

$$s_i^* = \mathrm{argmax}_{s_i} \mu(s_i)$$

$$\mu(s_i) \equiv \sum_{\{s_j\}_{j\neq i}} P(\{s_j\}_{j\neq i}, s_i|A_{ij})$$ ✓

How to compute the marginals?

- MCMC (Monte Carlo Markov chain, Gibbs sampling) general for discrete variables, issues with large equilibration time
- Variational mean field approximation - fast but often very wrong.
- Belief propagation, fast, in general better than mean field, and exact for large networks generated by the stochastic block model.

How to compute the marginals?

- MCMC (Monte Carlo Markov chain, Gibbs sampling) general for discrete variables, issues with large equilibration time
- Variational mean field approximation - fast but often very wrong.
- Belief propagation, fast, in general better than mean field, and exact for large networks generated by the stochastic block model.

Belief Propagation

BP in physics:

- Bethe, Peierls, Onsager'35
- Thouless-Anderson-Palmer'76
- Mezard, Parisi, Zecchina'02

BP in computer science:

- LDPC: Gallager'60
- Bayes inference: Pearl'82
- many generalizations ....

= message passing algorithm to find the clustering

= analysis tool (cavity method in physics) of the Bayes-optimal clustering in the limit of large networks

Belief Propagation for SBM

from Decelle, Krzakala, Moore, Zdeborova'11

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

$$h_{s_i} = \frac{1}{N} \sum_k \sum_{s_k} c_{s_k s_i} \psi_{s_k}^k$$

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

BP runs in linear time $O(cq^2 N)$

BP asymptotically exact here (the relevant part of the factor graph is tree-like). Remains to be proven rigorously

$$\mu^i(s_i) = \psi_{s_i}^i + o(1)$$

Bethe log-likelihood

$$\log Z_{\mathrm{BP}}(q,\{n_a\},\{c_{ab}\}) = \sum_i \log Z^i - \sum_{(i,j)\in E} \log Z^{ij} + \frac{cN}{2}$$

$$Z^{ij} = \sum_{a<b} c_{ab}(\psi_a^{i\to j}\psi_b^{j\to i} + \psi_b^{i\to j}\psi_a^{j\to i}) + \sum_a c_{aa}\psi_a^{i\to j}\psi_a^{j\to i}$$

$$Z^i = \sum_{t_i} n_{t_i} e^{-h_{t_i}} \prod_{j\in\partial i} \sum_{t_j} c_{t_j t_i} \psi_{t_j}^{k\to i}$$

-> model selection

-> expectation-maximization-like learning

How to use BP to analyze the Bayes optimal inference?

Iterate BP from two different initializations:

(1) Random $\psi^{i\to j}$

(2) Planted: $\psi^{i\to j}_{s\neq s^*} = 0 \quad \psi^{i\to j}_{s^*} = 1$

If you find two different fixed points, take the one with larger Bethe likelihoods.
This is what Bayes-optimal sampling would do.

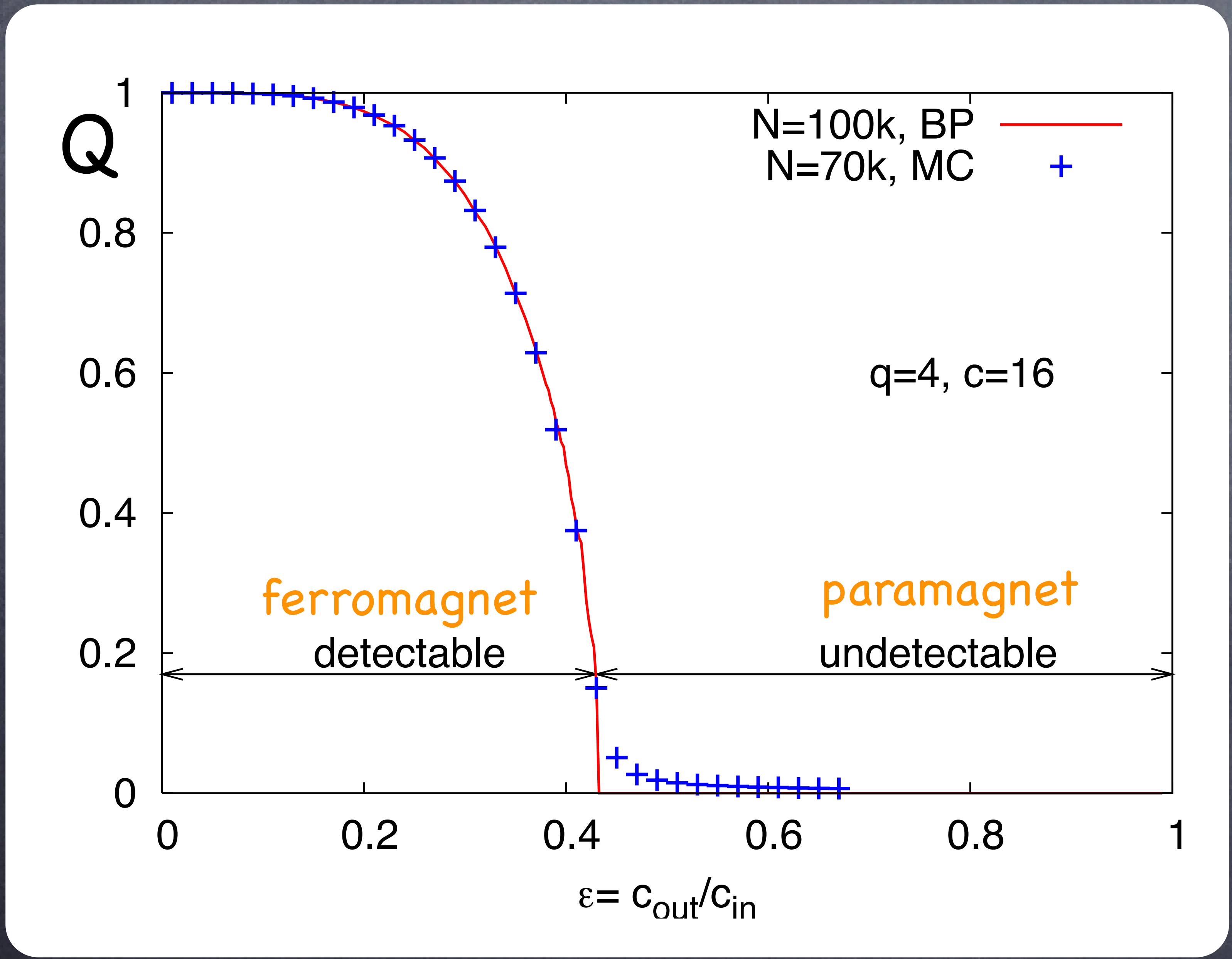
Holds for arbitrary $q, n_a, c_{ab}, N \to \infty$

Results

(from Decelle, Krzakala, Moore, Zdeborova'11)

$$Q = \max_{\pi} \frac{\sum_{i=1}^{N} \delta_{\pi(t_i), s_i^*}/N - 1/q}{1 - 1/q}$$

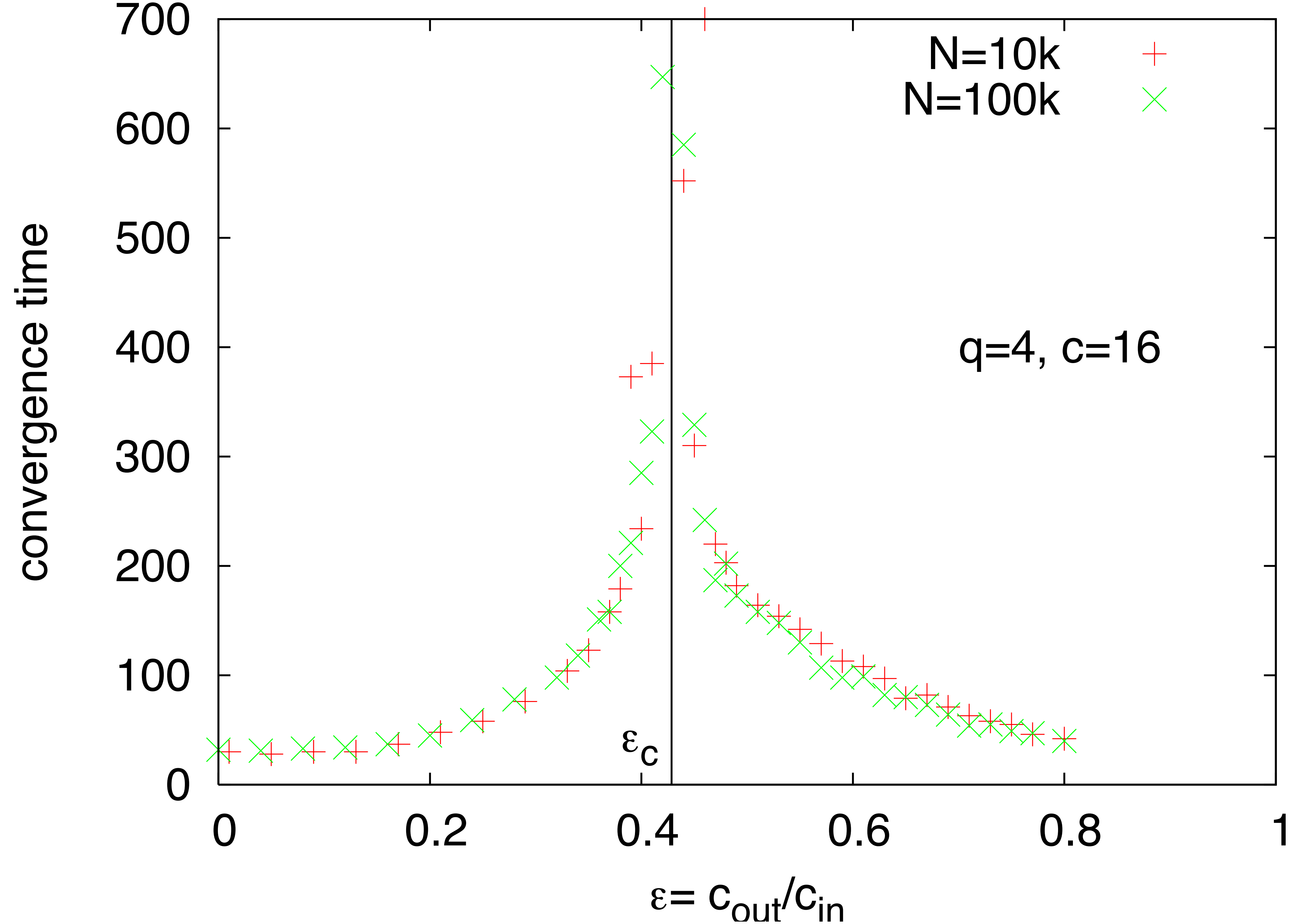
Overlap of the optimal estimation with the true labeling.



$$c_{aa} = c_{\text{in}}$$

$$c_{ab} = c_{\text{out}} \qquad a \neq b$$

Equilibration time



Locating the phase transition

(Thouless'86, critical temperature on Bethe-lattice spin glass, Kesten-Stigum'66 bound for Galton-Watson processes)

- When average degree in every group is the same: Stability of the uniform BP fixed point $\psi_a^{i\to j} = n_a$.
- Study the influence of a message change along a chain, one step:

$$\boxed{T^{ab} \equiv \frac{\partial \psi_a^{k_i}}{\partial \psi_b^{k_{i+1}}}\bigg|_{\psi_t = n_t} = n_a\left(\frac{c_{ab}}{c} - 1\right)}$$

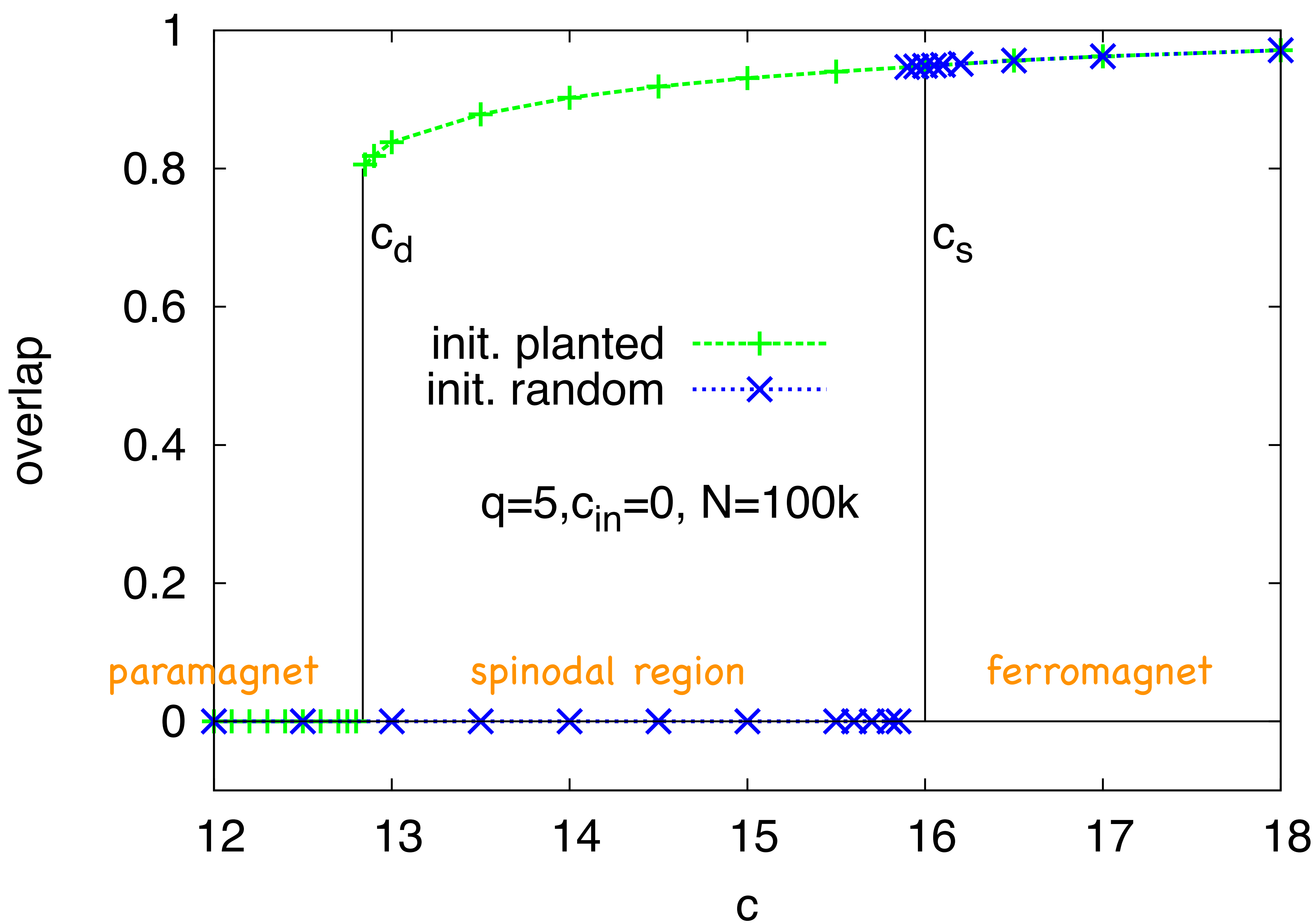
- Take largest eigenvalue of T. Variance of a message-change on the root due to changed boundary $\mathrm{var} \approx c^d \lambda^{2d}$ Stability $1 > c\lambda^2$. BP stays at the trivial fixed point.
- Special case of groups of equal sizes, and only in/out probabilities different: $|c_{\text{in}} - c_{\text{out}}| < q\sqrt{c}$

Proof of the phase transition for q=2

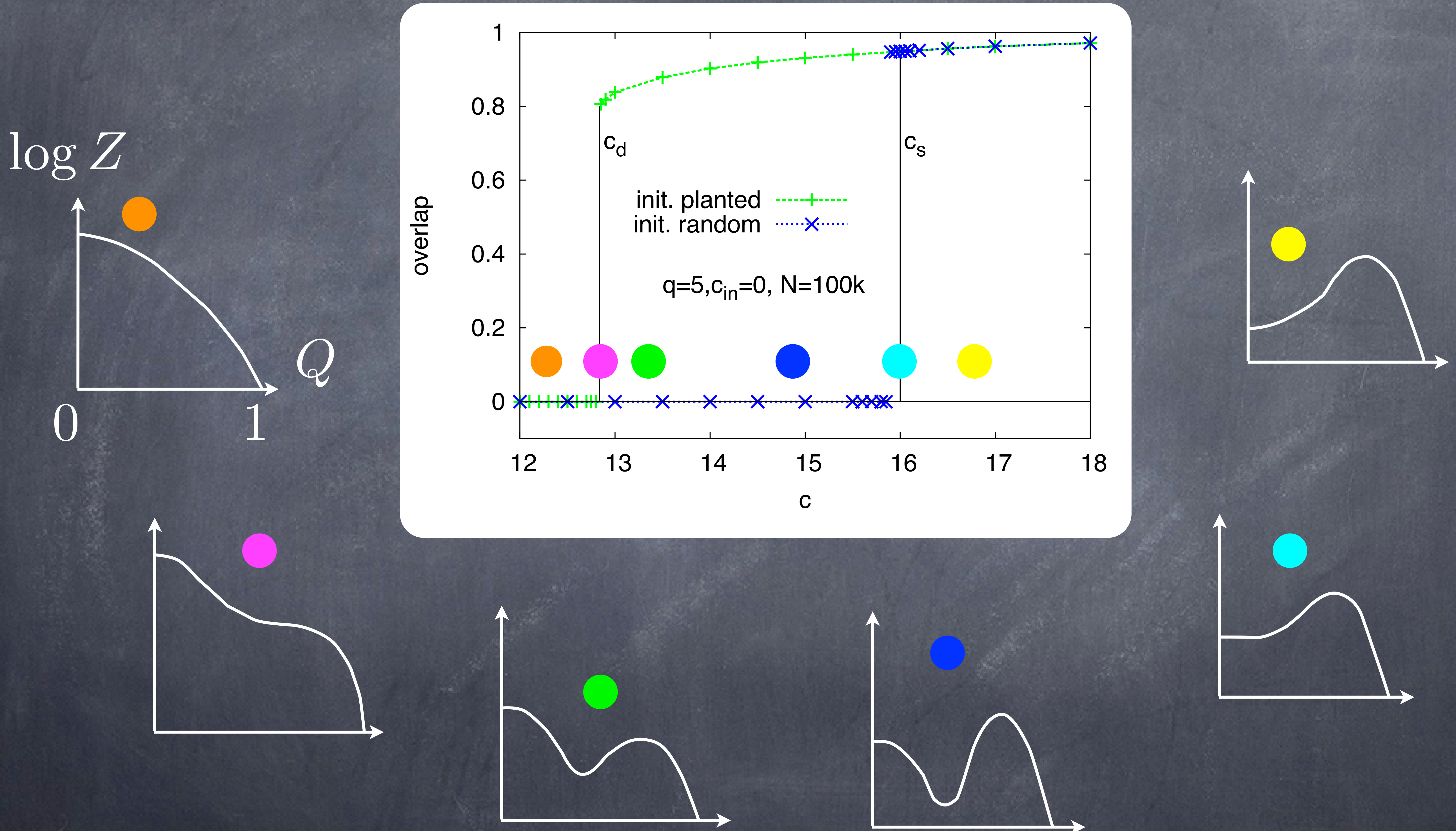
- The undetectable regime (Mossel, Neeman, Sly'12), SBM contiguous to Erdos-Renyi random graph.
- The detectable regime (Massoulie'13, Mossel, Neeman, Sly'13). Both proofs use a constructive but not very practical (polynomial) algorithm.

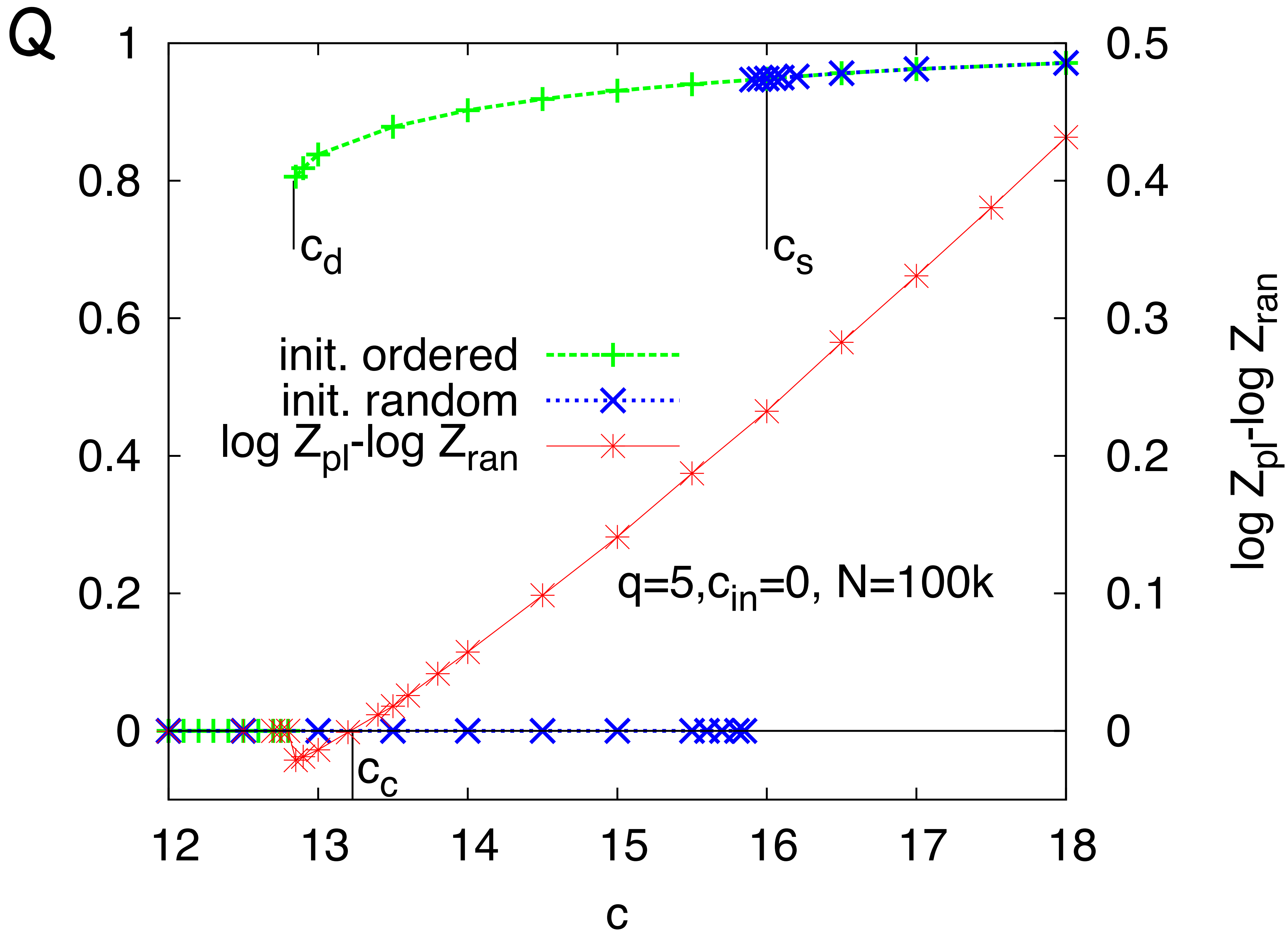
1st order phase transitions
even more interesting

$c_{\rm in} = 0$ dis-assortative network



1st order phase transition





Q
init. ordered
init. random
log Z_{pl}-log Z_{ran}
c_d
c_s
c_c
q=5,c_{in}=0, N=100k
c
log Z_{pl}-log Z_{ran}

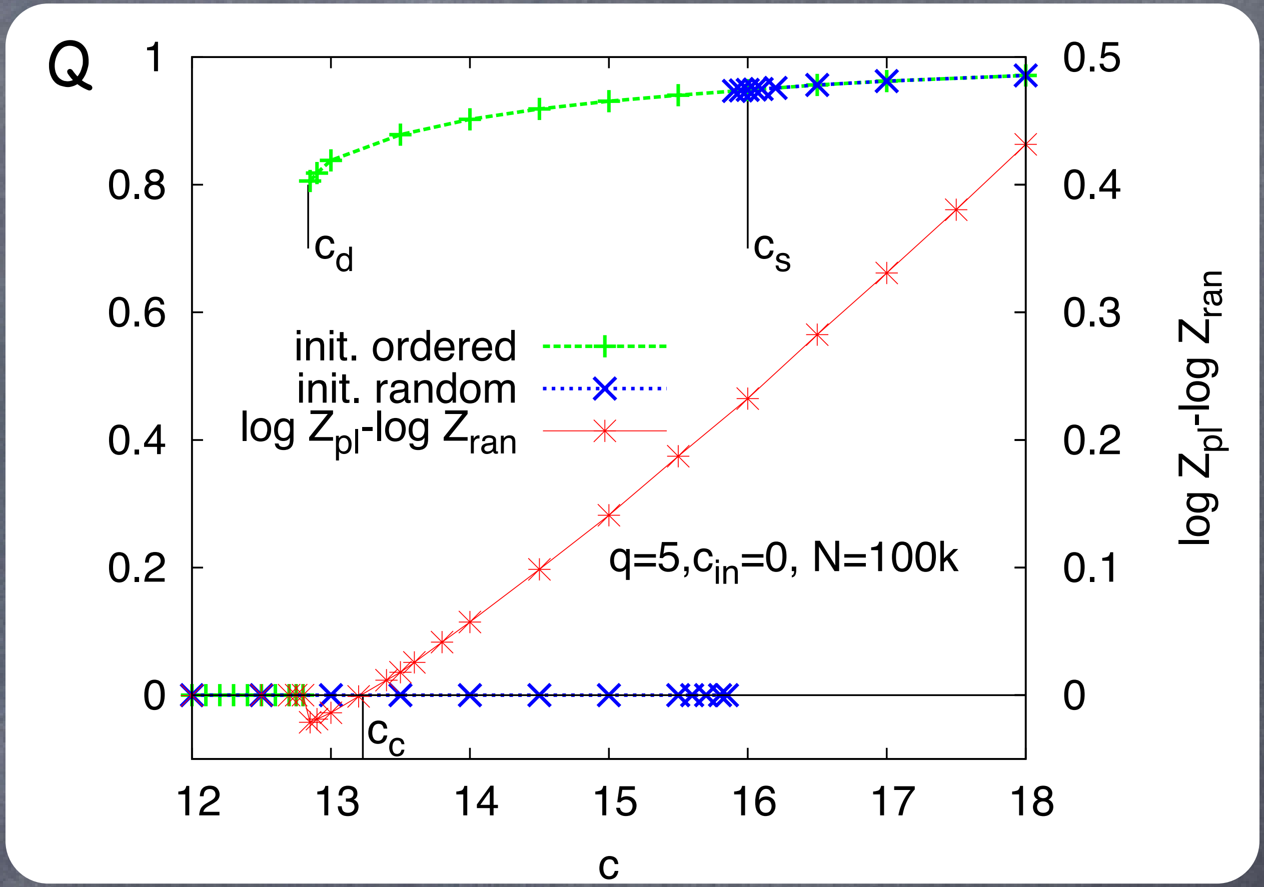
Algorithmic consequences

planted coloring:

$$c_{\text{in}} = 0$$

$$\lim_{q\to\infty} \frac{c_c(q)}{2q\log q} = 1$$

$$\lim_{q\to\infty} \frac{c_s(q)}{q^2} = 1$$



$c < c_c$ impossible

$c_c < c < c_s$ possible but hard

$c_s < c$ easy

The hard-easy phase transition

- Spinodal line of the 1st order phase transition. MCMC equilibration time and BP convergence time diverge.
- Algorithmic barrier in many inference problems: planted clique, compressed sensing, error correcting codes, planted constraint satisfaction, sparse PCA, matrix factorization,...
- Conjecture: Algorithmic barrier for a large class of algorithms. Including BP, Gibbs sampling (MCMC), spectral, stochastic local search. NOT including Gauss elimination.
- Open problem: Establish formal connections between this phase transition and performance of (some class of) polynomial algorithms.

Back to the optimal algorithms

- BP asymptotically optimal in the SBM and linear in N.

BUT

- Quadratic in the number of clusters
- Needs to know the parameters of the graph or learn them via iterative method full of local optima.
- Bothered by small loops in graphs.

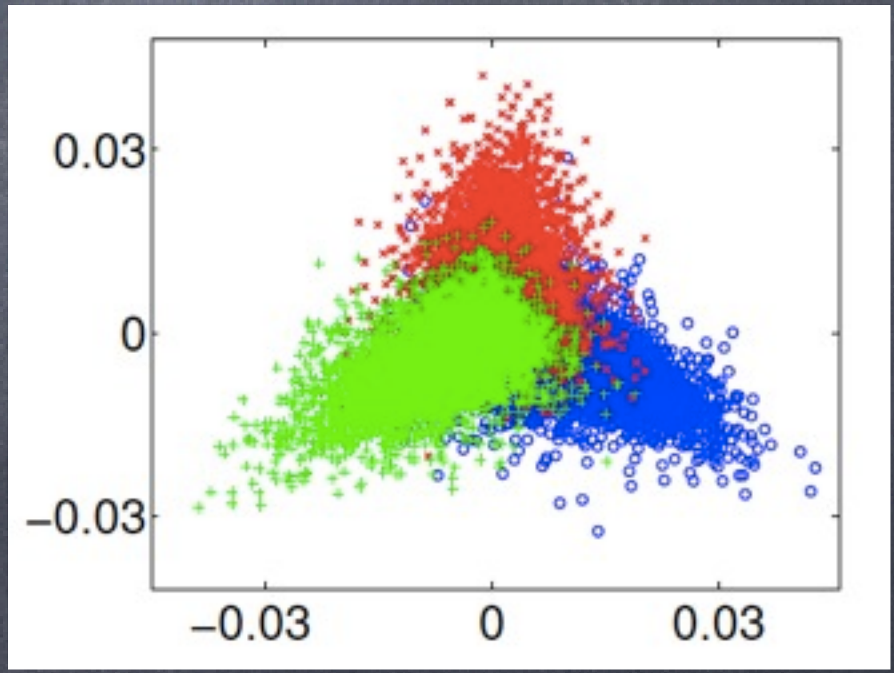
Is there a method that keeps the advantages and does not have these disadvantages?

Spectral clustering

- Compute q largest eigenvalues and their eigenvectors for a matrix associated to the graph.
- Cluster components of these eigenvectors, e.g. using k-means. For 2 groups - signs of the 2nd eigenvector.

Matrices:

Adjacency	A_{ij}
Laplacian	$L_{ij} = d_i \delta_{i,j} - A_{ij}$
Random walk	$Q_{ij} = \dfrac{A_{ij}}{d_i}$
Modularity	$M_{ij} = A_{ij} - \dfrac{d_i d_j}{2M}$



Trouble with sparse graphs

Wigner not good for sparse G

$$P(\lambda) = \frac{1}{2\pi c}\sqrt{4c-\lambda^2}$$

- The edge of the spectrum of sparse graphs is spoiled by nodes of large degree.
- ER graphs, largest degree ~ log(N)/log(log(N)).
- How to correct this? Remove largest degrees? Not good enough - loosing information.

Properties of the spectrum for large random sparse graphs

- Largest eigenvalue $\lambda_{\max}$ is the average excess degree, bulk inside circle of radius $\sqrt{\lambda_{\max}}$
- For graphs generated by the stochastic block model: If every group the same average degree then q-1 real eigenvalues $\mu = c\nu$, where $\nu \neq 0$ eigenvalue of

$$T_{ab} = n_a \left(\frac{c_{ab}}{c} - 1 \right)$$

Rigorous proof:

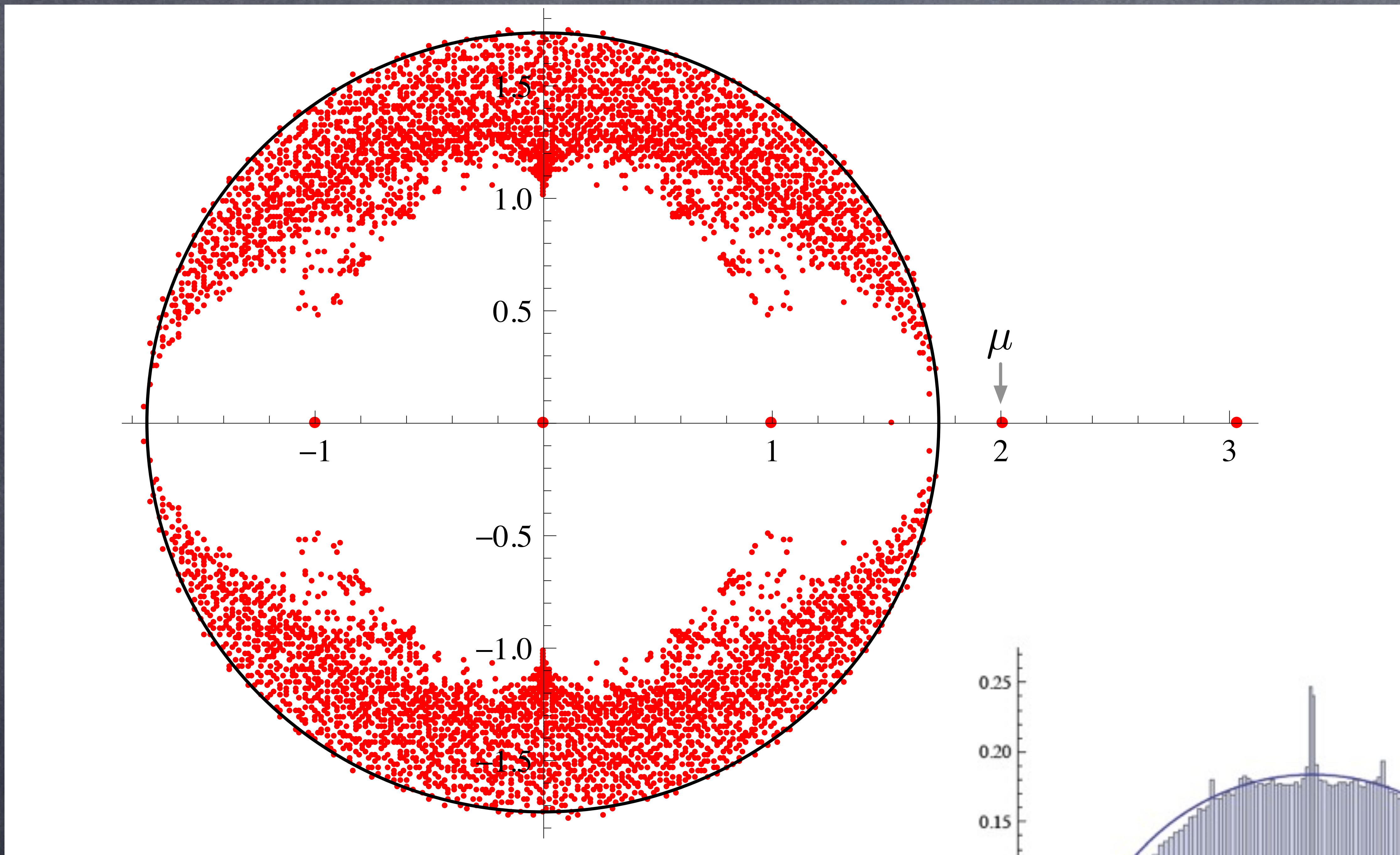
Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs

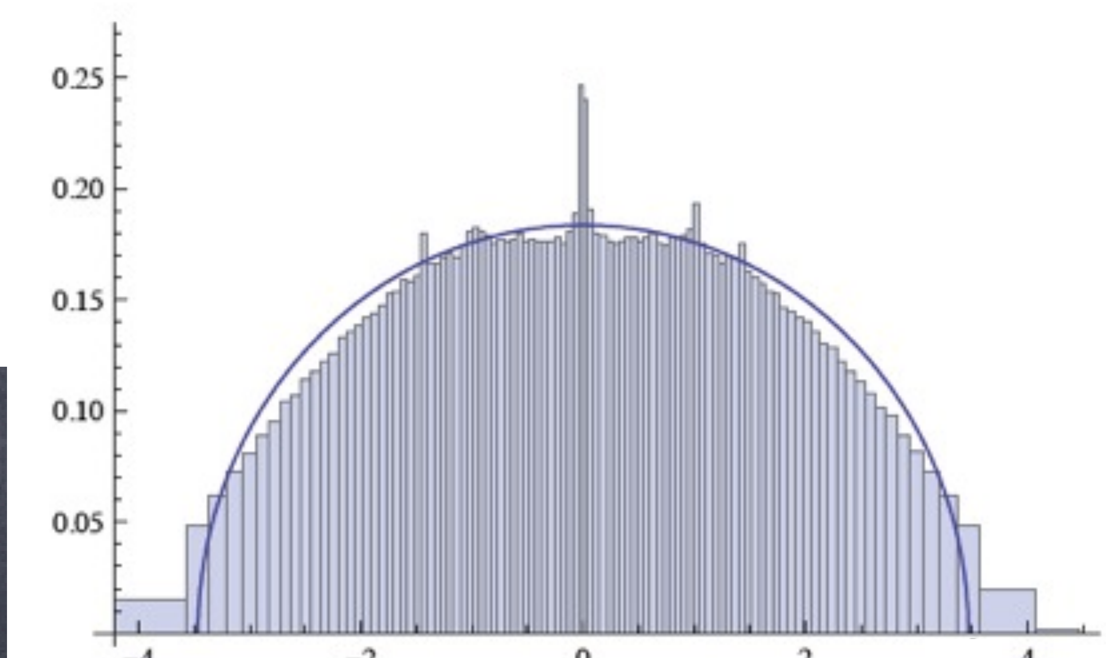
Charles Bordenave, Marc Lelarge, Laurent Massoulié

(Submitted on 24 Jan 2015)

A non-backtracking walk on a graph is a directed path such that no edge is the inverse of its preceding edge. The non-backtracking matrix of a graph is indexed by its directed edges and can be used to count non-backtracking walks of a given length. It has been used recently in the context of community detection and has appeared previously in connection with the Ihara zeta function and in some generalizations of Ramanujan graphs. In this work, we study the largest eigenvalues of the non-backtracking matrix of the Erdos-Renyi random graph and of the Stochastic Block Model in the regime where the number of edges is proportional to the number of vertices. Our results confirm the "spectral redemption conjecture" that community detection can be made on the basis of the leading eigenvectors above the feasibility threshold.

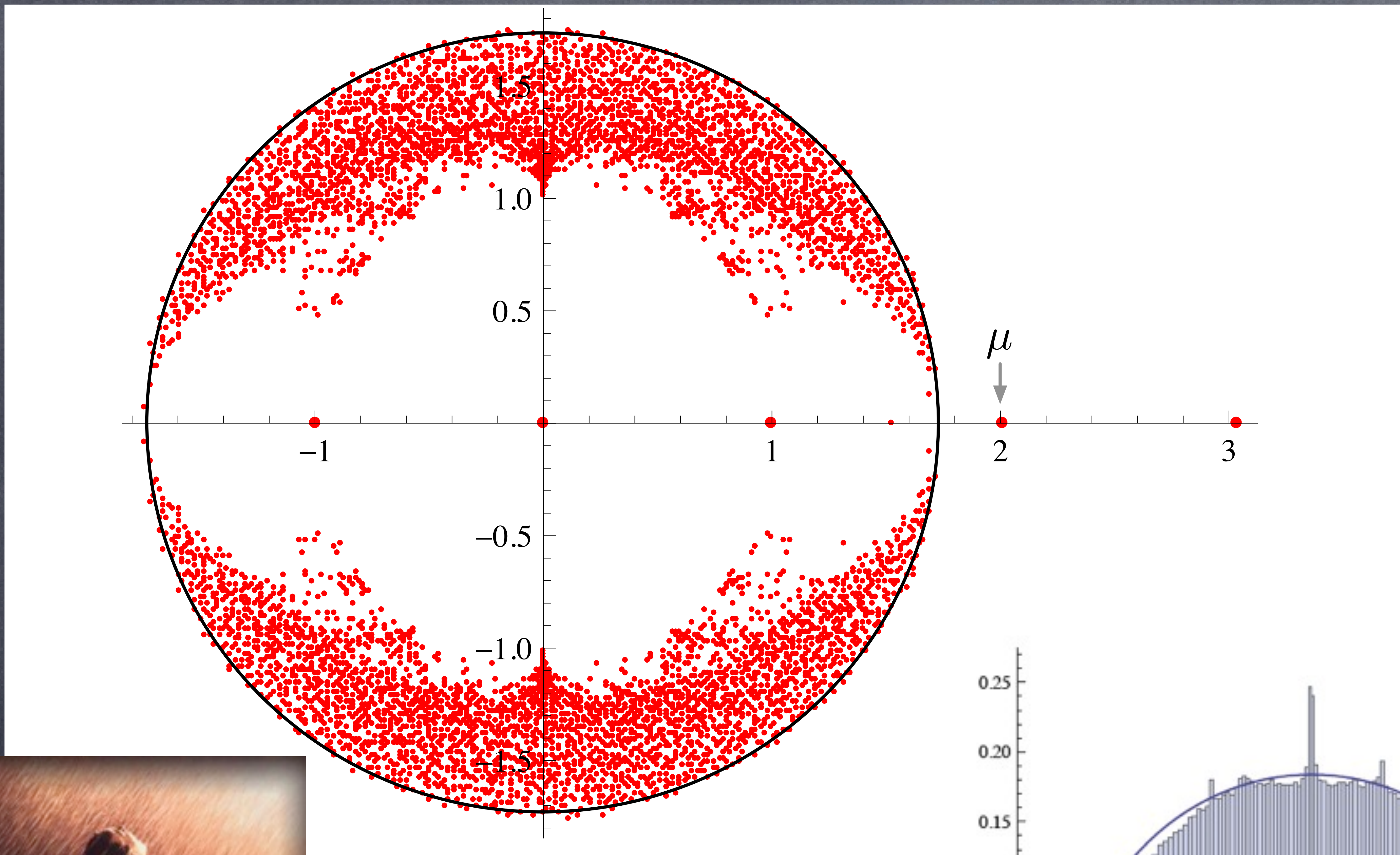
Spectral Redemption

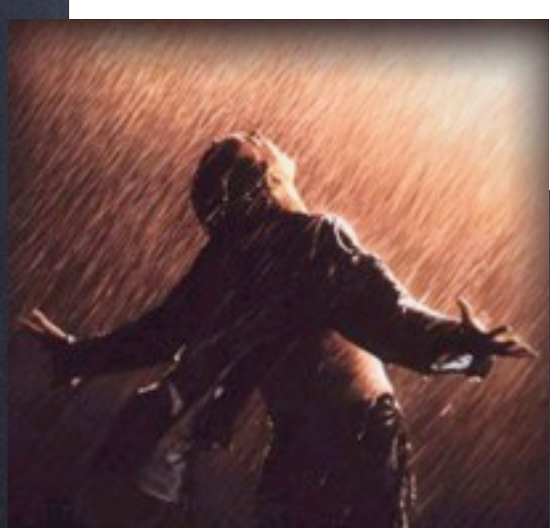


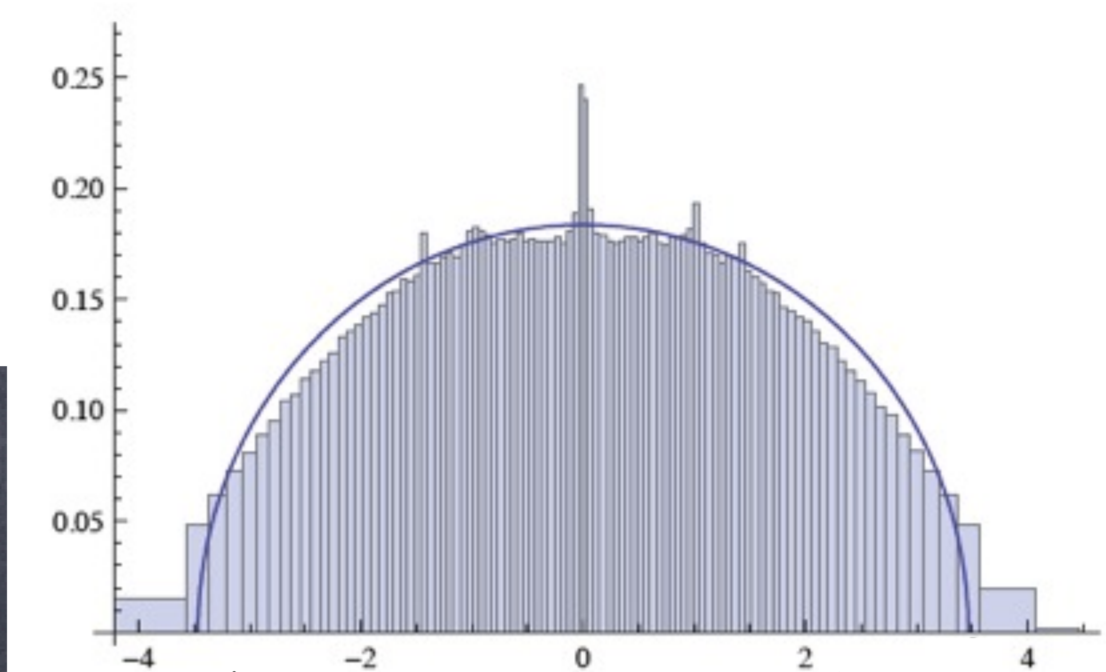


adjacency matrix spectrum

Spectral Redemption

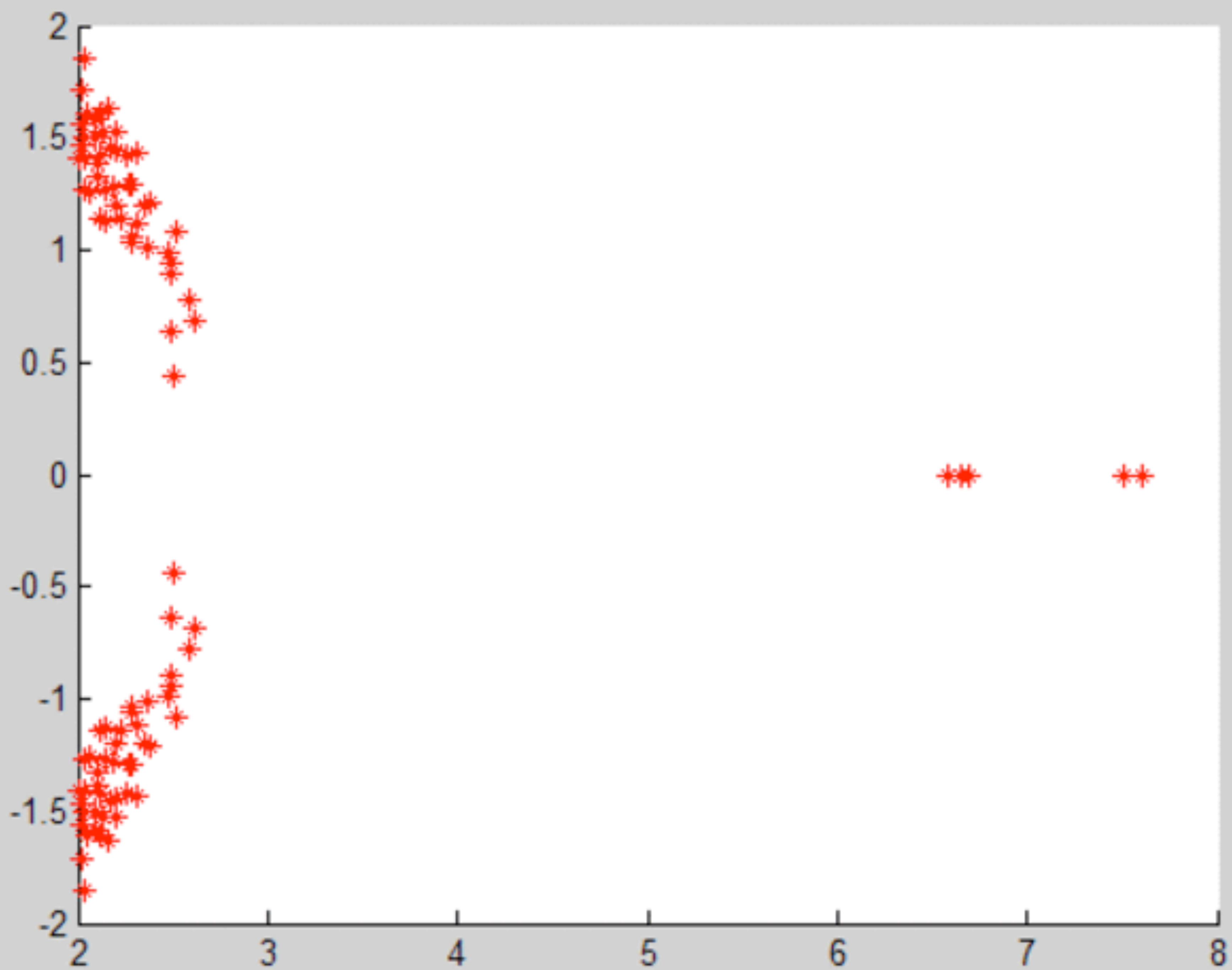




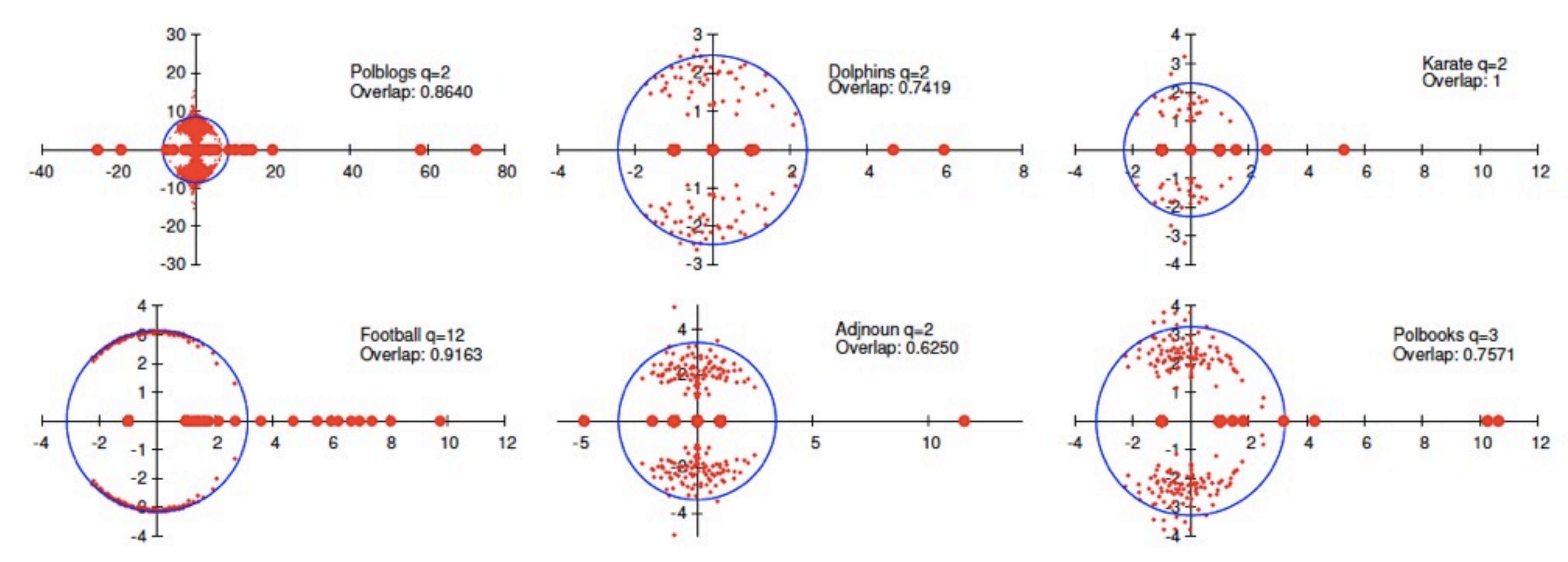


adjacency matrix spectrum

Spectrum of the non-backtracking matrix as $c_{\mathrm{in}}/c_{\mathrm{out}}$ decreases (fixed N, c)

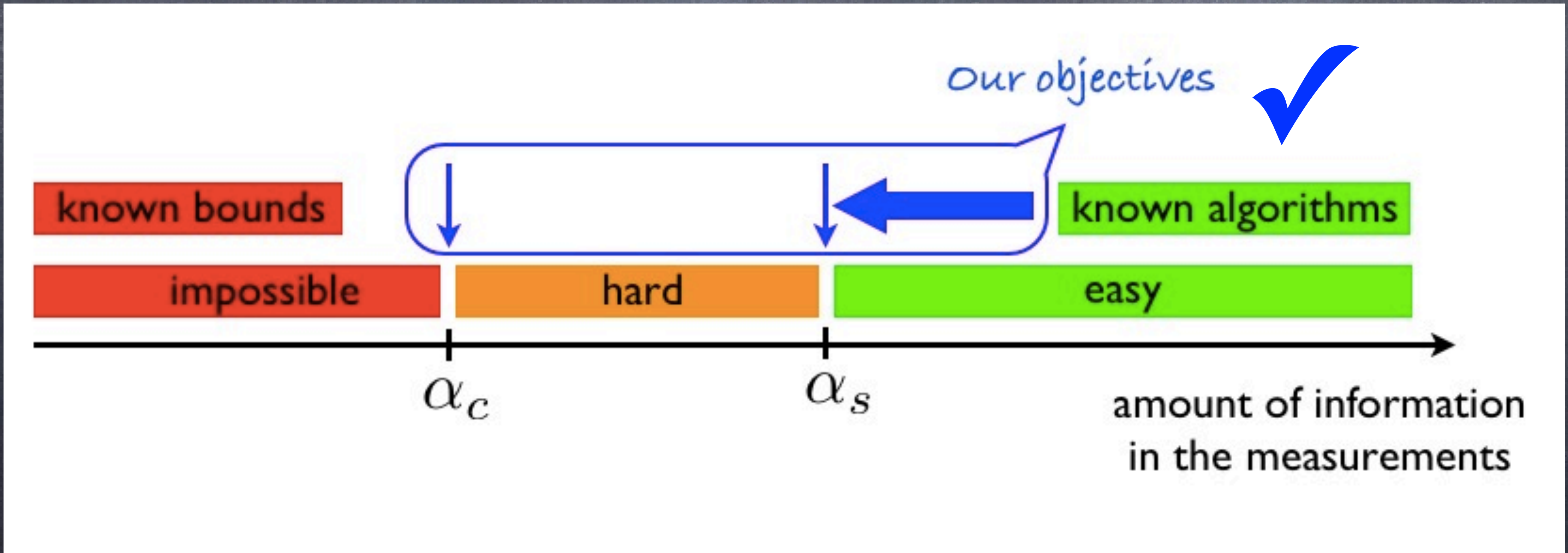


Spectra of some real networks



Statistical physics of inference problems.

- **Phase transitions** in inference are significant for **algorithmic average hardness** impossible/possible or hard/easy. The same pattern in many problems.
- **Design of new algorithms.** Message passing, non-backtracking for clustering of sparse networks ....
- Precise **conjectures** waiting for proofs.



This talk is based on:

- A. Decelle, F. Krzakala, C. Moore, LZ, Phase transition in the detection of modules in sparse networks, Phys. Rev. Lett. 107, 065701 (2011).

- A. Decelle, F. Krzakala, C. Moore, LZ, Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications, Phys. Rev. E 84, 066106 (2011).

- F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, LZ, P. Zhang, Spectral clustering of Sparse Networks, PNAS 110, 20935 (2013).

- Implementations available at: http://mode_net.krzakala.org/

This talk is based on:

- A. Decelle, F. Krzakala, C. Moore, LZ, Phase transition in the detection of modules in sparse networks, Phys. Rev. Lett. 107, 065701 (2011).
- A. Decelle, F. Krzakala, C. Moore, LZ, Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications, Phys. Rev. E 84, 066106 (2011).
- F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, LZ, P. Zhang, Spectral clustering of Sparse Networks, PNAS 110, 20935 (2013).
- Implementations available at: http://mode_net.krzakala.org/

Thank you for your attention!

Belief Propagation for SBM

from Decelle, Krzakala, Moore, Zdeborova'11

$$Z = \sum_{\{q_i\}} \prod_{i=1}^{N} n_{q_i} \prod_{i \neq j} p_{q_i q_j}^{A_{ij}} (1 - p_{q_i q_j})^{1 - A_{ij}}$$

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

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

BP runs in linear time $O(cq^2 N)$

BP asymptotically exact here (the relevant part of the factor graph is tree-like and no RSB for optimal Bayesian inference.) $\mu^i(q_i) = \psi_{q_i}^i + o(1)$

Belief Propagation for SBM

from Decelle, Krzakala, Moore, Zdeborova'11

$$Z = \sum_{\{q_i\}} \prod_{i=1}^{N} n_{q_i} \prod_{i \neq j} p_{q_i q_j}^{A_{ij}} (1 - p_{q_i q_j})^{1 - A_{ij}}$$

BP iteration equations

$$\psi_{q_i}^{i \to j} = \frac{1}{Z^{i \to j}} n_{q_i} \prod_{k \neq i,j} \left[\sum_{q_k} p_{q_i q_k}^{A_{ik}} (1 - p_{q_i q_k})^{1 - A_{ik}} \psi_{q_k}^{k \to i} \right]$$

Little problems:

Factor graph is fully connected = many short loops.

BP runs in N^2 steps - not good for N= 1 million.

$$p_{ab} = \frac{c_{ab}}{N} \qquad\qquad N \to \infty, c = O(1)$$

Beliefs on non-edges (A_ij=0) do not depend in the leading order on the target-node.

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

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

BP marginals: $\psi_{q_i}^{i} = \frac{1}{Z^{i}} n_{q_i} e^{-h_{q_i}} \prod_{j\in\partial i} \left[\sum_{q_j} c_{q_j q_i} \psi_{q_j}^{j\to i}\right]$

Final BP for SBM

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

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

BP runs in linear time $O(cq^2 N)$

BP asymptotically exact here (the relevant part of the factor graph is tree-like and no RSB for optimal Bayesian inference.)

$$\mu^i(q_i) = \psi_{q_i}^i + o(1)$$

Learning parameters

Learning

Given the graph, what is the best estimate for

$$\{n_a, p_{ab}\}$$

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

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

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

Stationarity conditions

(expectation maximization learning):

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

EM learning with BP

$$n_a = \frac{1}{N} \sum_i \psi_a^i$$

$$c_{ab} = \frac{1}{N} \frac{1}{n_b n_a} \sum_{(i,j)\in E} \frac{c_{ab}(\psi_a^{i\to j}\psi_b^{j\to i} + \psi_b^{i\to j}\psi_a^{j\to i})}{Z^{ij}}$$