Average hardness and phase transitions in random CSP

Lenka Zdeborová
(IPhT, CNRS & CEA, Saclay, France)
with
F. Krzakala, M. Mezard, A. Coja-Oghlan,
F. Ricci-Tersenghi, A. Montanari, G. Semerjian ....

Motivation:
What makes problems hard?

Motivation: What makes problems hard?

Computational complexity
worst vs average

- K-SAT = satisfiability of Boolean formulas: NP complete (Cook 1971) - concerns worst case computational complexity.
- Average computational complexity - in what time on average can a large fraction of instances be solved.

Historical note

- 1971 - Cook proves K-SAT to be NP-complete.
- 1979 - Golberg shows that if a variable is included in a clause with fixed probability this ensemble is on average polynomial.
- Until 1991 basically all computer science believes that NP-complete problems might in fact all be on average easy.

1991

Where the *Really* Hard Problems Are

Peter Cheeseman
RIACS*

Bob Kanefsky
Sterling Software

William M. Taylor
Sterling Software

Artificial Intelligence Research Branch
NASA Ames Research Center, Mail Stop 244-17
Moffett Field, CA 94035, USA
Email: <last-name>@ptolemy.arc.nasa.gov

Peter Cheeseman, Bob Kanefsky, and William M. Taylor. Where the really hard problems are. In J. Mylopoulos and R. Reiter, editors, *Proceedings of IJCAI-91*, pages 331-337, San Mateo, CA, 1991. Morgan Kaufmann.

Abstract

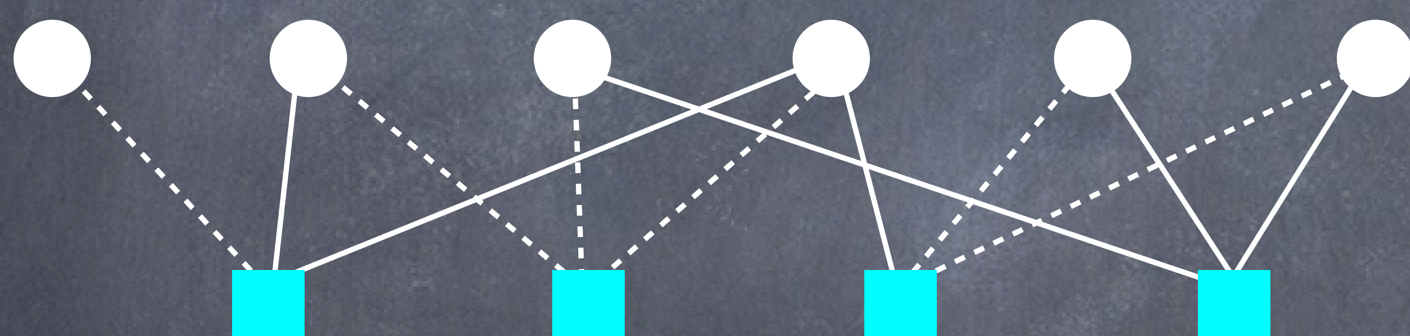
It is well known that for many NP-complete problems, such as K-Sat, etc., typical cases are easy to solve; so that computationally hard cases must be rare (assuming $P \neq NP$). This paper shows that NP-complete problems can be summarized by at least one "order parameter", and that the hard problems occur at a critical value of such a parameter. This critical value separates two regions of charac-

so says nothing about the difficulty of typical instances. However, this situation raises the question "where are the really hard instances of NP problems?". Can a subclass of problems be defined that is typically (exponentially) hard to solve, or do worst cases appear as rare "pathological cases" scattered unpredictably in the problem space?

In this paper we show that for many NP problems one or more "order parameters" can be defined, and hard instances occur around particular critical values of these order parameters. In addition, such critical values form a boundary that separates the space of problems into two regions. One region is underconstrained, so the density of solutions is high, thus making it relatively easy to find a solution. The other region is overconstrained and very unlikely to contain a solution. If there are solu-

Random K-SAT

- N variables
- randomly choose M K-uples of variables
- negate with probability 1/2

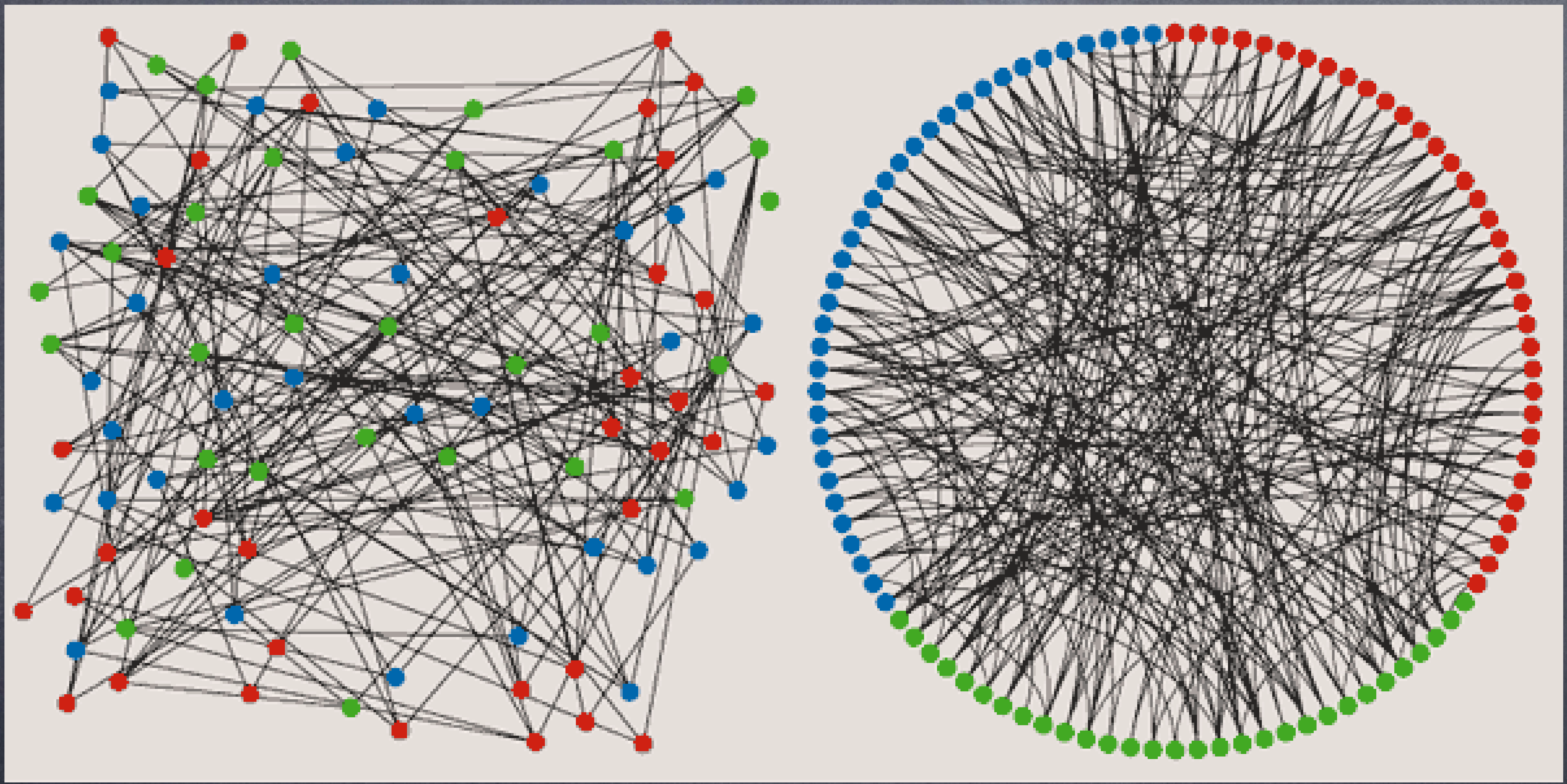


$$N = 6, M = 4, K = 3$$

$$\alpha = \frac{M}{N}$$

$$N \to \infty$$
$$M \to \infty$$

Definition of graph coloring



q=3: number of colors

Random Graph Coloring

- Erdos-Renyi random graph: Every edge present with probability $p=c/(N-1)$.

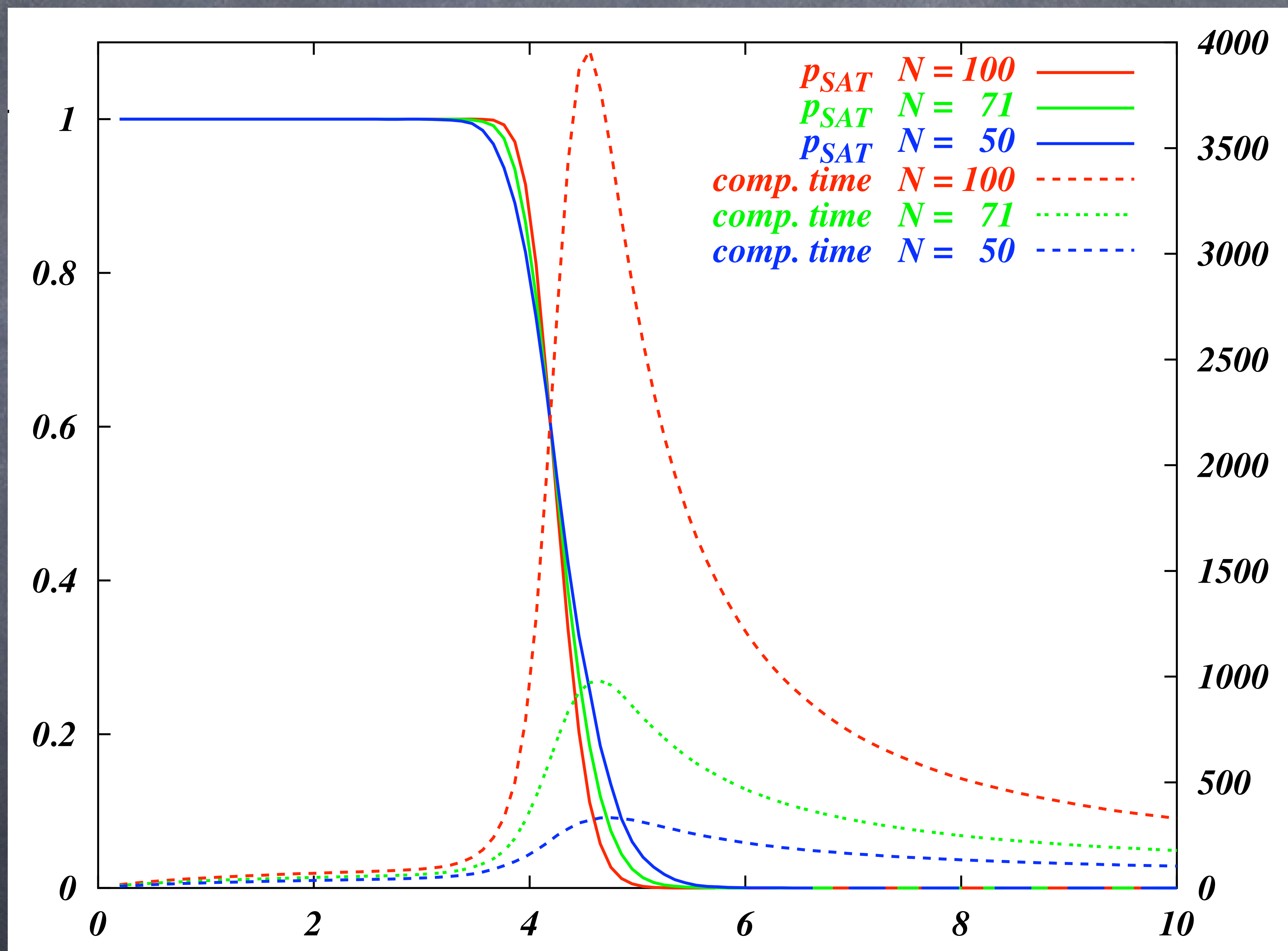
- Random regular graph: every node has degree r.

- Planted random graph: Fix a random color for every vertex, put M edges randomly only among different colors. Forget the "planted" coloring.

Where the really hard problems are?

(Cheeseman, Kanefsky, Taylor'91; Mitchell, Selman, Levesque'92)

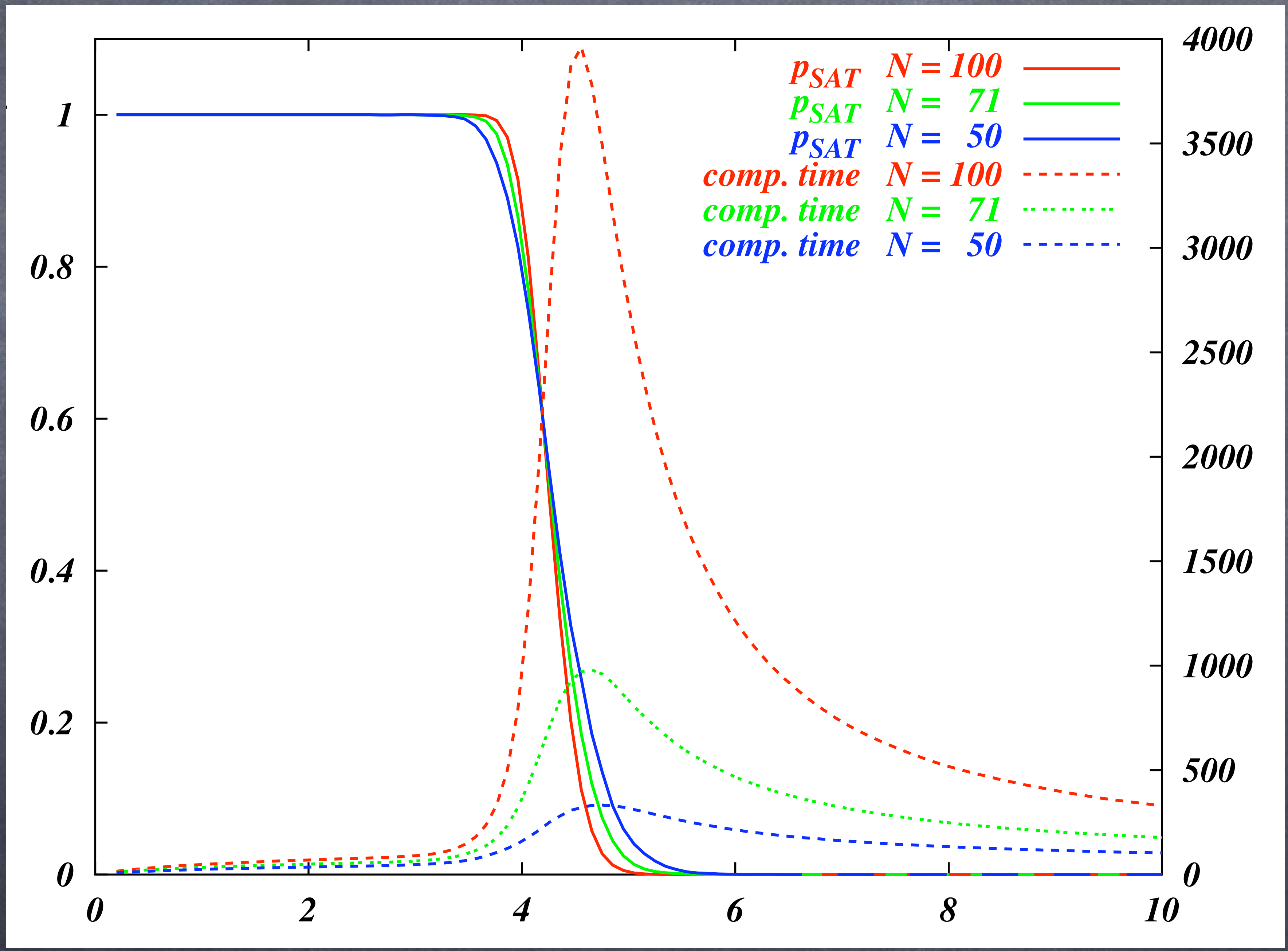
probability of colorability



average degree c

Where the really hard problems are?

(Cheeseman, Kanefsky, Taylor'91; Mitchell, Selman, Levesque'92)



time to decide

average degree c

Large number of colors

$q \to \infty$ $\qquad c_s \approx 2q \ln q - \ln q$

1st & 2nd moment on something smart *(Coja-Oghlan, Vilenchik, 2013)*

$$\mathbb{E}(Z) = q^N \left(1 - \frac{1}{q}\right)^{\frac{cN}{2}}$$

Naive algorithm which works for connectivities $c < q \ln q$:

Repeatedly pick a random vertex and assign it a random color not assigned to any of its neighbours.

Large number of colors

$q \to \infty$ $$c_s \approx 2q \ln q - \ln q$$

1st & 2nd moment on something smart (Coja-Oghlan, Vilenchik, 2013)

$$\mathbb{E}(Z) = q^N \left(1 - \frac{1}{q}\right)^{\frac{cN}{2}}$$

Naive algorithm which works for connectivities $c < q \ln q$:

Repeatedly pick a random vertex and assign it a random color not assigned to any of its neighbours.

An open question (for 30 years):

Is there a polynomial algorithm which would provably color graphs of connectivity $c \sim (1+\epsilon) q \ln q$ for some $\epsilon > 0$?

Physics comes into the game

- Monasson, Zecchina'97 realized random K-SAT = spin glass. Hence replica and cavity method developed for spin glasses useful to describe its properties.
- Technical problems in the method for sparse graphs resolved by Mezard, Parisi in 2001.
- Mezard, Parisi, Zecchina`2002
  - Computed the SAT/UNSAT transition,
  - Predicted clustering of solutions causing hardness
  - Invented survey propagation - best to solve random SAT instances for already more than 10 years.

Conjecture: random r-regular graph is q-colorable iff $\Sigma \geq 0$

$$\Sigma = \ln\left[\sum_{i=0}^{q-1}(-1)^i\binom{q}{i+1}(1-(i+1)\eta)^r\right] - \frac{r}{2}\ln\left(1-q\eta^2\right)$$

where η is the largest solution of

$$\eta = \frac{\sum_{i=0}^{q-1}(-1)^i\binom{q-1}{i}(1-(i+1)\eta)^{r-1}}{\sum_{i=0}^{q-1}(-1)^i\binom{q}{i+1}(1-(i+1)\eta)^{r-1}}$$

Mezard, Parisi, Zecchina, Weigt, Pagnani,
Krzakala, Ricci-Tersenghi, Montanari 2002-2004.

More phase transitions

L. Zdeborova, F. Krzakala, Phys. Rev. E 2007, EPL 2007, PRL 2009, etc.

(I) Clustering of solutions

(II) Condensation

(III) Spinodal transition

(IV) Freezing of solutions

The clustering transition

- Consider a random walk (try to flip a color at random, if still a valid coloring accept, if not try again) among colorings starting from a coloring chosen uniformly at random from all of them.

- Conjecture about the clustering threshold: For $c < c_d$ the walk will go to distance close to (q-1)/q in a constant number of steps per node. For $c > c_d$ it will stay closer than 1/2 forever (in large N limit).

- The set of solutions divides in exponentially many exponentially large clusters (= basins of attractions of the random walk).

The clustering transition

- Consider a random walk (try to flip a color at random, if still a valid coloring accept, if not try again) among colorings starting from a coloring chosen uniformly at random from all of them. Monitor the Hamming distance from the starting configuration.

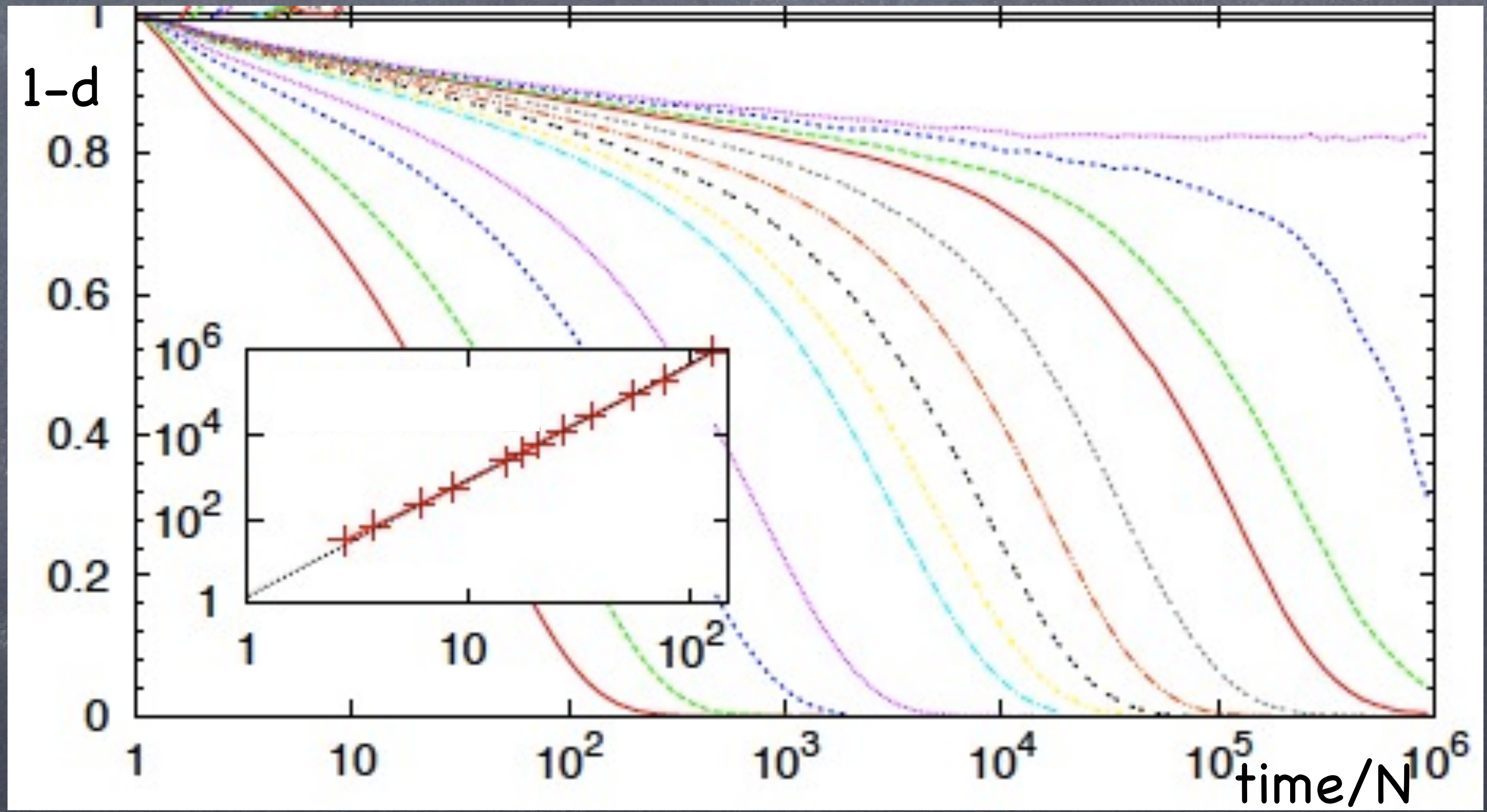
$$\tau = (c - c_d)^{-\frac{1}{2}}$$

$$c_d(q=3) = 4$$

$$c_d(4) = 8.35$$

$$c_d(5) = 12.84$$

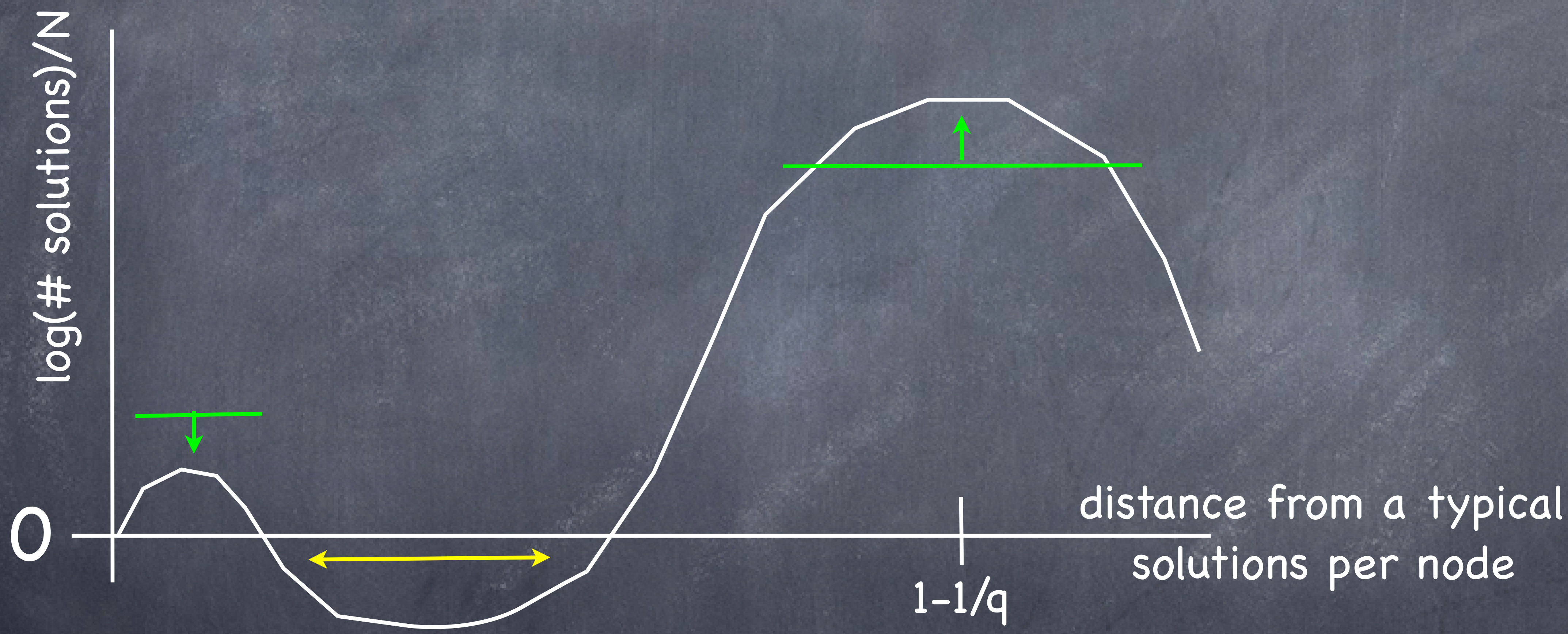
$$c_d(q \to \infty) = q \ln q$$



Close relation with reconstruction on trees: construct a configuration starting on the root, do the leaves contain any information abotu the root?

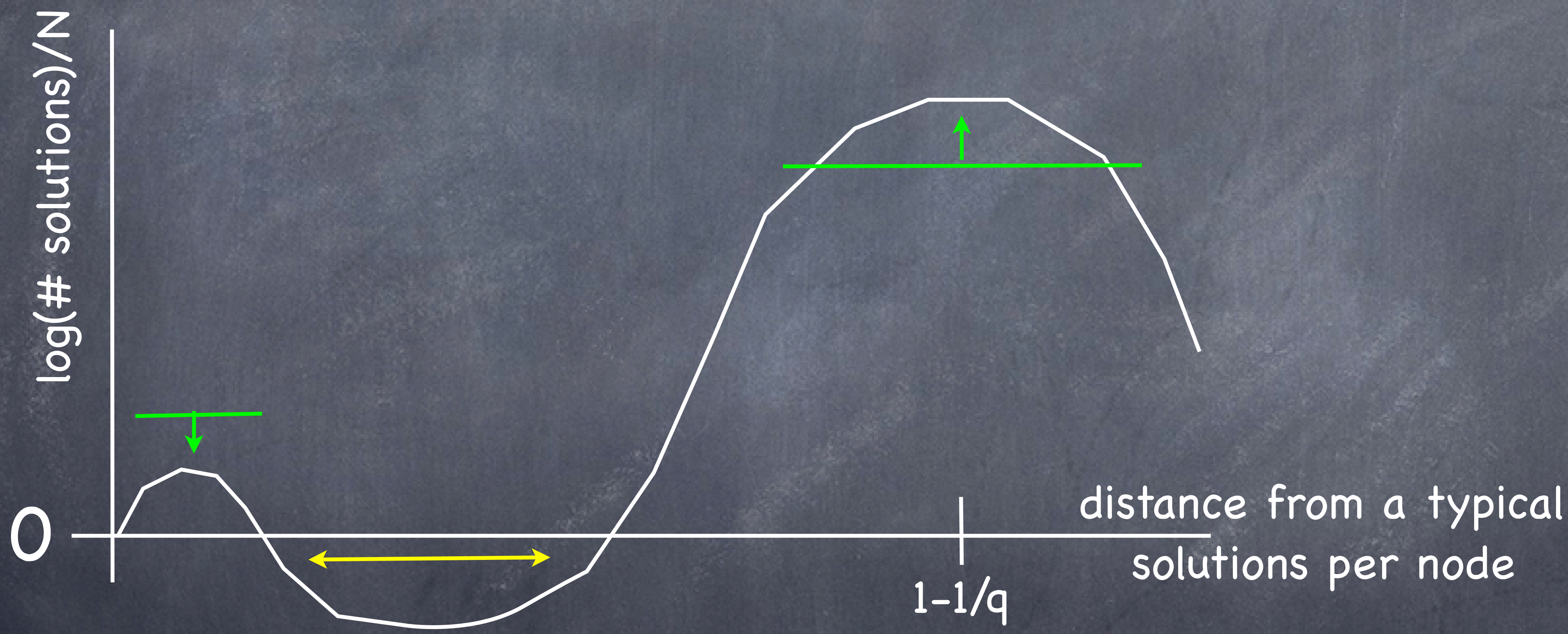
Clustering rigorously

- Large q in coloring, large K in K-SAT: Existence of exponentially many exponentially large geometrical clusters proven (Mora, Mezard, Zecchina'05; Achlioptas, Ricci-Tersenghi'05, Achlioptas, Coja-Oghlan'08).



Clustering rigorously

- Large q in coloring, large K in K-SAT: Existence of exponentially many exponentially large geometrical clusters proven (Mora, Mezard, Zecchina'05; Achlioptas, Ricci-Tersenghi'05, Achlioptas, Coja-Oghlan'08).



random walk clusters $\neq$ geometrical cluster

Condensation transition

$$s(c) \equiv \lim_{N\to\infty} \mathbb{E}(\ln(Z+1))$$

Conjecture: at the condensation transition s(c) is non-analytic (2nd derivative dis-continuous).

$$\ln \mathbb{E}(Z) = \mathbb{E}(\ln(Z+1)) + o(N) \qquad \text{for } c < c_c$$

$$\exists \epsilon > 0 : \ \ln \mathbb{E}(Z) > \mathbb{E}(\ln(Z+1)) + \epsilon N \ \ \text{for } c > c_c$$

Beyond condensation almost all solutions belong to a finite number of clusters.

$$c_d(q=3) = 4$$

$$c_d(4) = 8.35$$

$$c_d(5) = 12.84$$

$$c_d(q\to\infty) = q\ln q$$

$$c_c(q=3) = 4$$

$$c_c(4) = 8.46$$

$$c_c(5) = 13.23$$

$$c_c(q\to\infty) = 2q\ln q$$

In random Not-All-Equal SAT (2-coloring of k-hyper-graphs) the condensation transition exists.

(A. Coja-Oghlan, LZ, arxiv, 2011, SODA 2012)

$$r_{cond} = 2^{k-1}\ln 2 - \ln 2$$

Theorem 1.1 *There is a constant $k_0 \geq 3$ such that for all $k \geq k_0$ and $r < r_{cond}$ the random hypergraph $H_k(n,m)$ is 2-colorable w.h.p. and*

$$\ln Z \sim \ln \mathrm{E}[Z] \qquad w.h.p. \tag{3}$$

Theorem 1.3 *There exist a constant $k_0 \geq 3$ and a sequence $\varepsilon_k \to 0$ such that for any $k \geq k_0$ there are $\delta_k > 0, \zeta_k > 0$ such that the following two statements are true.*

1. *W.h.p. $H_k(n,m)$ is 2-colorable for all $r < r_{cond} + \varepsilon_k + \delta_k$.*
2. *For any density r with $r_{cond} + \varepsilon_k < r < r_{col}$ we have*

$$\ln Z < \ln \mathrm{E}[Z] - \zeta_k n \qquad w.h.p. \tag{4}$$

Conjecture 1.6 *There is $\varepsilon_k \to 0$ such that $r_{col} \sim 2^{k-1}\ln 2 - (\frac{\ln 2}{2} + \frac{1}{4}) + \varepsilon_k$.*

(Proof of 1.6 in: A. Coja-Oghlan, K. Panagiotou, arxiv 2011)

Two main ideas of the proof:

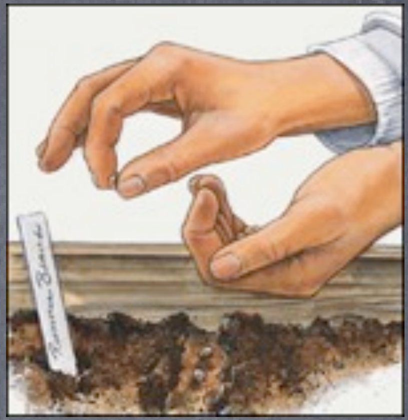
- The planted ensemble is very similar to the random ensemble (high probability properties of one are high probability properties of the other) before the condensation transition, and the planted ensemble is easier to analyze.

- In the large k regime clusters look like small "subcubes", only fraction $2^{-k/2}$ of variables not frozen, and their values are almost independent.

- In the proof use subcubes to bound the expected size of the planted cluster and look when this becomes larger than the total expected number of solutions.

Planted Coloring

(Krzakala, Zdeborova'09)



Generating planted instances: Fix a configuration. Choose constraints randomly such that all (but fraction p) are satisfied by the fixed configuration.

Conjecture: Planted coloring on average easy for

$$c > (q-1)^2$$

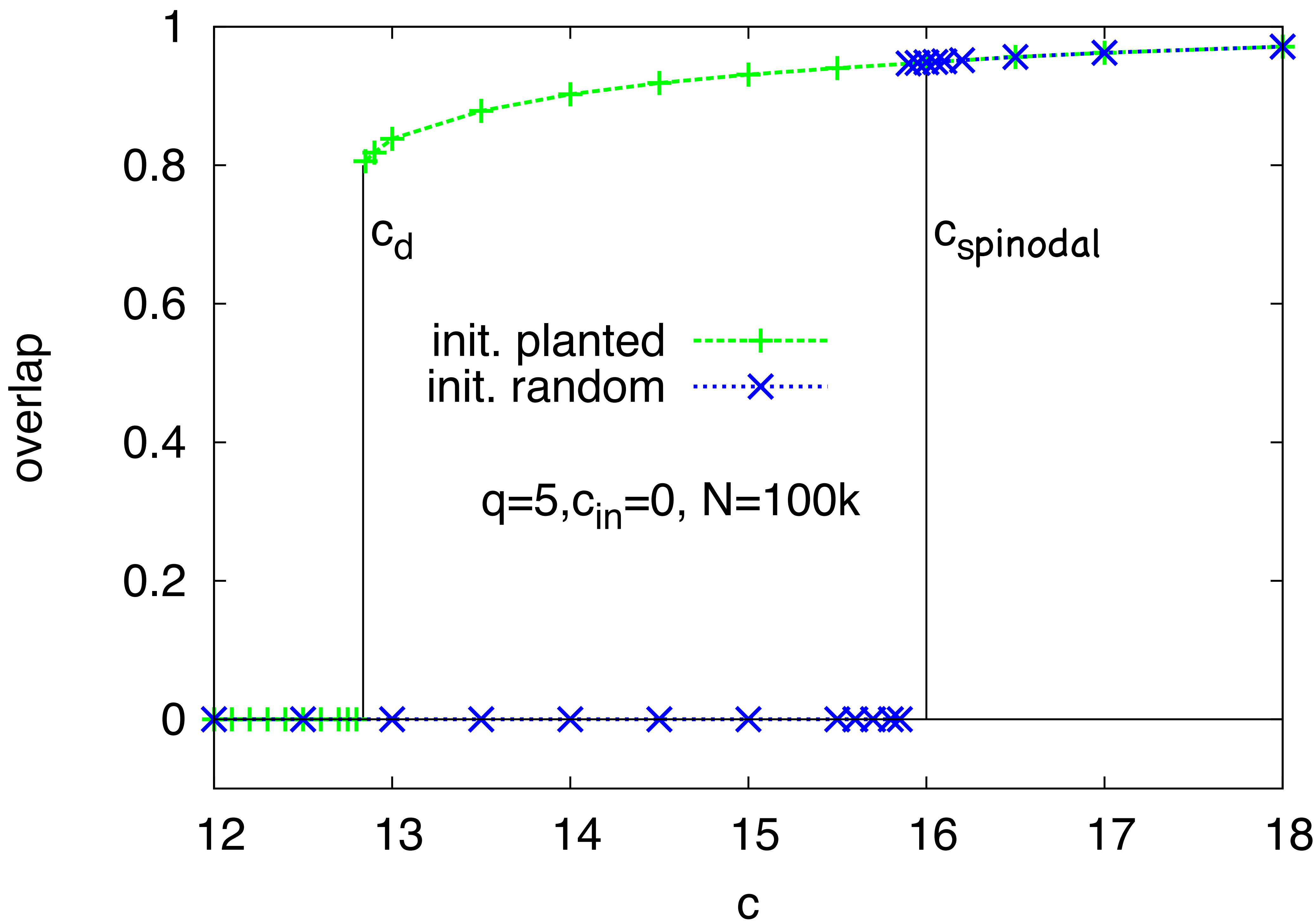
Using Belief Propagation

$$\psi_{s_i}^{i\to j} = \frac{1}{Z^{i\to j}} \prod_{k\in\partial i\setminus j} (1-\psi_{s_i}^{k\to i})$$

Rigorous bound: for q>q_0 there is a constant s.t. planted coloring easy on average for $c > \mathrm{const}.q^2$

(Coja-Oghlan, Mossel, Vilenchik, 2009)

Planted 5-coloring



Remind: $c_c = 13.23, c_s = 13.67$

Algorithmic consequences

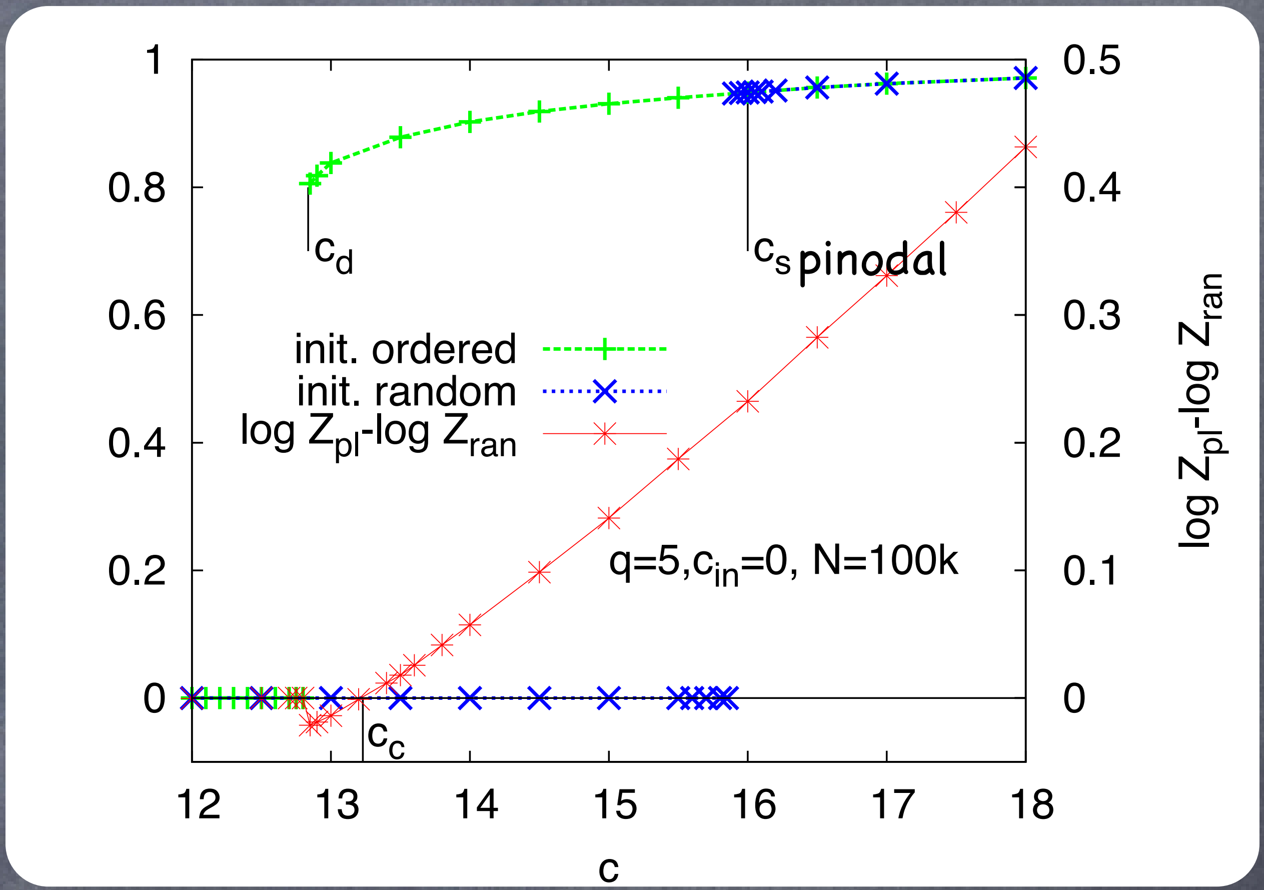
$c < c_c$

inference impossible

planted = random

proof q=2

(Mossel, Neeman, Sly'11)



$c_c < c < c_{\text{spinodal}}$

inference possible but hard

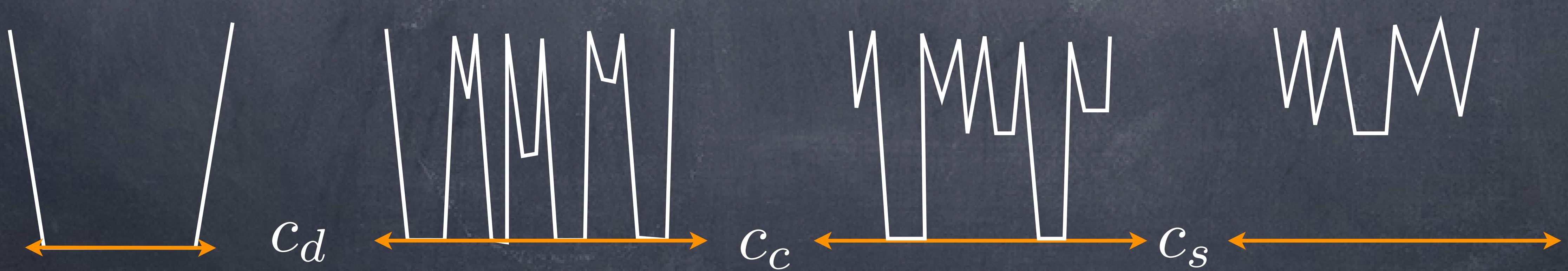
$c_{\text{spinodal}} < c$

inference easy

Does clustering or condensation make the coloring hard?

Hard to sample $\neq$ Hard to solve

- Algorithms (belief propagation or stochastic local search) find solutions even in the glassy phase - empirical evidence everywhere.
- In 3-coloring condensation c=4, algorithms provably work up to 4.03 (Achlioptas, Moore'03)



Freezing of variables

- Def.: Variable is frozen in a cluster (= set of colorings) if it takes the same color in the whole set. Cluster is frozen if is has a finite fraction of frozen variables. Rigidity transition - typical solution belongs to a frozen cluster.
- Coloring belongs to a frozen cluster iff ($N \to \infty$) it has a non-trivial whitening. Whitening: If a node has a neighbor that does not have the other q-1 colors on its other neighbors, turn this node white and iterate.

q	c_d	c_r	c_c	c_s
3	4	4.66(1)	4	4.687(2)
4	8.353(3)	8.83(2)	8.46(1)	8.901(2)
5	12.837(3)	13.55(2)	13.23(1)	13.669(2)
6	17.645(5)	18.68(2)	18.44(1)	18.880(2)
7	22.705(5)	24.16(2)	24.01(1)	24.455(5)
8	27.95(5)	29.93(3)	29.90(1)	30.335(5)
9	33.45(5)	35.658	36.08(5)	36.490(5)
10	39.0(1)	41.508	42.50(5)	42.93(1)

Recent proofs about freezing: (Molloy'12)

$$c_c \simeq 2q \ln q - \ln q - 2 \ln 2$$

= rigorous lower bound Coja-Oghlan, Vilenchik 2013.

$$c_s \simeq 2q \ln q - \ln q - 1$$

= rigorous upper bound Coja-Oghlan 2013.

$$c_r \simeq q(\ln q + \ln \ln q + 1)$$

= rigorous proof Molloy 2012.

$$c_{\text{spinodal}} = (q-1)^2$$

bounded by Coja-Oghlan, Mossel, Vilenchik, 2009

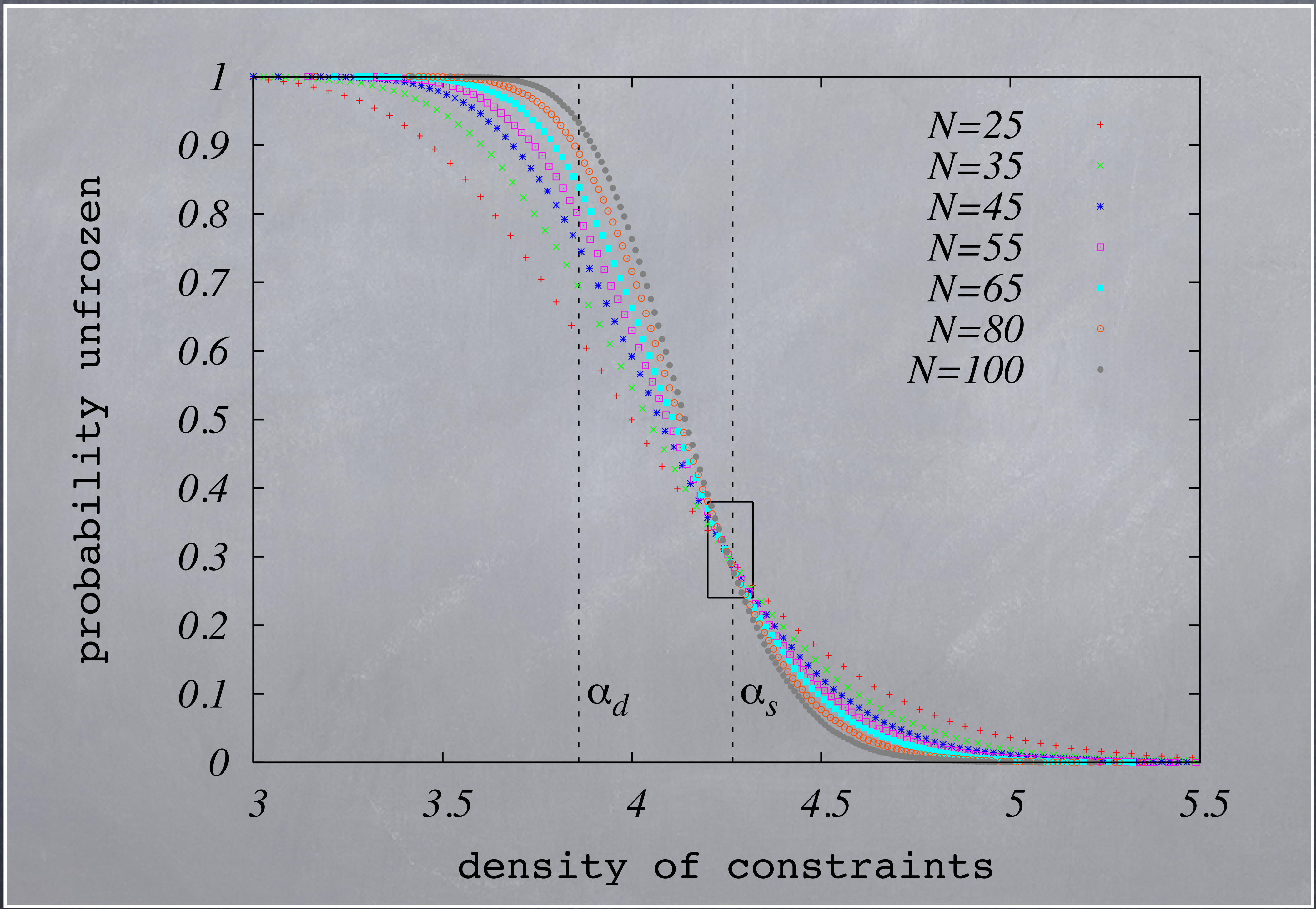
q	c_d	c_r	c_c	c_s
3	4	4.66(1)	4	4.687(2)
4	8.353(3)	8.83(2)	8.46(1)	8.901(2)
5	12.837(3)	13.55(2)	13.23(1)	13.669(2)
6	17.645(5)	18.68(2)	18.44(1)	18.880(2)
7	22.705(5)	24.16(2)	24.01(1)	24.455(5)
8	27.95(5)	29.93(3)	29.90(1)	30.335(5)
9	33.45(5)	35.658	36.08(5)	36.490(5)
10	39.0(1)	41.508	42.50(5)	42.93(1)

Conjecture (Zdeborova, Krzakala 2007): Freezing of variables responsible for the onset of algorithmic hardness for a large class of algorithms.

- Algorithms never find frozen solutions. Empirically but also predicted by the only known explanations of why algorithms work in clustered region (state following, Krzakala, Zdeborova`2010).

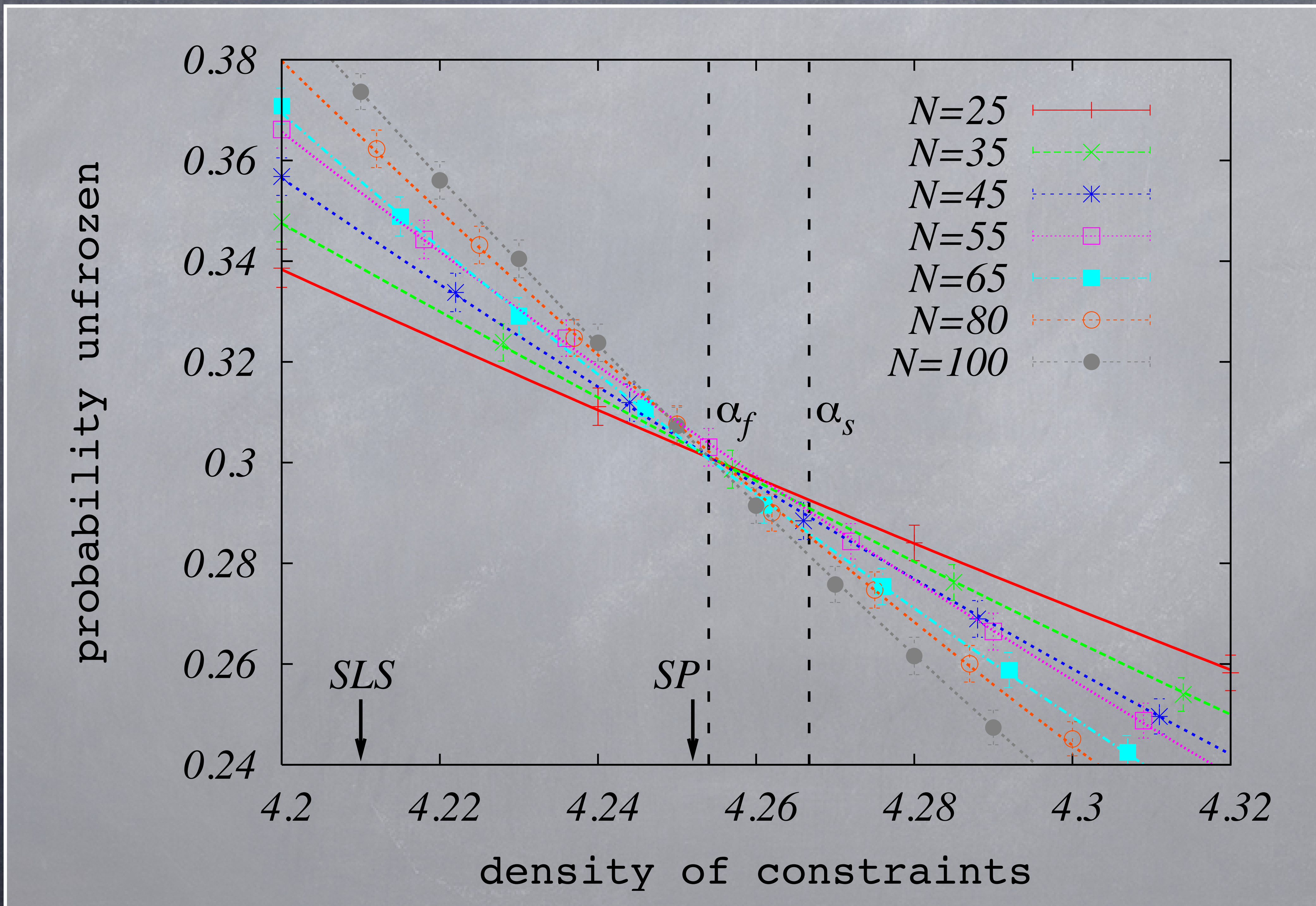
Probability that an unfrozen solution exists in 3-SAT

(Ardelius, Zdeborova'08)



Zoom at the freezing transition

(Ardelius, Zdeborova'08)



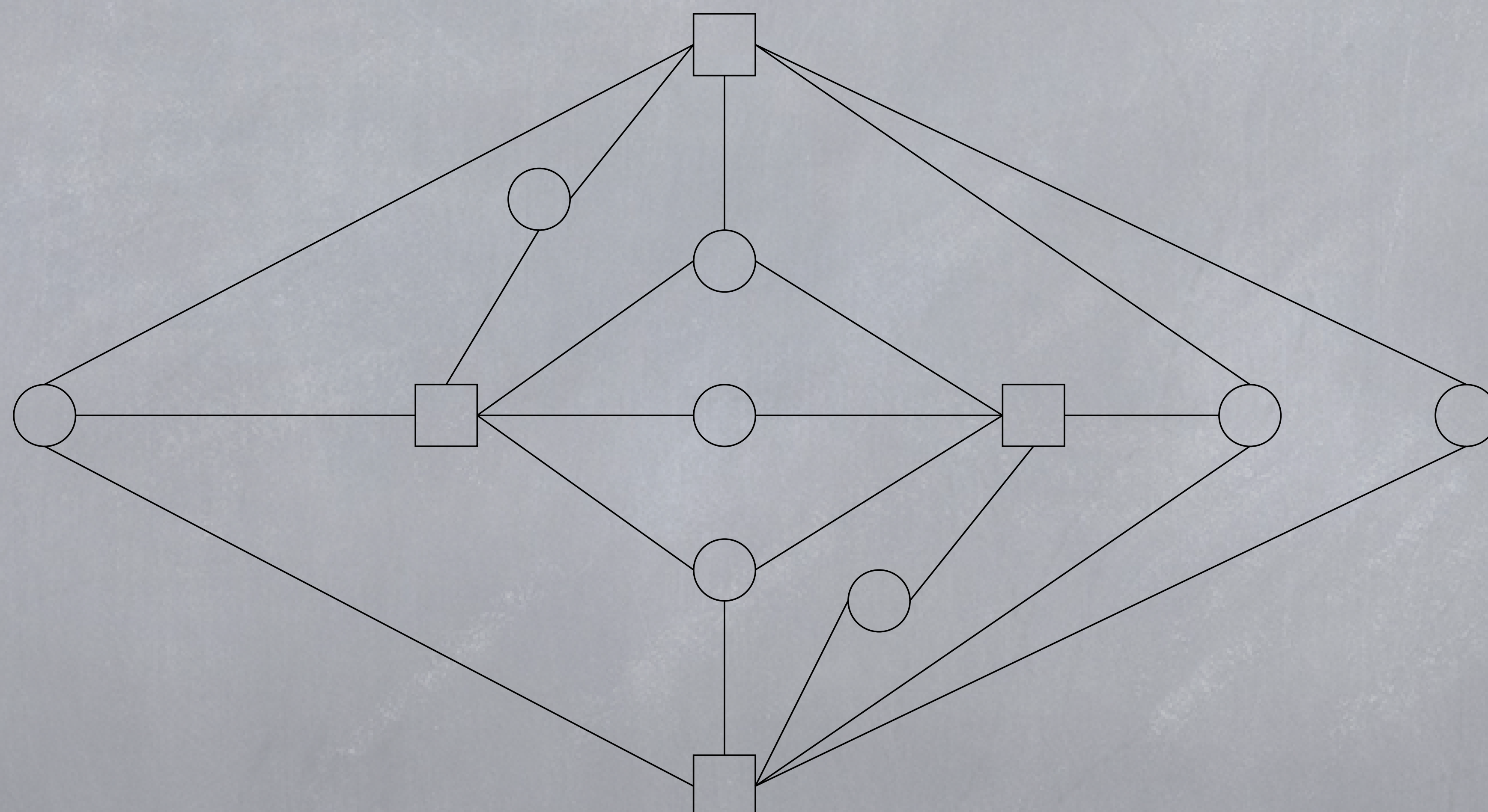
Locked CSP

(Zdeborova, Mezard'08)

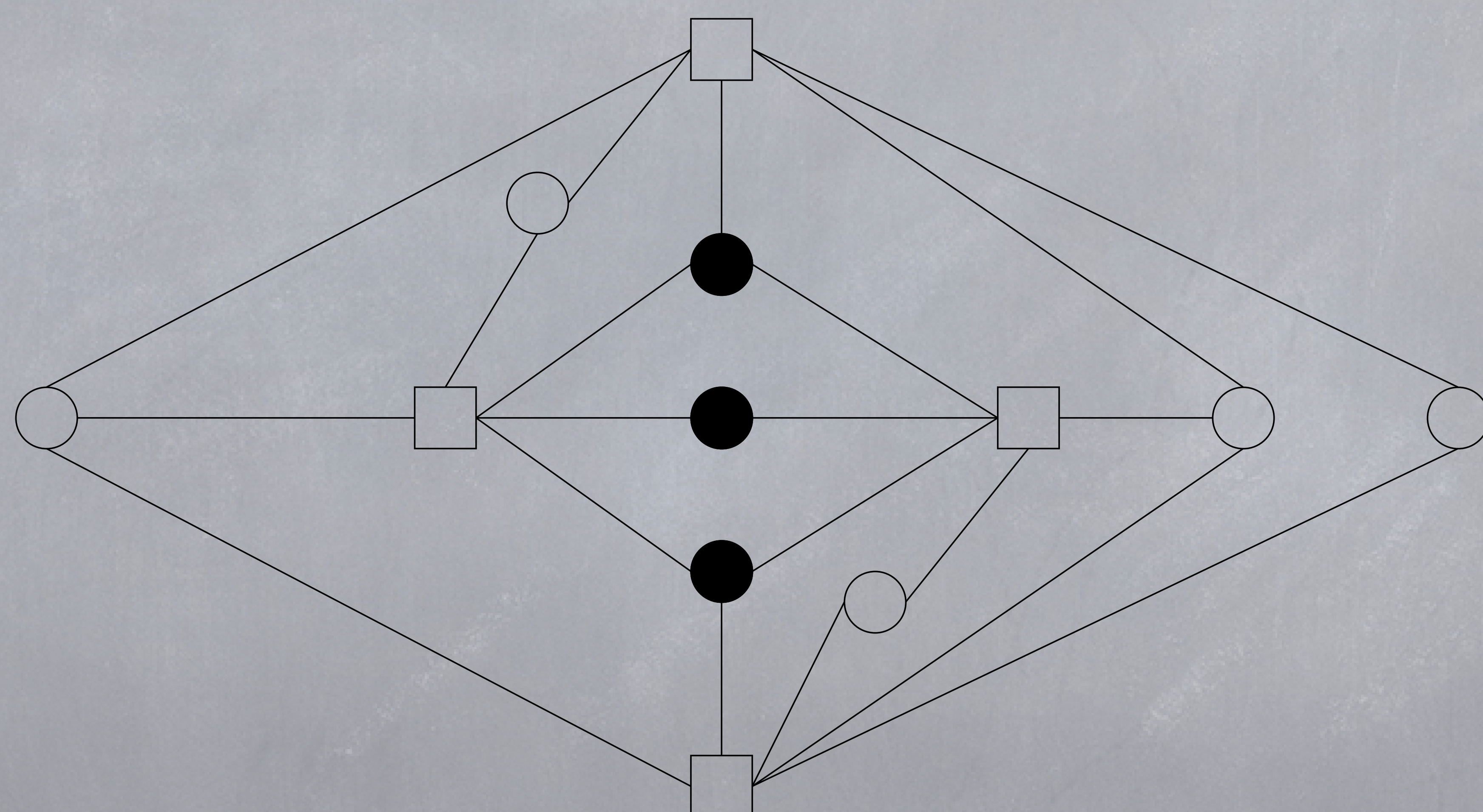
- Definition: A closed loop of variables has to be flipped to go from one solution to another.
- Examples: XOR-SAT on the core, 1-in-K SAT without leaves.

In locked CSP clusters are point like, always frozen.

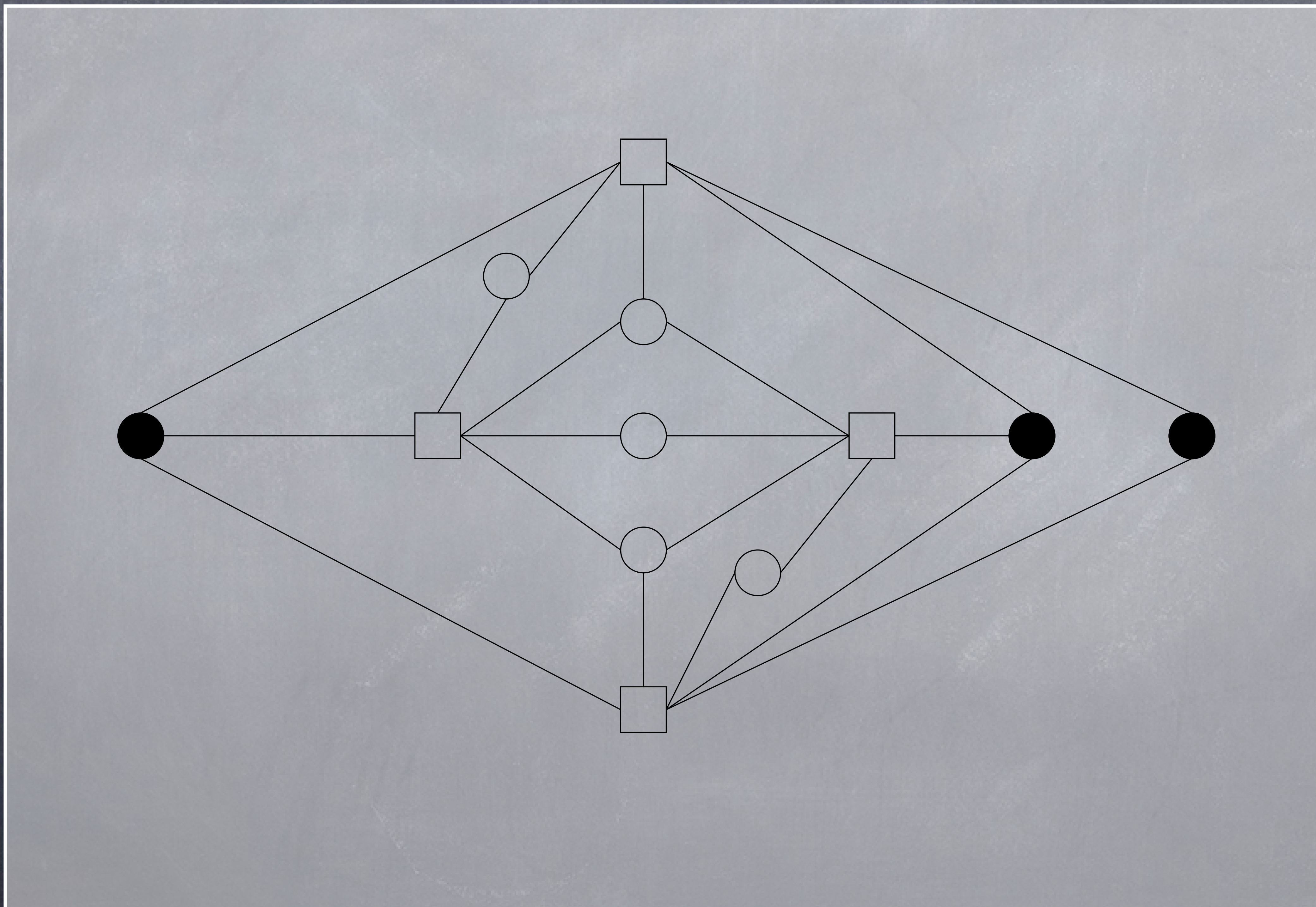
Example: 1-or-3-in-5 SAT



Example: 1-or-3-in-5 SAT

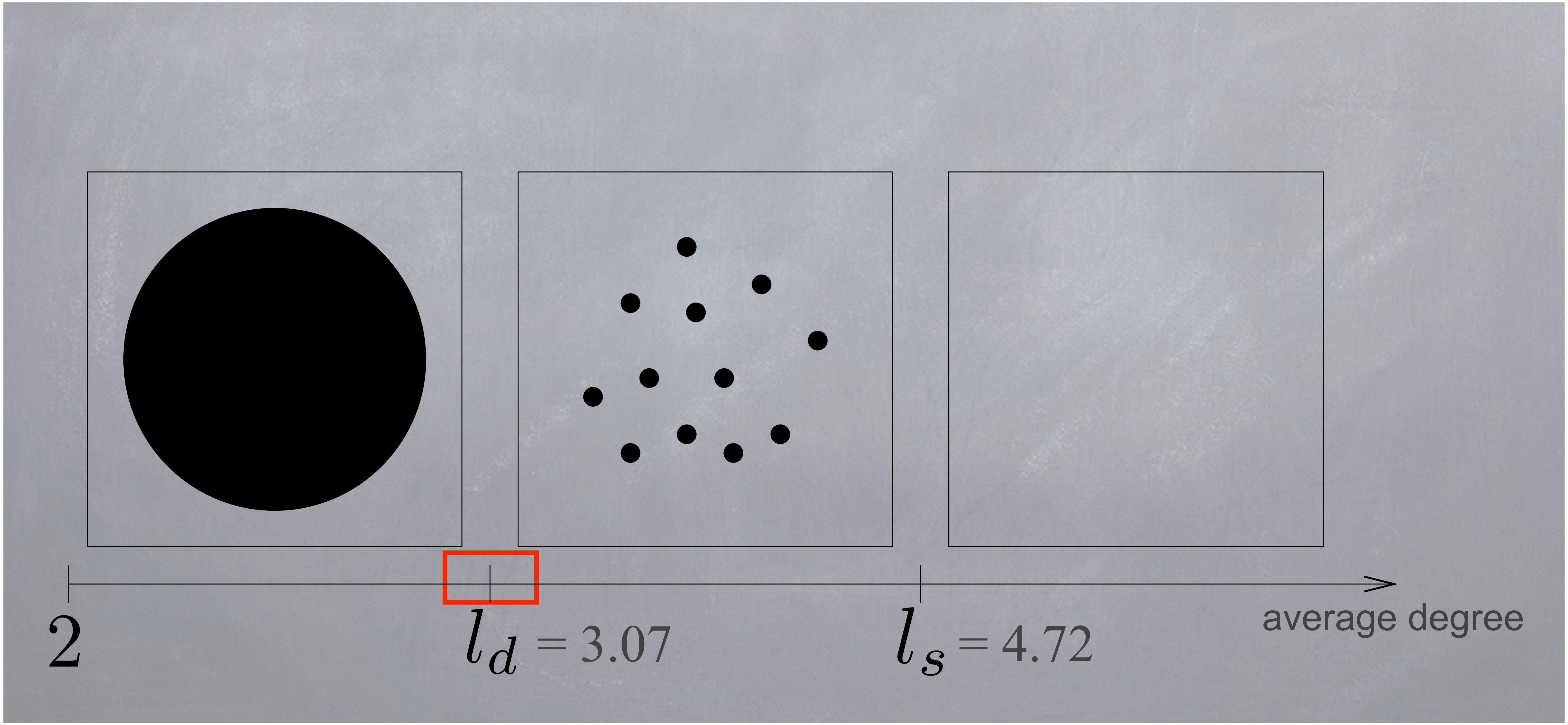


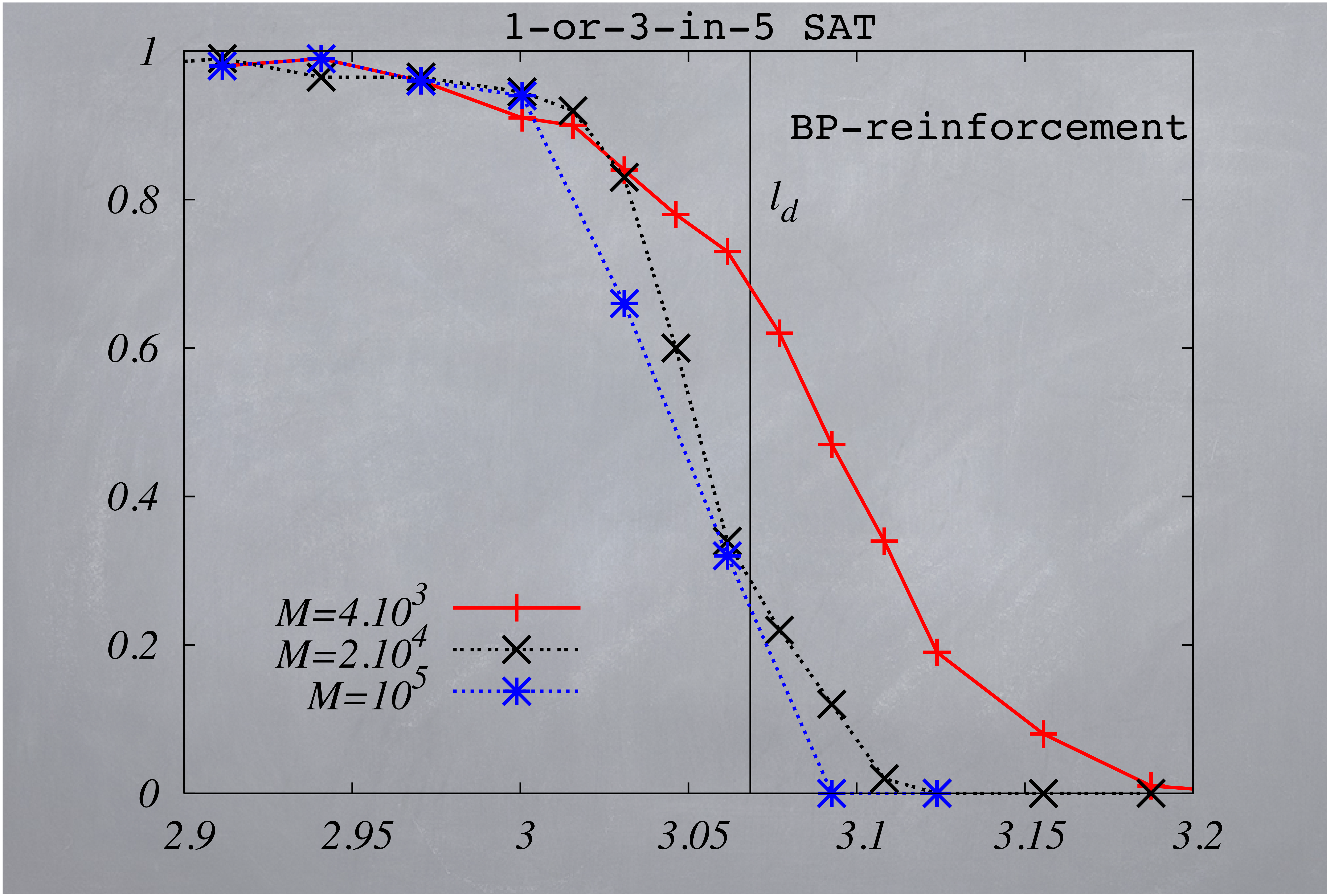
Example: 1-or-3-in-5 SAT



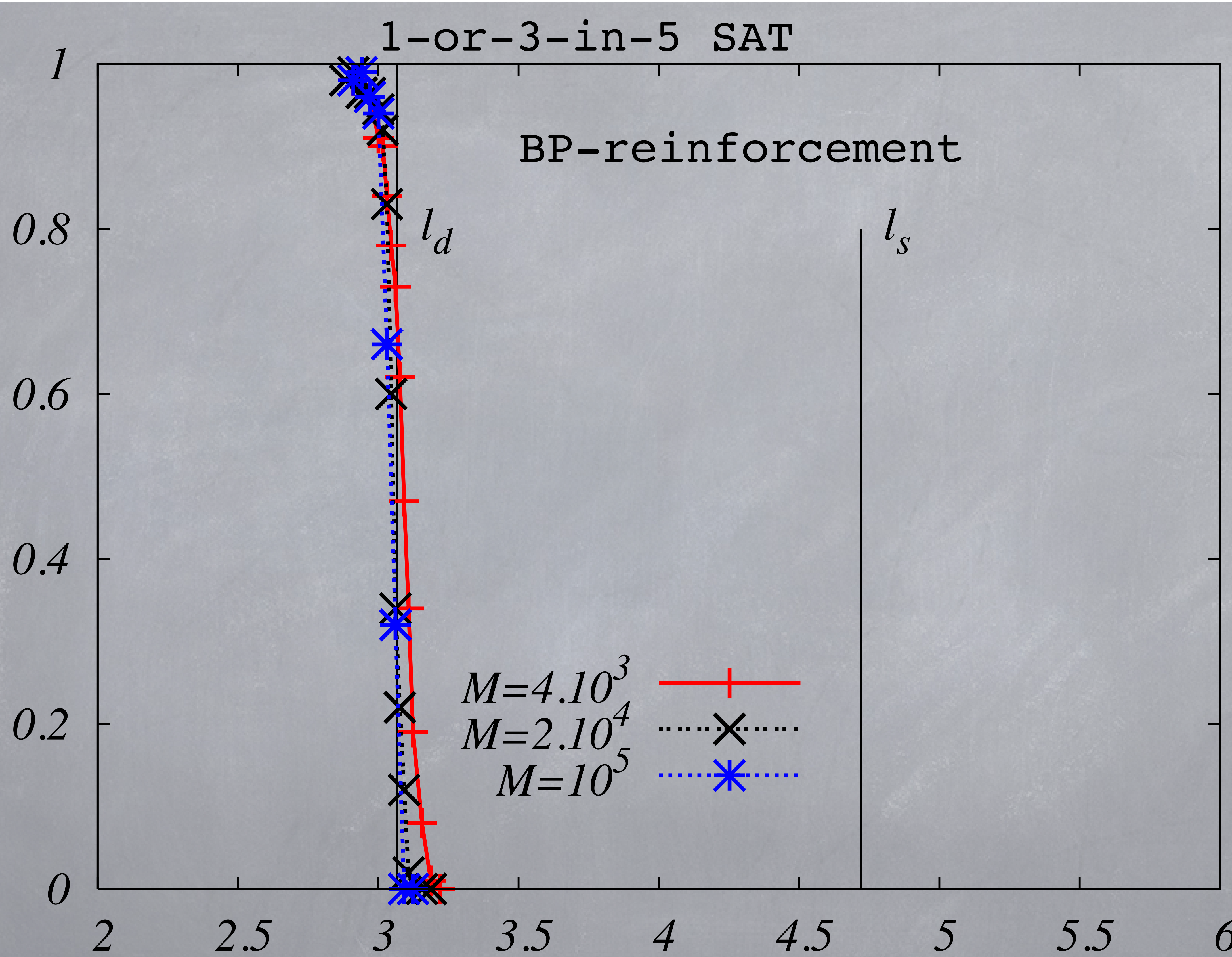
Locked CSP

(1) clustering = freezing
(2) For symmetric ones SAT threshold computed from 2nd moment

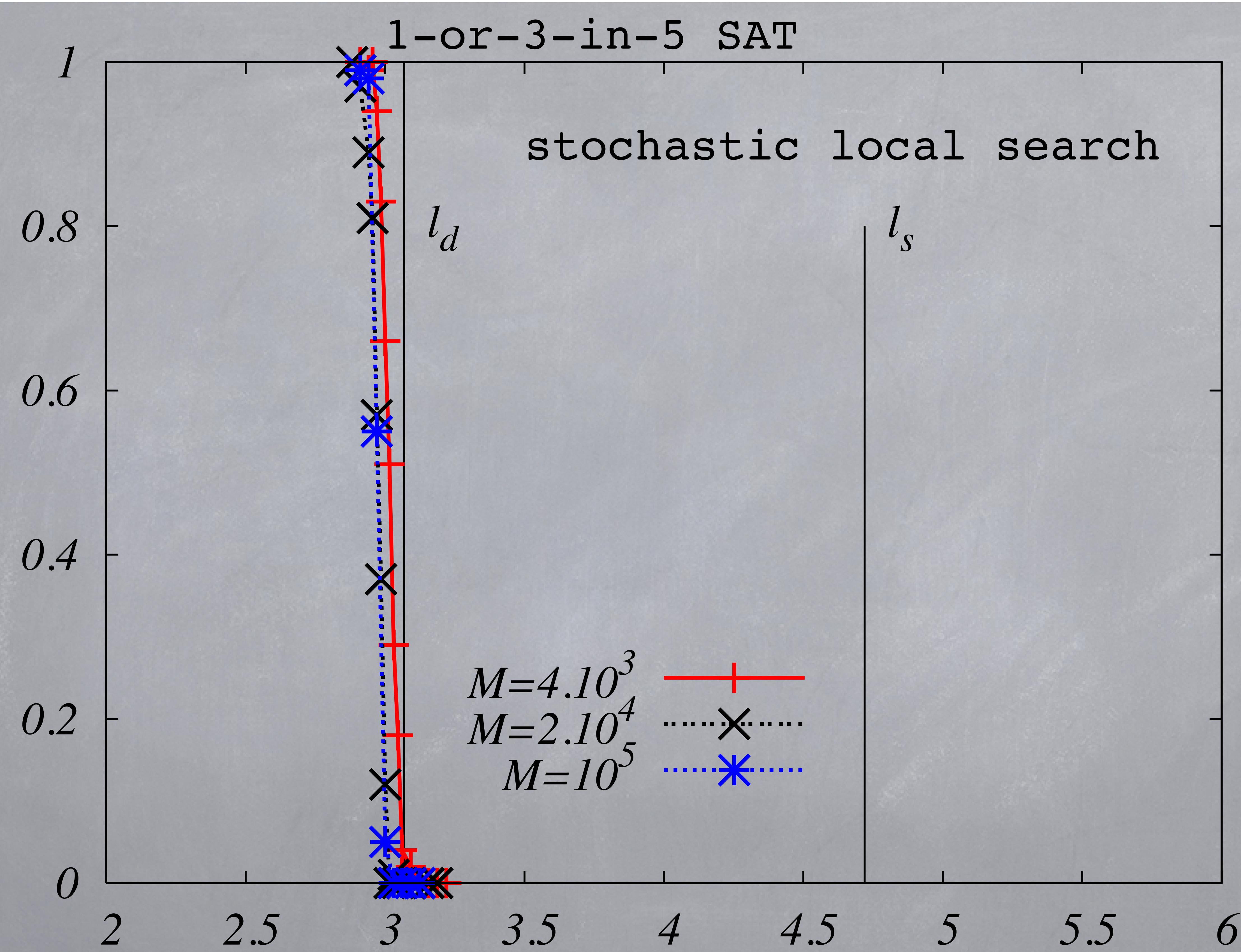




Tuesday, June 18, 13



1-or-3-in-5 SAT
BP-reinforcement
l_d
l_s
$M=4.10^3$
$M=2.10^4$
$M=10^5$
1
0.8
0.6
0.4
0.2
0
2
2.5
3
3.5
4
4.5
5
5.5
6


1-or-3-in-5 SAT
stochastic local search
l_d
l_s
$M=4.10^3$
$M=2.10^4$
$M=10^5$
1
0.8
0.6
0.4
0.2
0
2
2.5
3
3.5
4
4.5
5
5.5
6

- F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, L. Zdeborová: *Gibbs States and the Set of Solutions of Random Constraint Satisfaction Problems* Proc. Natl. Acad. Sci. 104, 10318 (2007). preprint: cond-mat/0612365 .

- L. Zdeborová, F. Krzakala: *Phase transition in the Coloring of Random Graphs* Phys. Rev. E 76, 031131 (2007). preprint: arxiv:0704.1269 .

- L. Zdeborová, M. Mézard: *Locked constraint satisfaction problems* Phys. Rev. Lett. 101, 078702 (2008). preprint: arXiv:0803.2955 .

- PhD thesis: L. Zdeborová: *Statistical Physics of Hard Optimization Problems* Acta Physica Slovaca 59, No.3, 169-303 (2009) preprint: arXiv:0806.4112 .

- F. Krzakala, L. Zdeborová: *Hiding Quiet Solutions in Random Constraint Satisfaction Problems* Phys. Rev. Lett. 102, 238701 (2009). preprint: arXiv:0901.2130 .

- A. Coja-Oghlan, L. Zdeborová: *The condensation transition in random hypergraph 2-coloring,* SODA 2012, preprint: arXiv:1107.2341v1.

Thank you for your attention!

Planted COL

(Krzakala, Zdeborova'09)

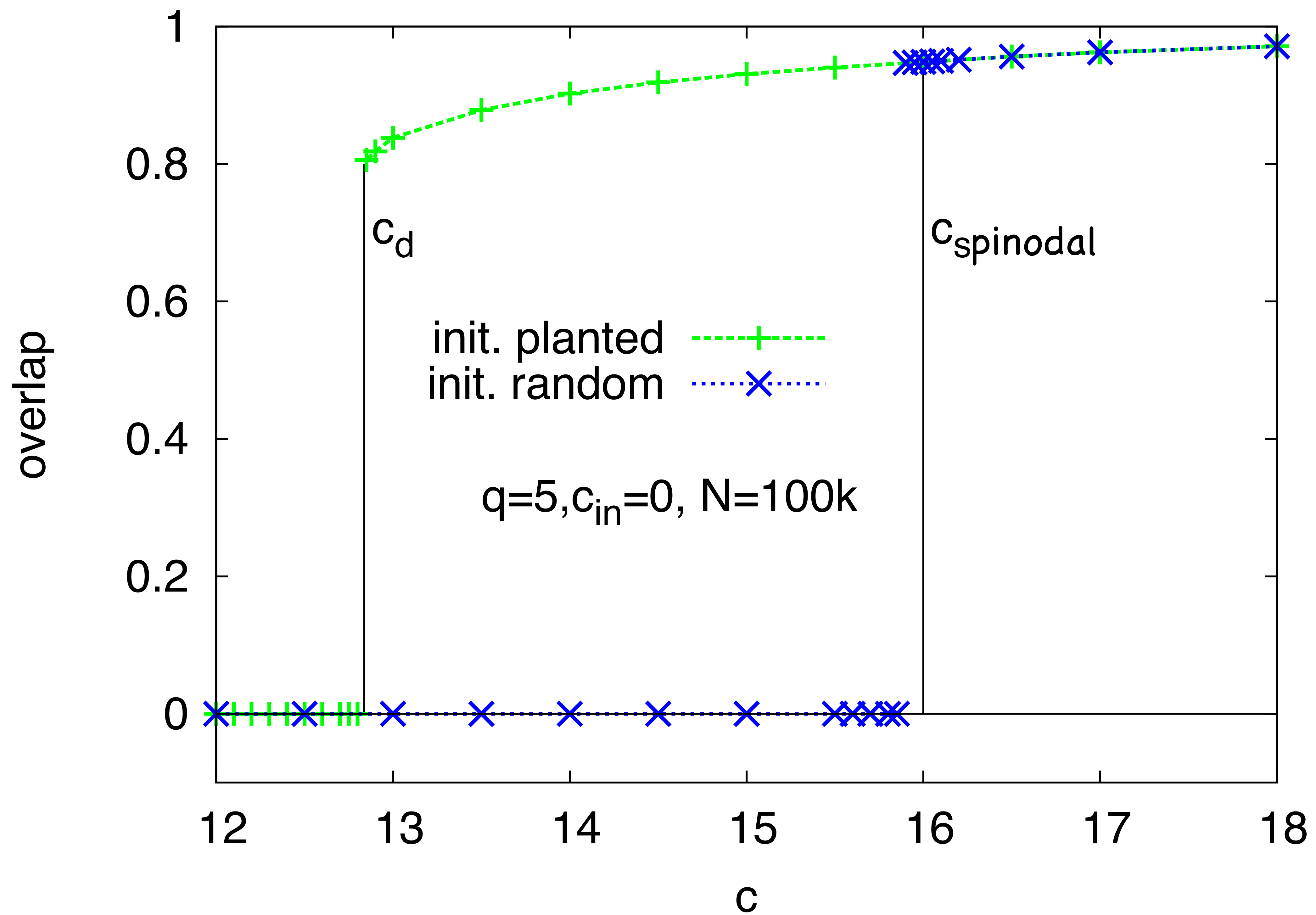
Generating planted instances:

- Fix a configuration.
- Choose constraints randomly such that all (but fraction p) are satisfied by the fixed configuration.

Planted 3-COL easy - not a generic situation.

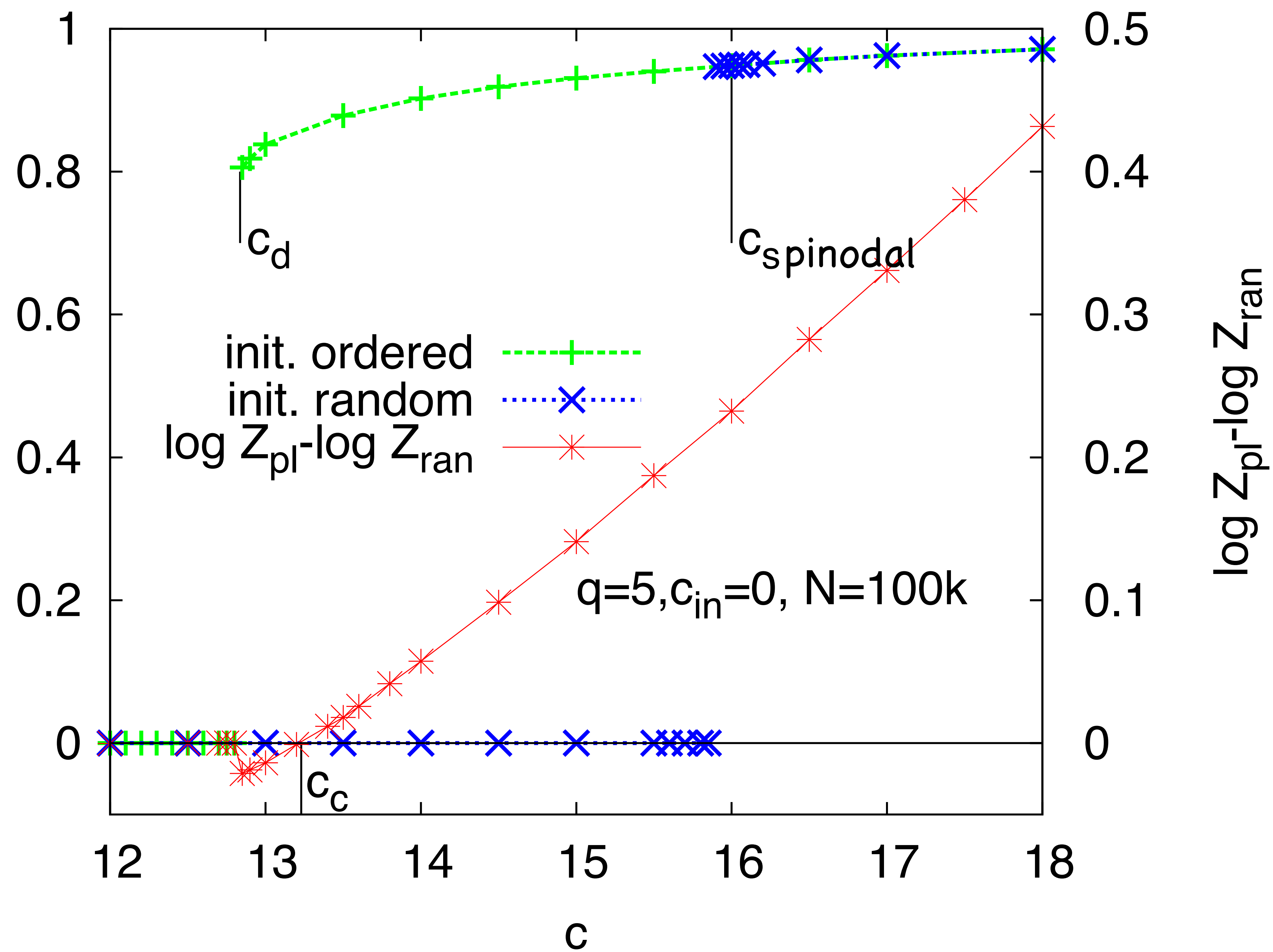
Properties (hardness) of planted CSP generalize to many inference problems (LDPC, community detection, compressed sensing, etc.)

Planted 5-coloring



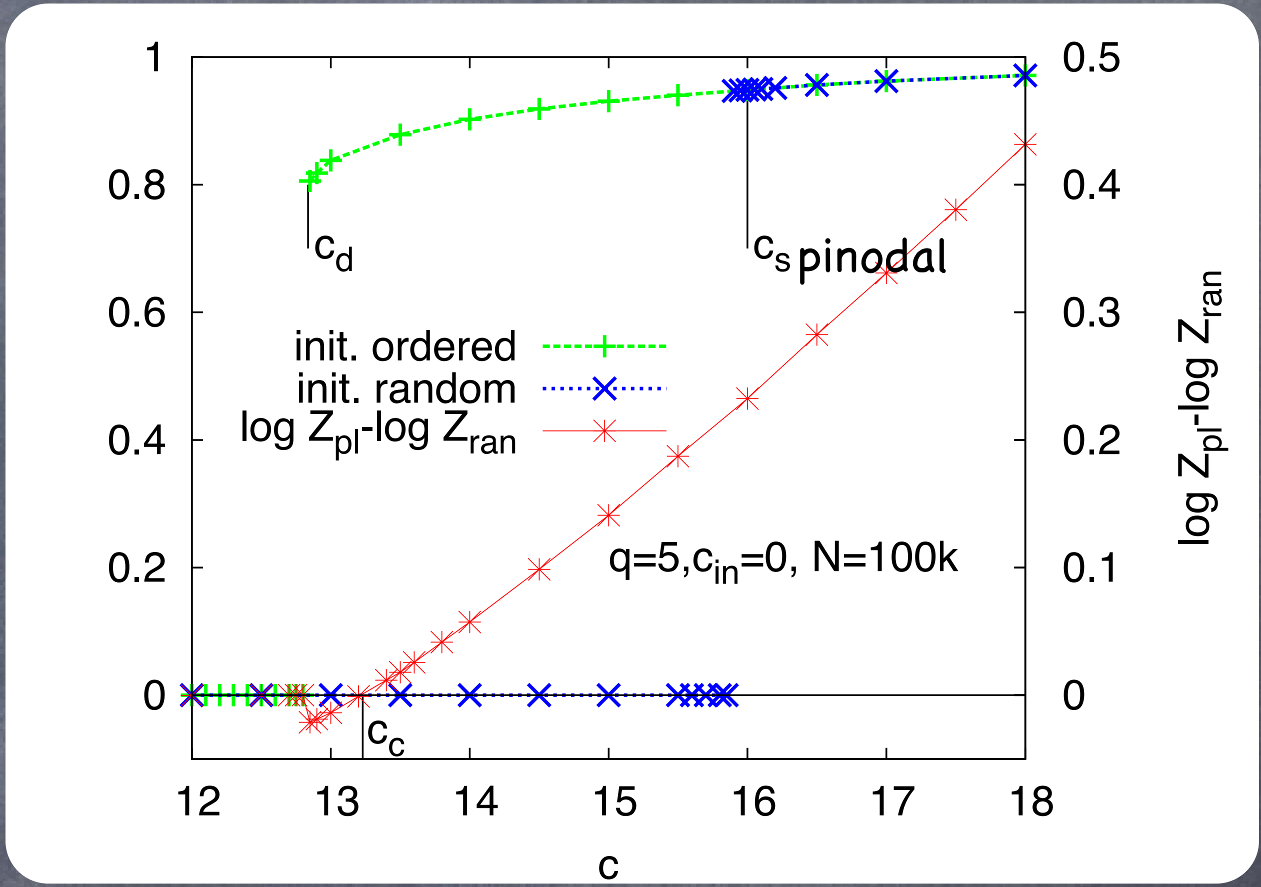
Remind: $c_c = 13.23, c_s = 13.67$

Tuesday, June 18, 13



init. ordered
init. random
log Z_pl-log Z_ran
c_d
c_spinodal
c_c
q=5,c_in=0, N=100k
c
log Z_pl-log Z_ran
0
0.2
0.4
0.6
0.8
1
0.1
0.2
0.3
0.4
0.5
12
13
14
15
16
17
18

Algorithmic consequences



$c < c_c$
inference impossible
planted = random

$c_c < c < c_{\mathrm{spinodal}}$
inference possible but hard

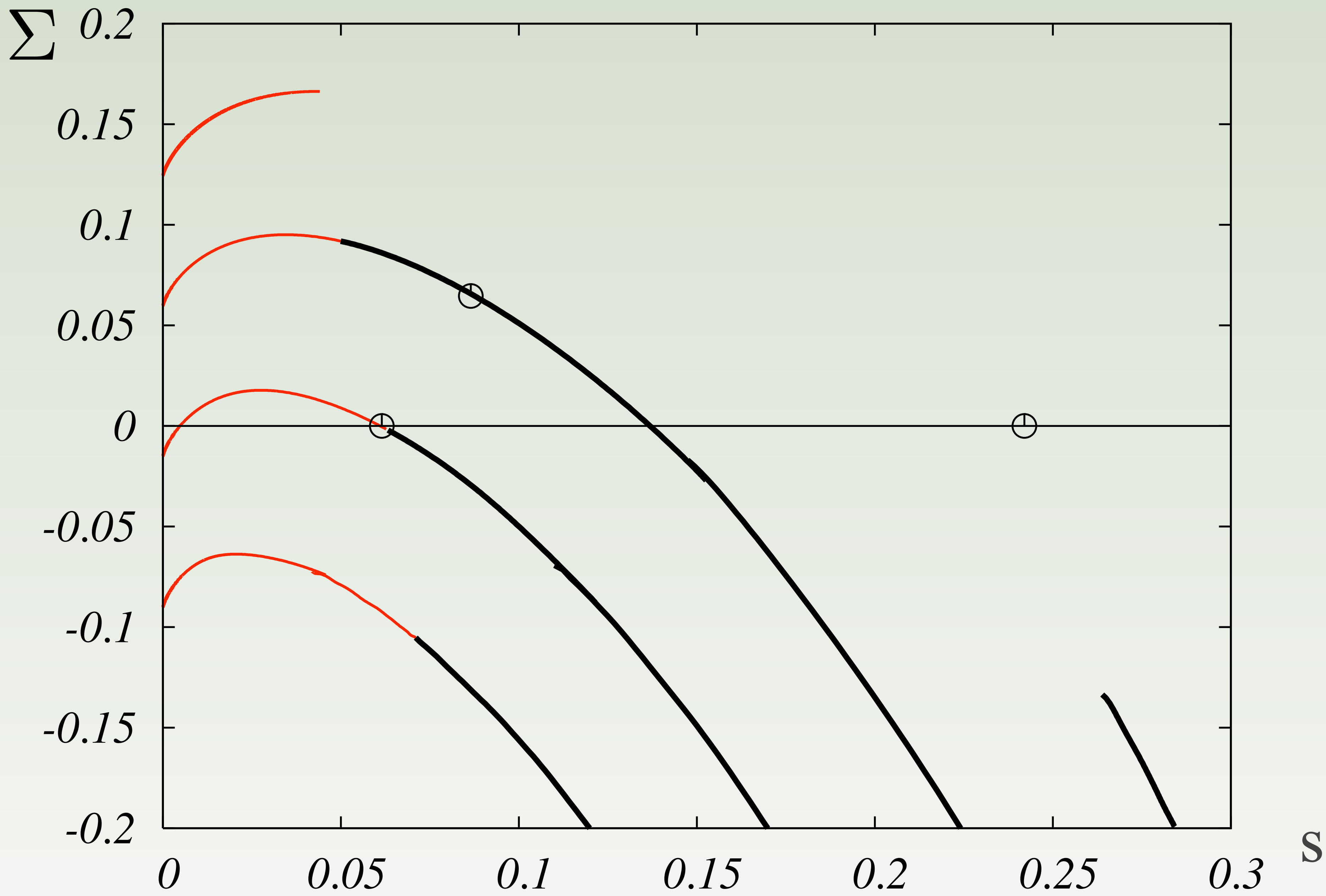
$c_{\mathrm{spinodal}} < c$
inference easy

Planting: a proof technique

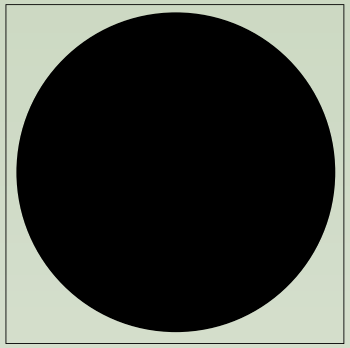
- Since (a) planted ensemble = random ensemble before the condensation transition, and (b) planted ensemble is easier to analyze. => One can prove clustering, freezing and condensation (Coja-Oghlan, et al 2010-ongoing)

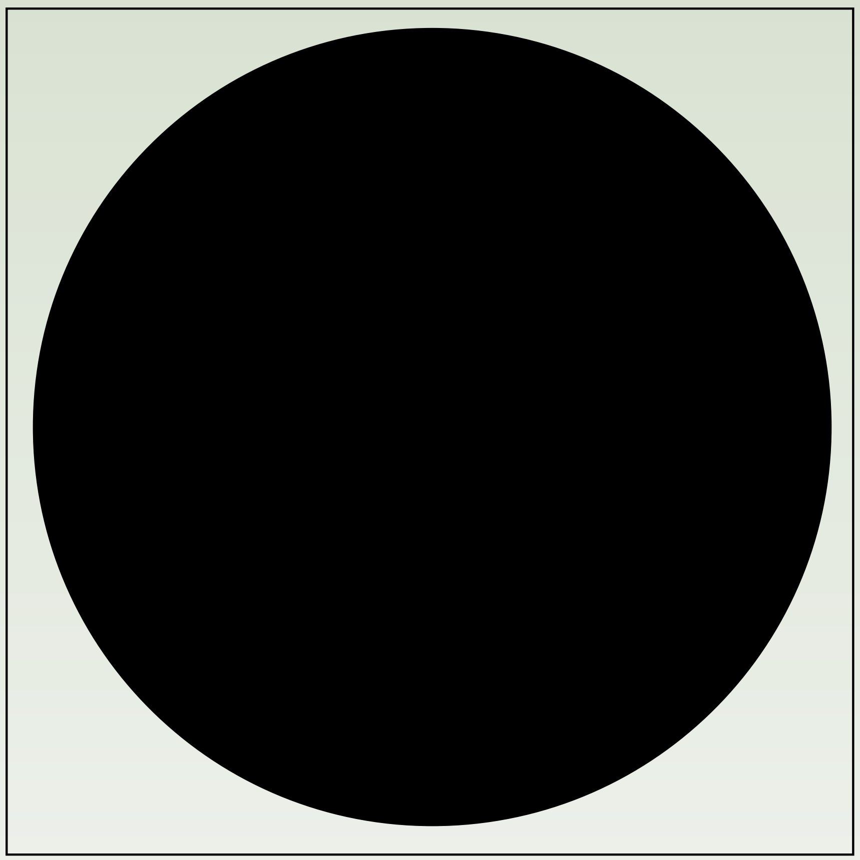
Learning from $\Sigma(s)$

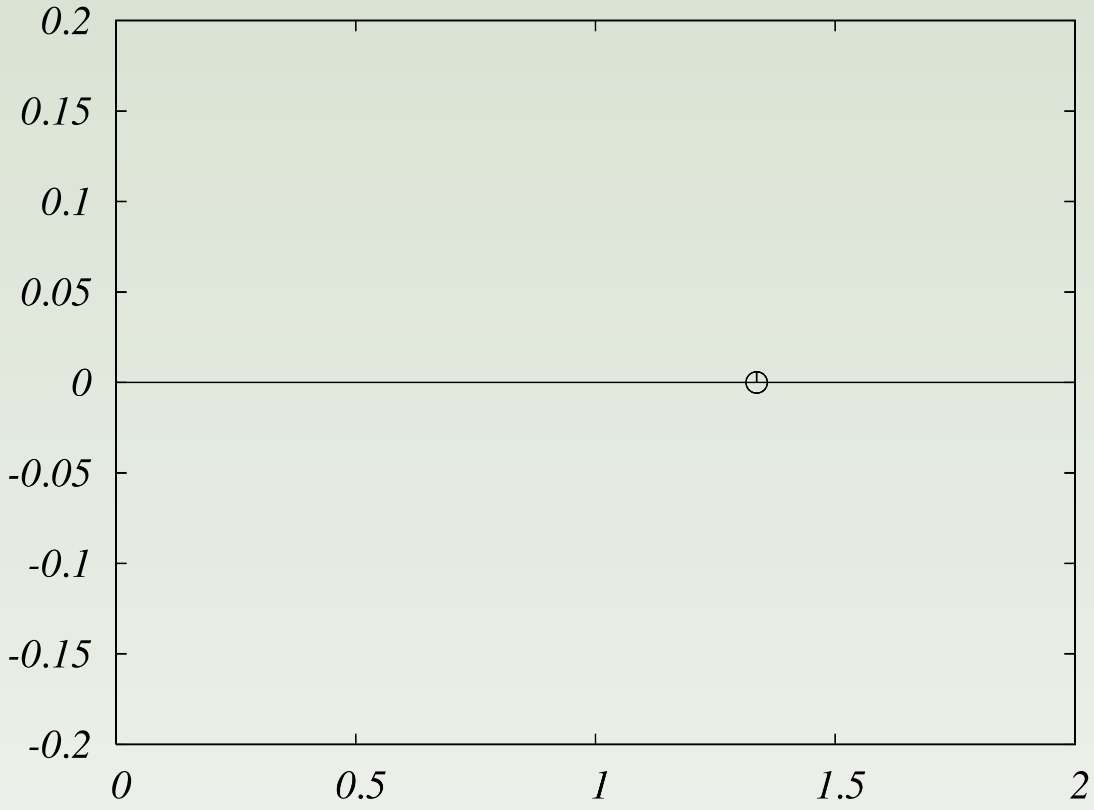
Example of 6-coloring, connectivities 17, 18, 19, 20 (from top).



Tuesday, June 18, 13

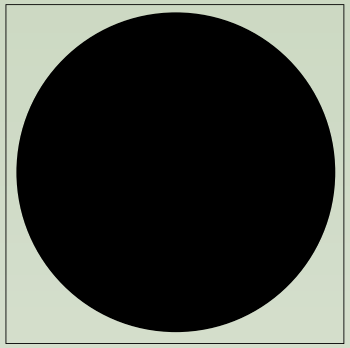


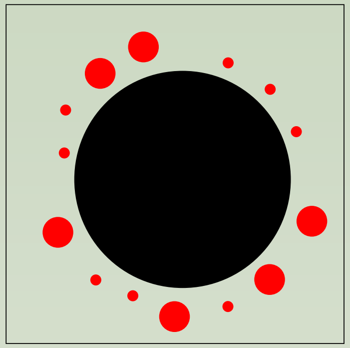


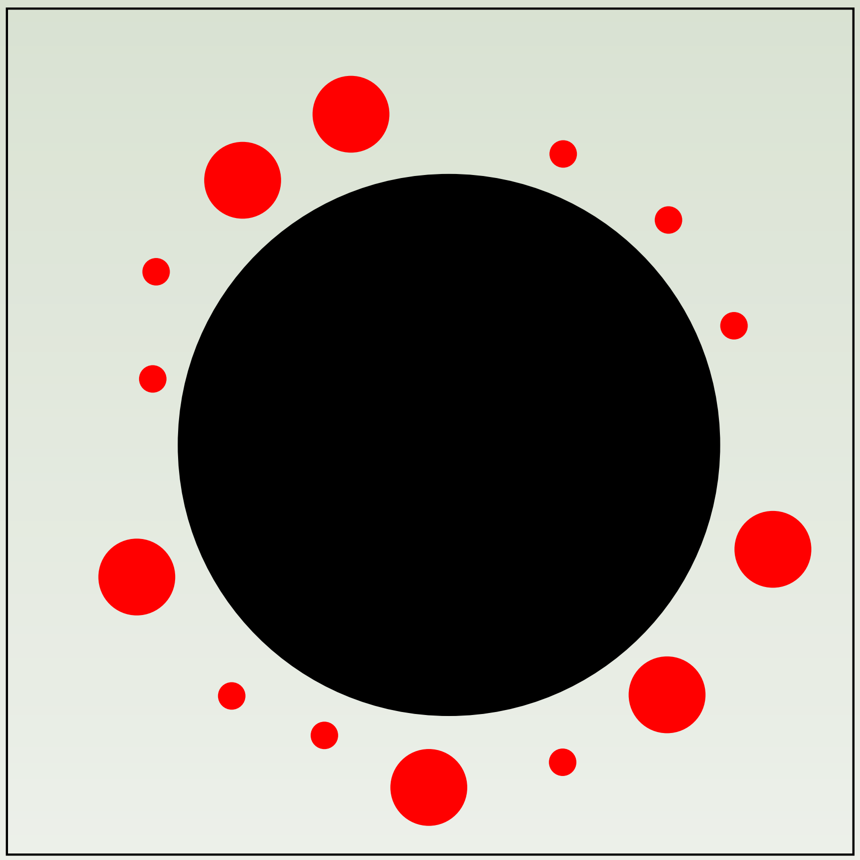


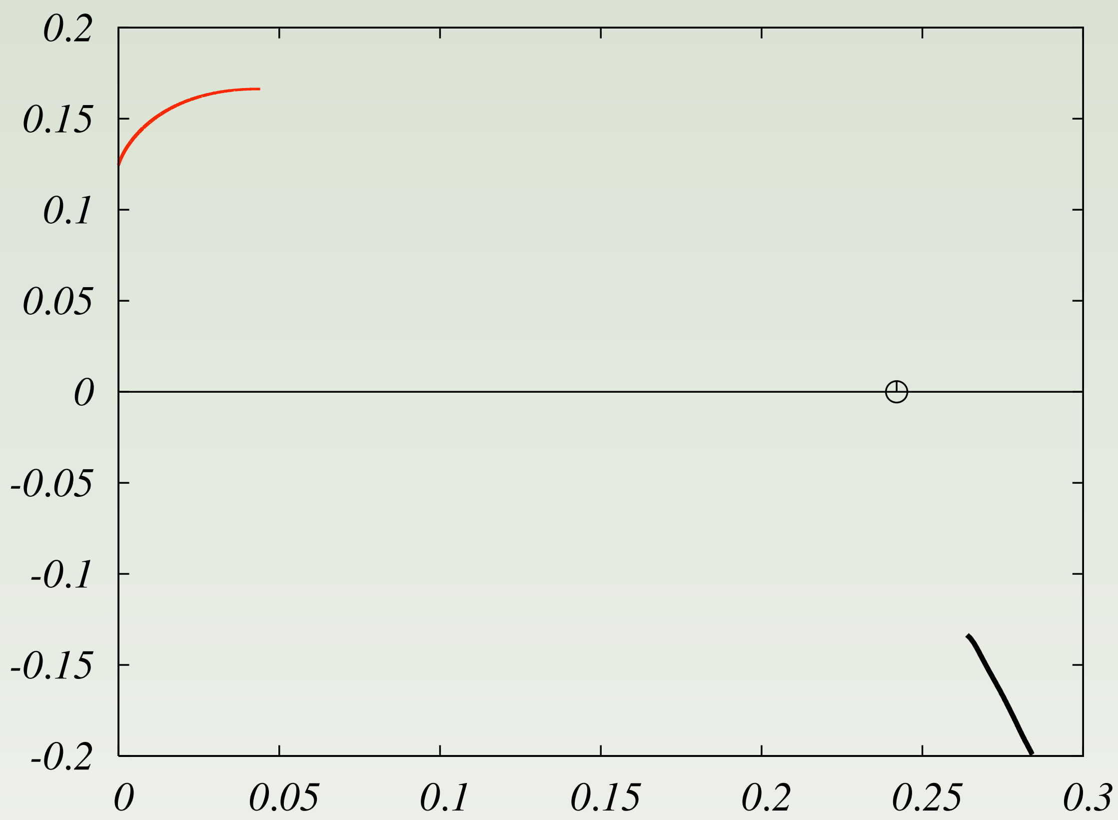
6 coloring of regular random graph

very low connectivity



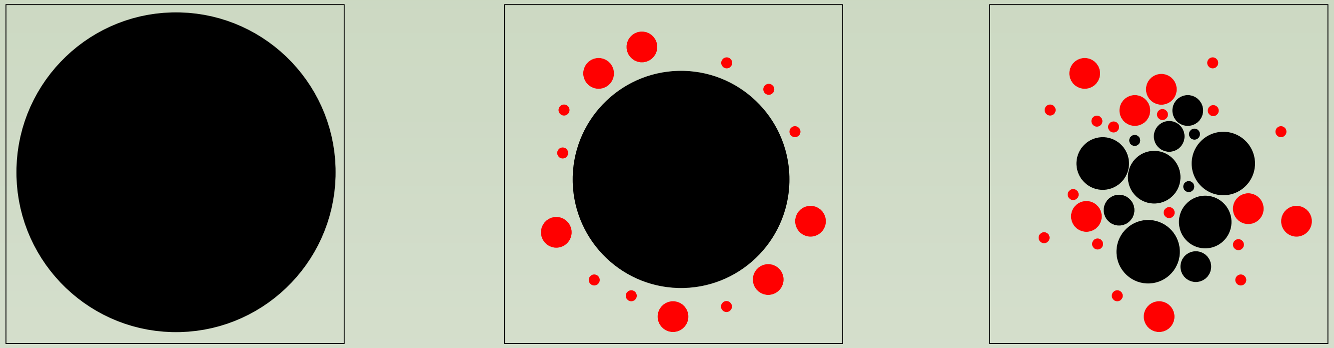


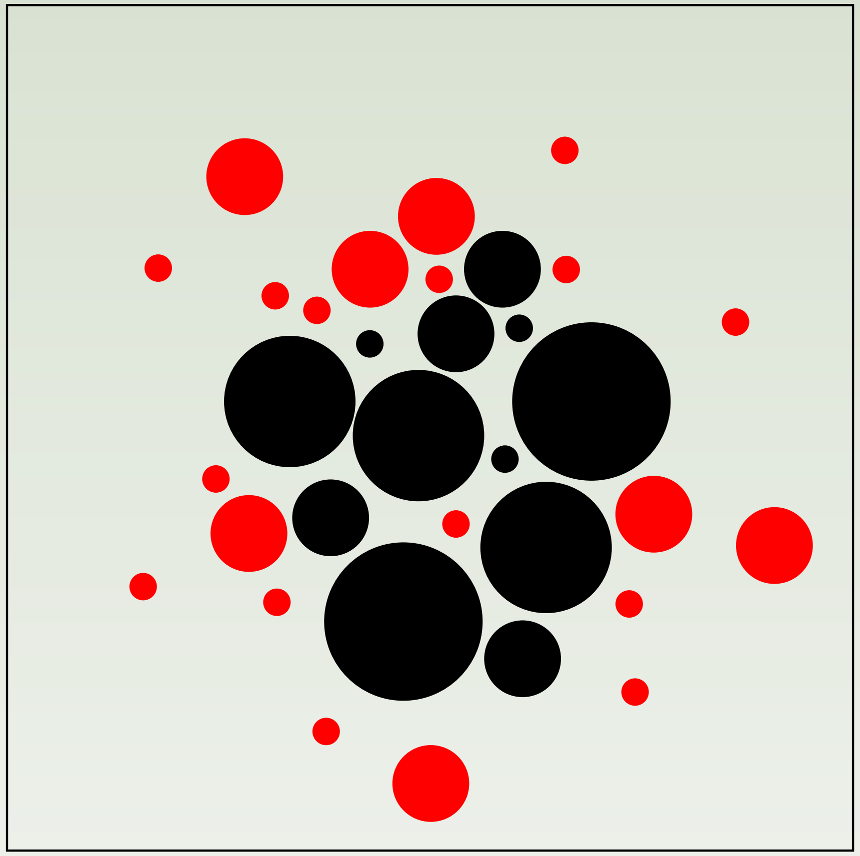


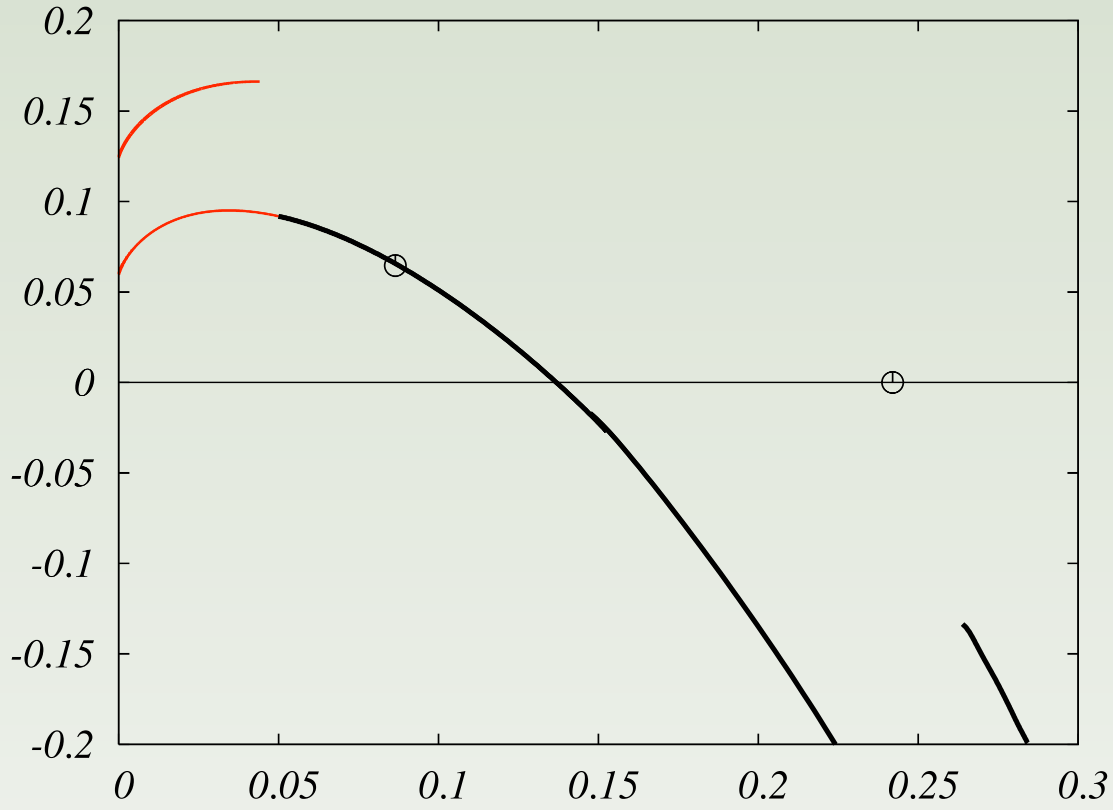


6 coloring of regular random graph

connectivity c=17



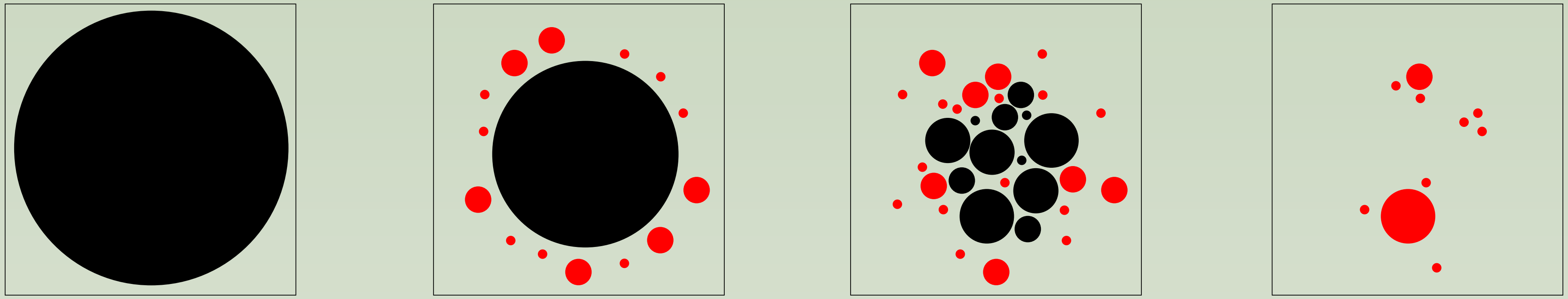


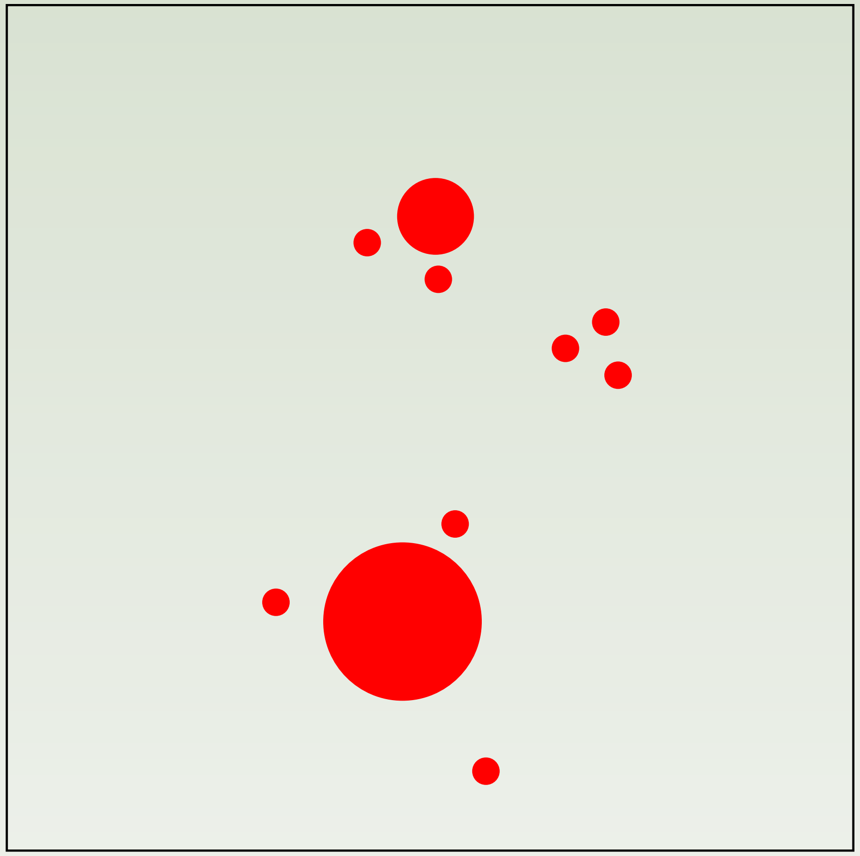


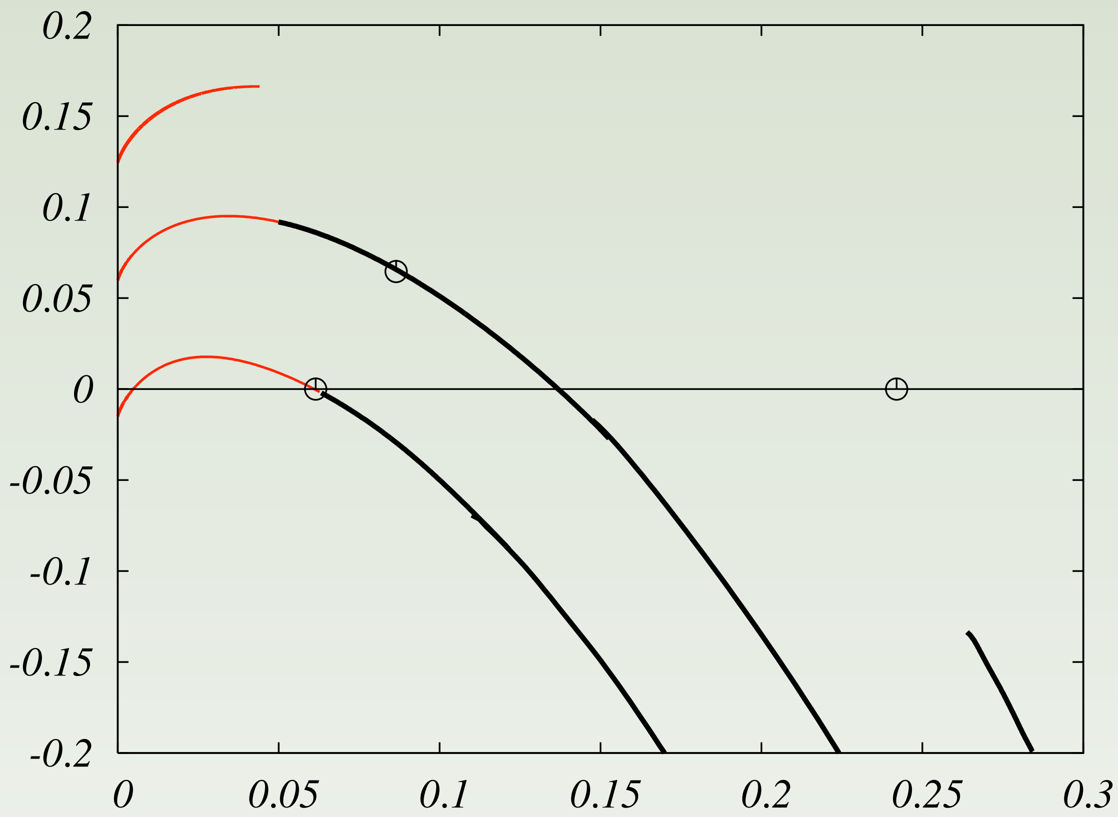
6 coloring of regular random graph

connectivity c=18

Tuesday, June 18, 13

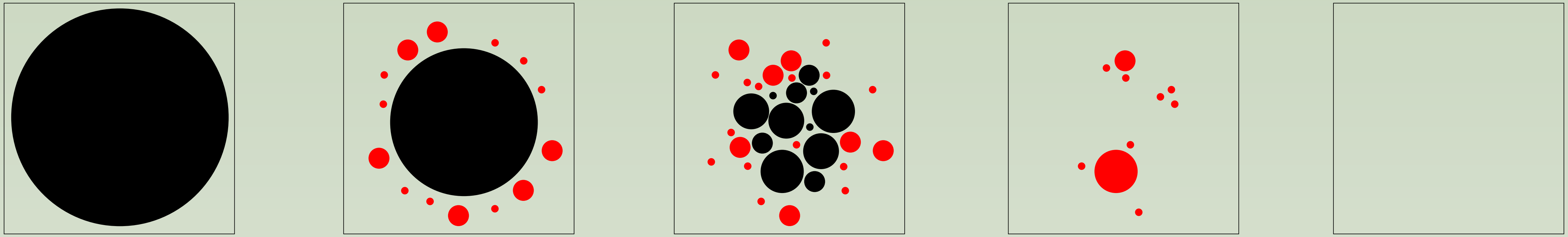


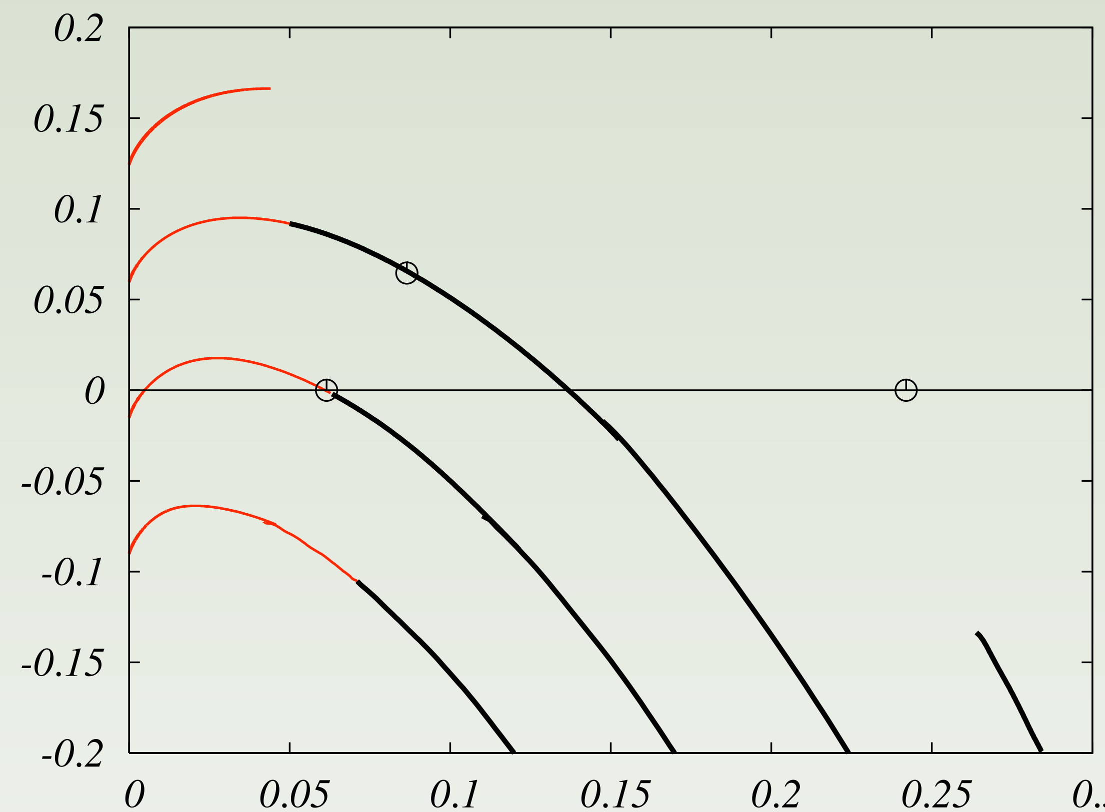




6 coloring of regular random graph

connectivity c=19





6 coloring of regular random graph

connectivity c=20