Statistical Physics and Machine learning 101, (part 3)

Florent Krzakala

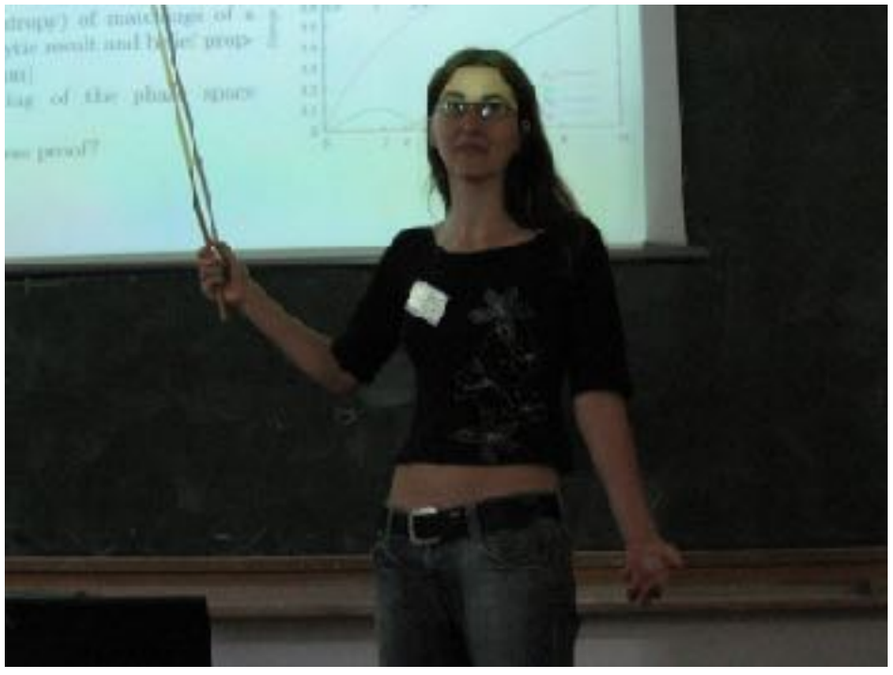
Institut Universitaire de France

Les Houches winter workshop (2003)

Learning the cavity method from Marc Mezard, in 2003 in a winter session in Les Houches

Les Houches summer school (2006)

2 young promising students



A distinguished professor

A notable visiting spouse during an "historic" moment

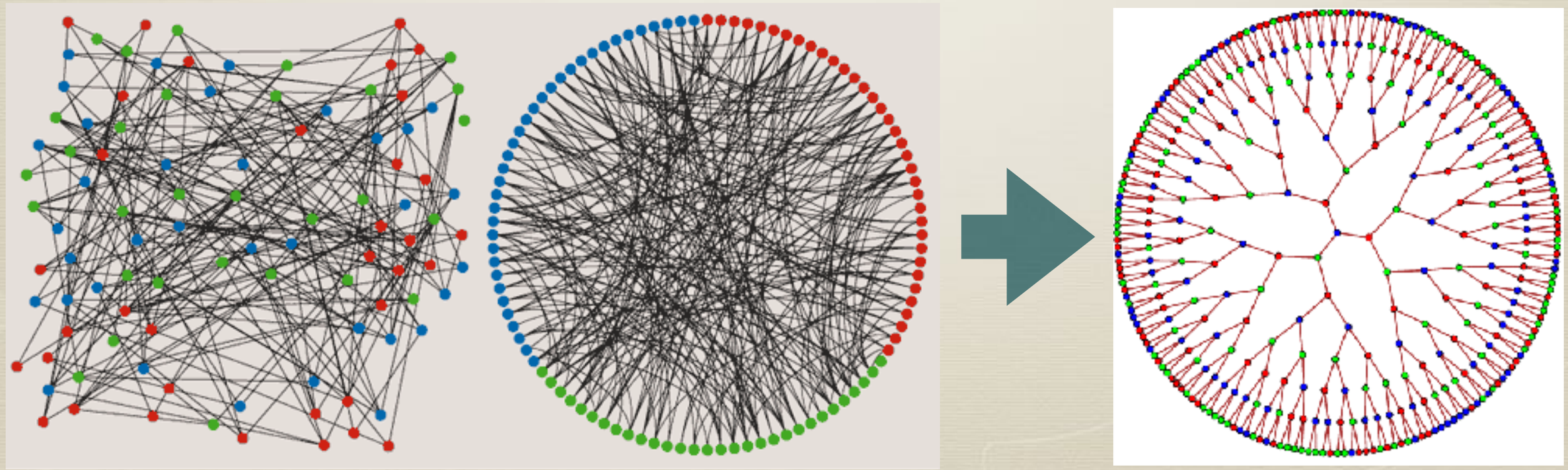
Statistical physics of colouring

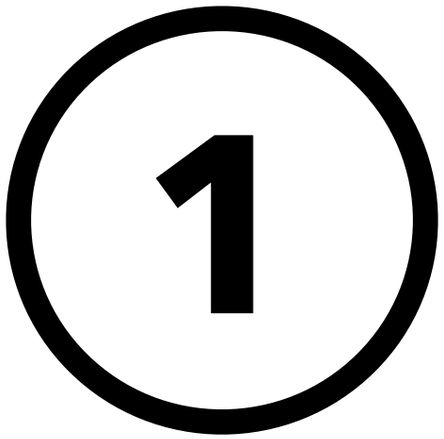
Coloring = zero temperature behavior of an anti-ferromagnetic Potts model.

Cost function = Hamiltonian: $\mathcal{H} = \sum_{<ij>} \delta(s_i, s_j) \qquad s_i = 1, 2, \ldots, q$

Spin glass-like problem : The random graph plays the role of disorder

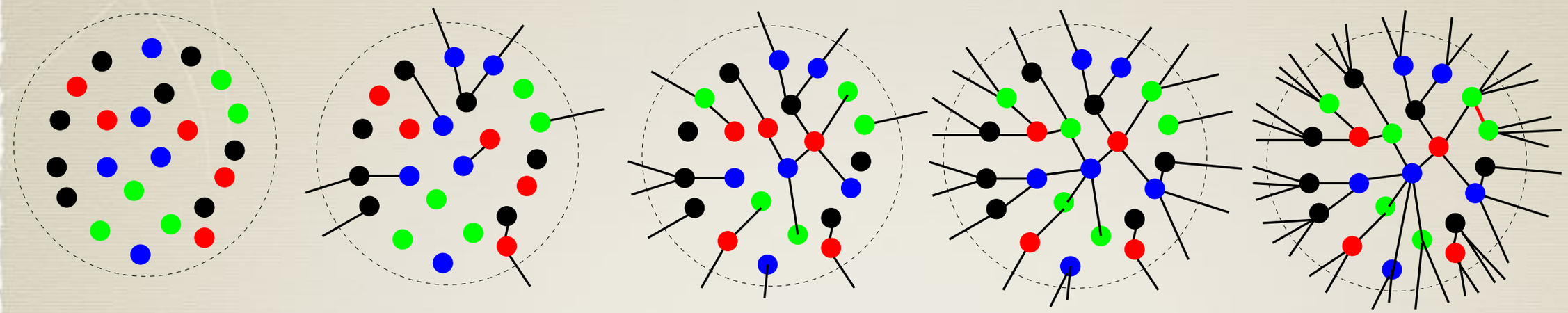
Tree-like local structure of the graph: as N grows the neighborhood of a random vertex is almost surely a tree up to distance $\log_c N$.

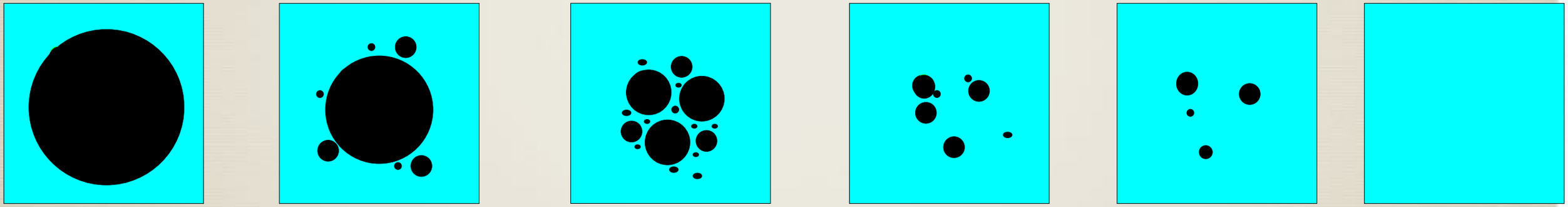




The space of solution of random Constraint satisfaction problems

The Phase Space



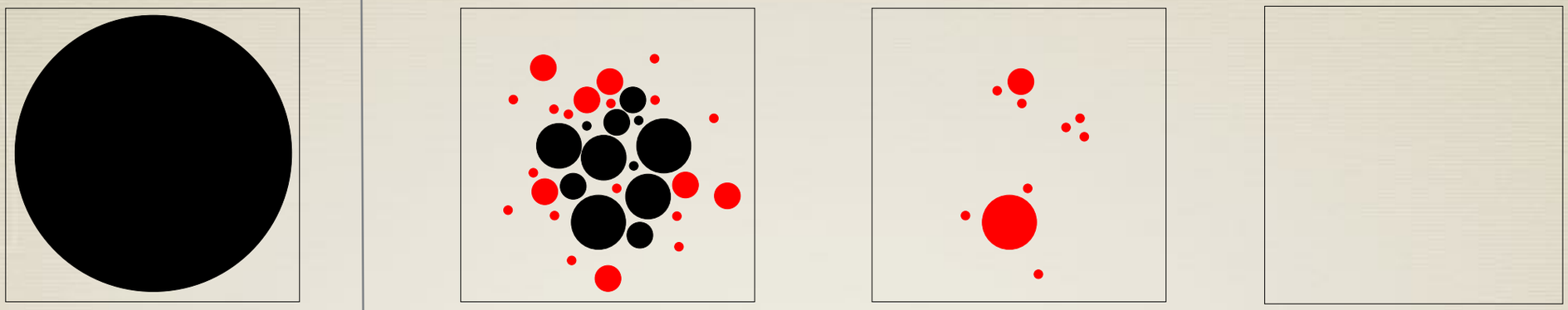


Increasing average degree
(more constraints, less freedom)

*[Zdeborova, **FK** '07; **FK**, Montanari, Ricci-Tersenghi, Semerjian, & Zdeborová '07]*

Many phase transitions

Average degree

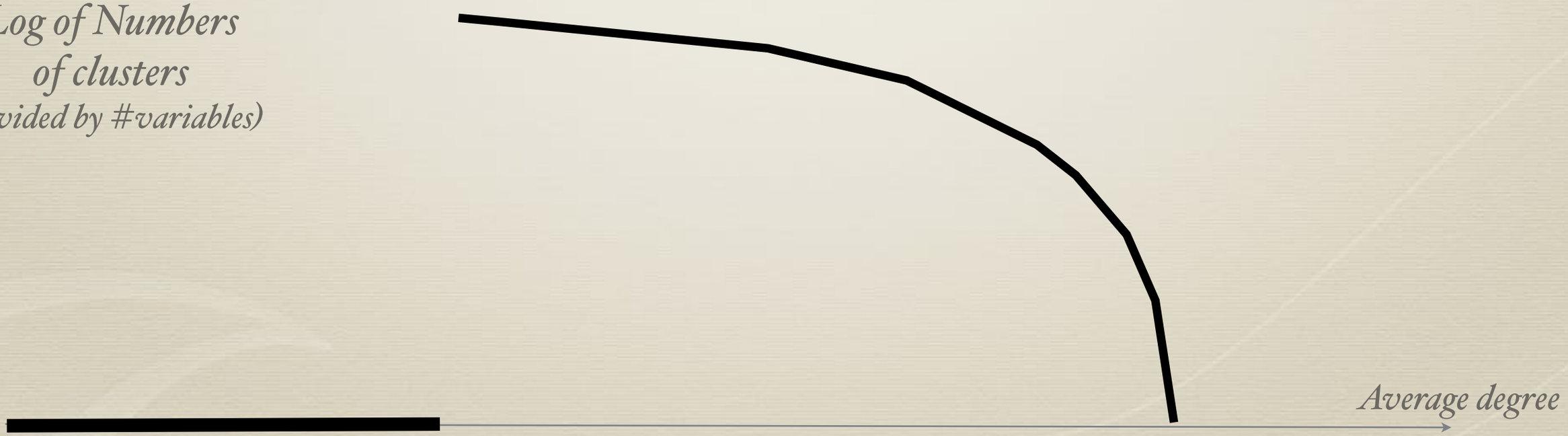


c_d

- **Clustering/Dynamic transition**
- **The phase space splits into exponentially many states**

$$c_d(3) = 4,\ c_d(4) = 8.35,\ c_d(5) = 12.84$$

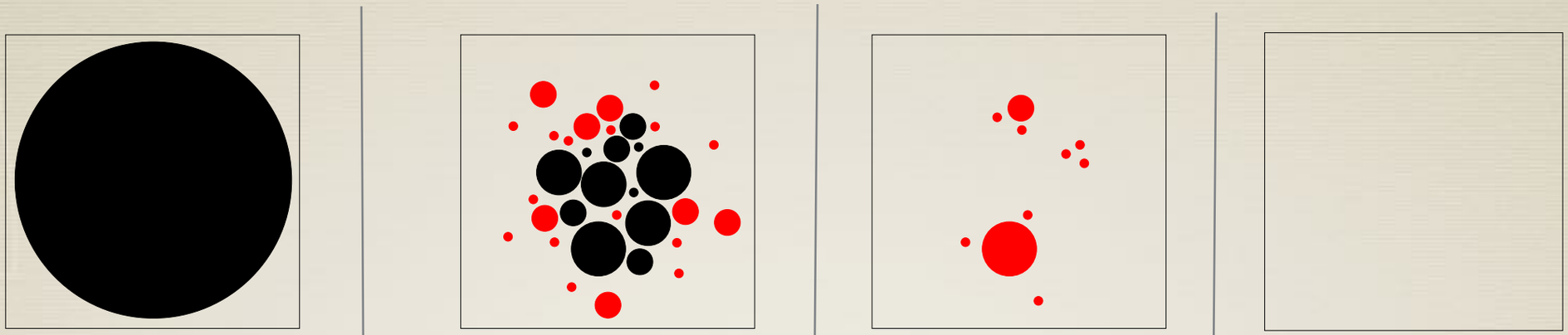
Log of Numbers of clusters (divided by #variables)



Average degree

Many phase transitions

Average degree



Clustering/Dynamic transition

The phase space splits into exponentially many states

$$c_d(3) = 4,\ c_d(4) = 8.35,\ c_d(5) = 12.84$$

Condensation transition

Entropy dominated by finite number of the largest states.

$$c_c(3) = 4,\ c_c(4) = 8.46,\ c_c(5) = 13.23$$

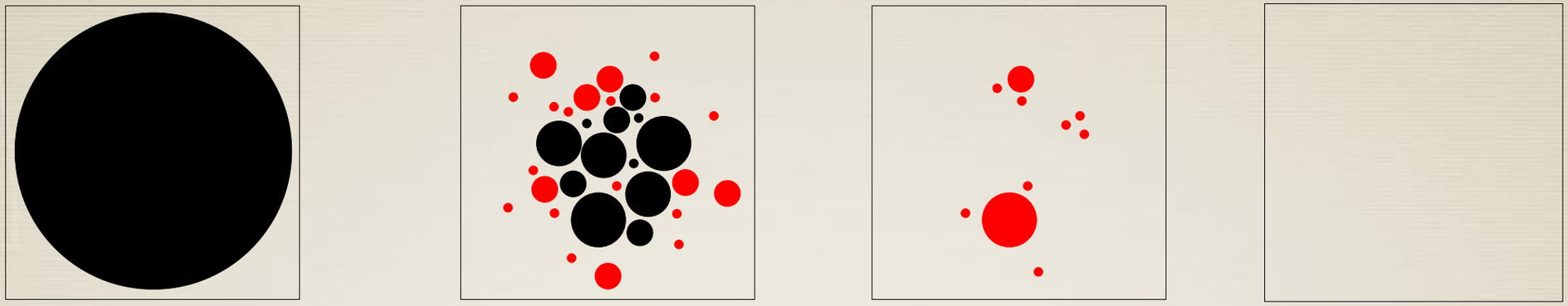
COL/UNCOL transition

No more clusters, uncolorable phase

$$c_s(3) = 4.69,\ c_s(4) = 8.90,\ c_s(5) = 13.67$$

Many phase transitions

Average degree



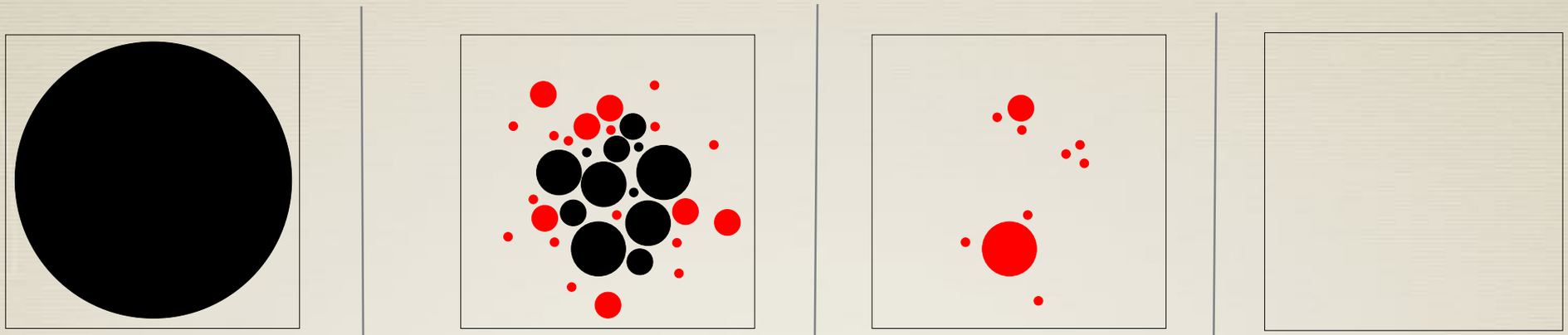
Log of Numbers of clusters (divided by #variables)

Log of Numbers needed to cover 99,99% of solutions (divided by #variables)

Average degree

Many phase transitions

Average degree



Clustering/Dynamic transition

The phase space splits into exponentially many states

$$c_d(3) = 4,\ c_d(4) = 8.35,\ c_d(5) = 12.84$$

Condensation transition

Entropy dominated by finite number of the largest states.

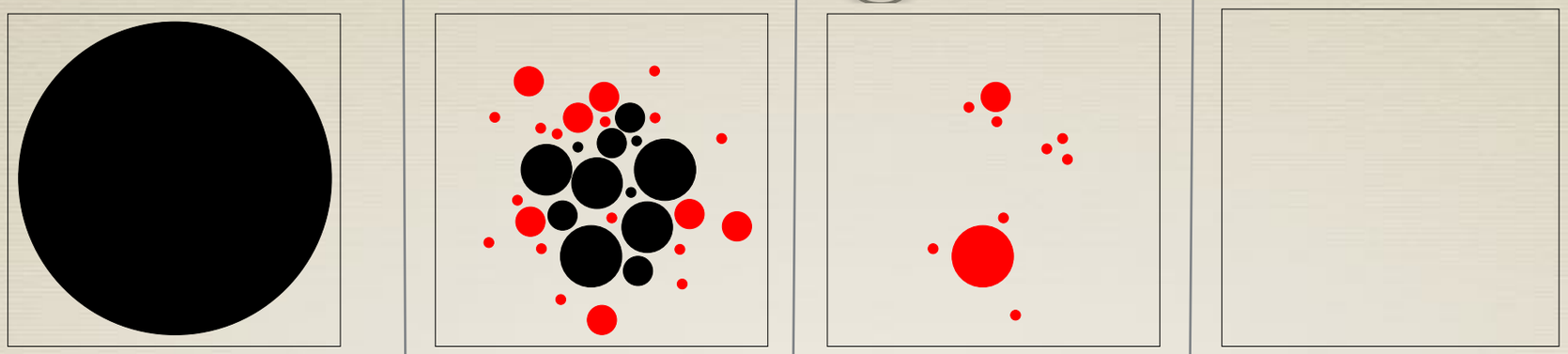
$$c_c(3) = 4,\ c_c(4) = 8.46,\ c_c(5) = 13.23$$

COL/UNCOL transition

No more clusters, uncolorable phase

$$c_s(3) = 4.69,\ c_s(4) = 8.90,\ c_s(5) = 13.67$$

Soft and Rigid clusters



Two types of clusters are found

Soft or "unfrozen" clusters

All variables are allowed at least two different colors in the cluster

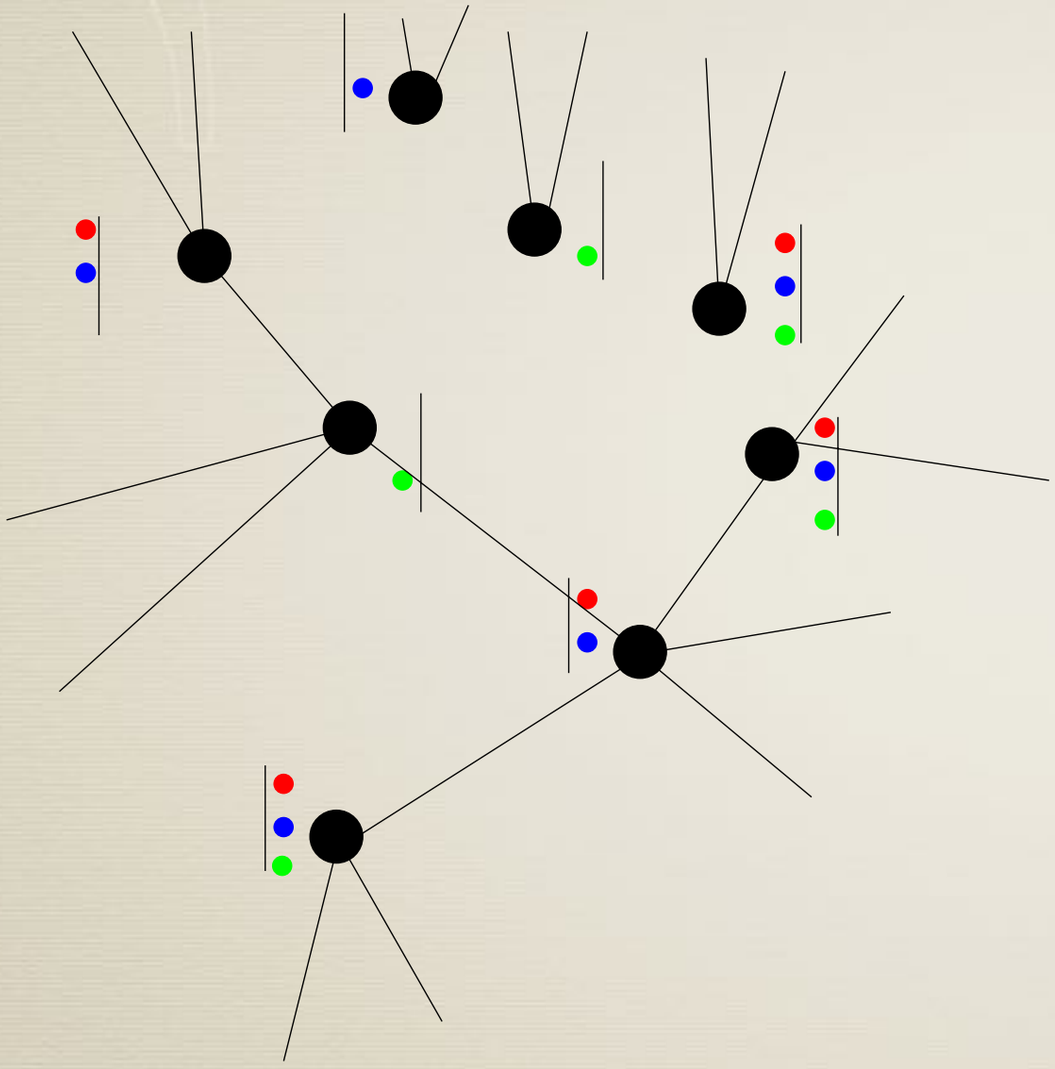
Rigid or "frozen" clusters

*A finite fraction of variables are allowed **only one color** in all solutions belonging to the cluster: we say that these variables "freeze"*

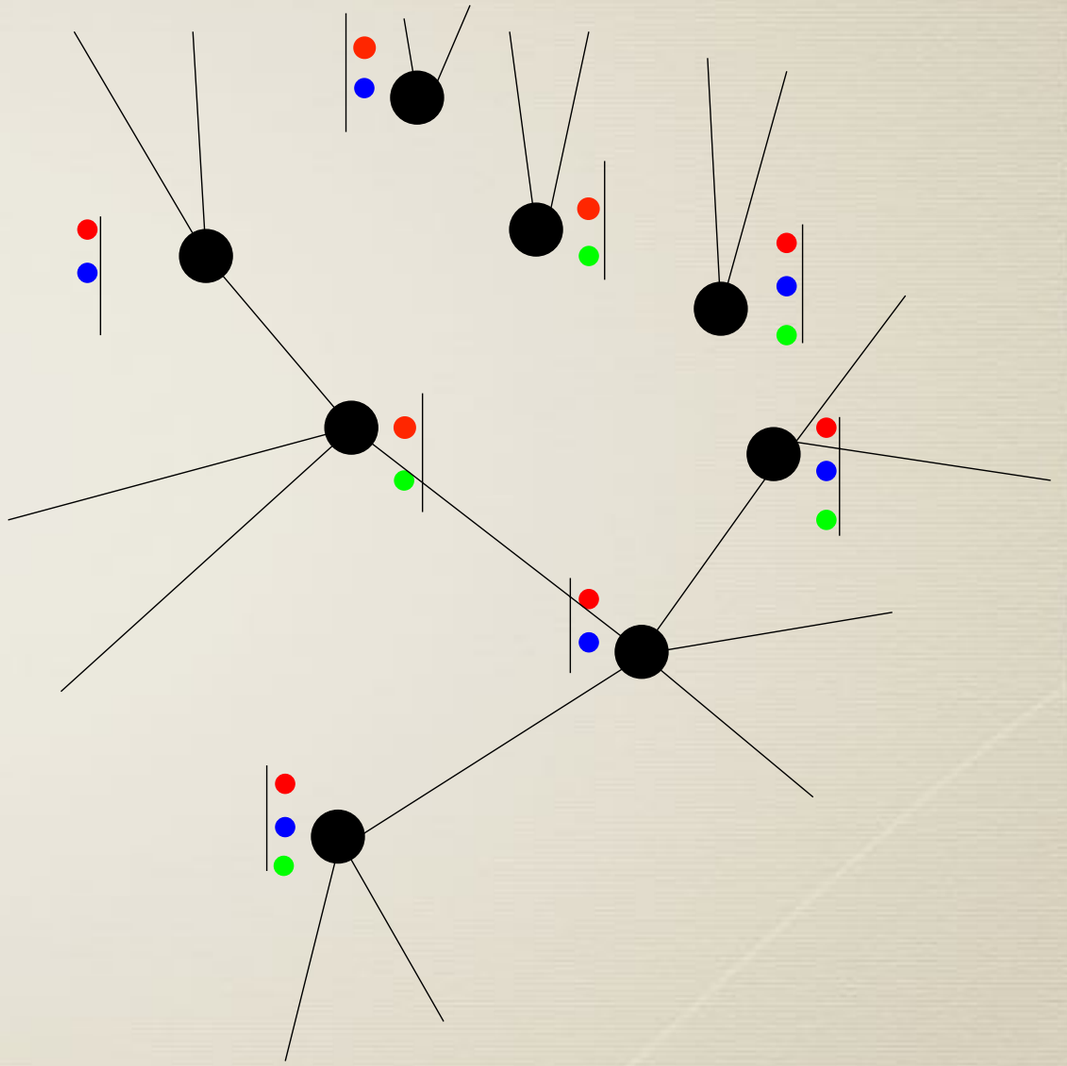
The set of solutions in a frozen cluster has a **rigid backbone:**

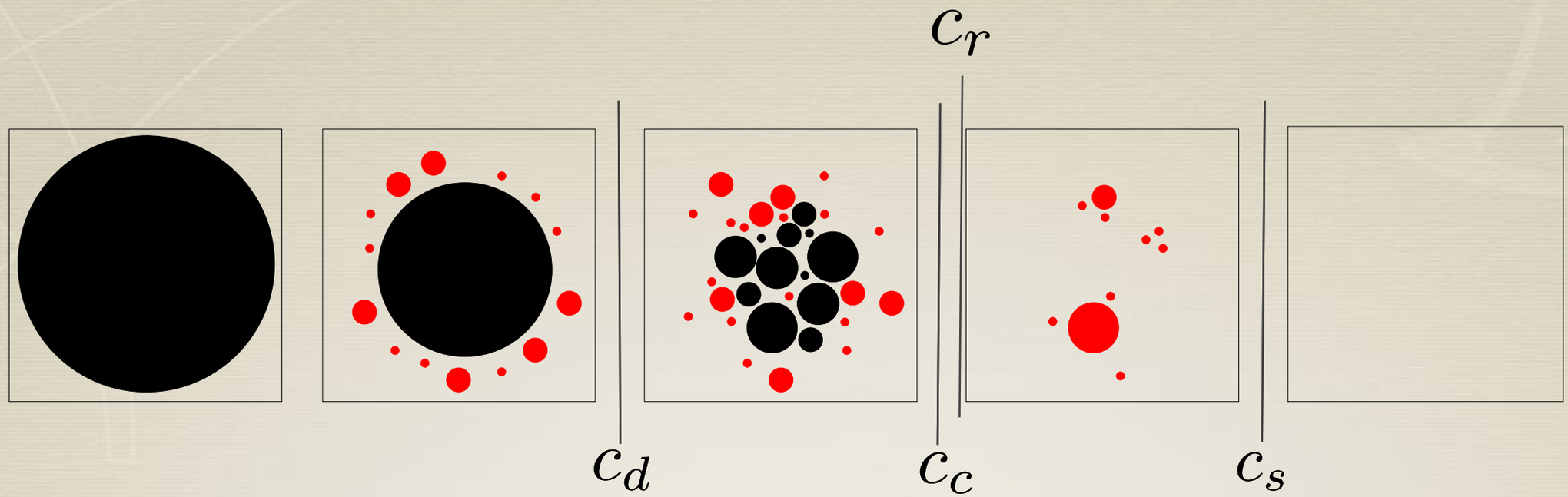
A finite fraction of variables have the same values in all solutions

Allowed colors in a frozen “rigid” cluster



Allowed colors in a unfrozen “soft” cluster





Clustering/Dynamic transition

- The phase space splits into exponentially many states

$c_d(3) = 4,\ c_d(4) = 8.35,\ c_d(5) = 12.84$

Condensation transition

- Entropy dominated by finite number of the largest states.

$c_c(3) = 4,\ c_c(4) = 8.46,\ c_c(5) = 13.23$

COL/UNCOL transition

- No more clusters, uncolorable phase

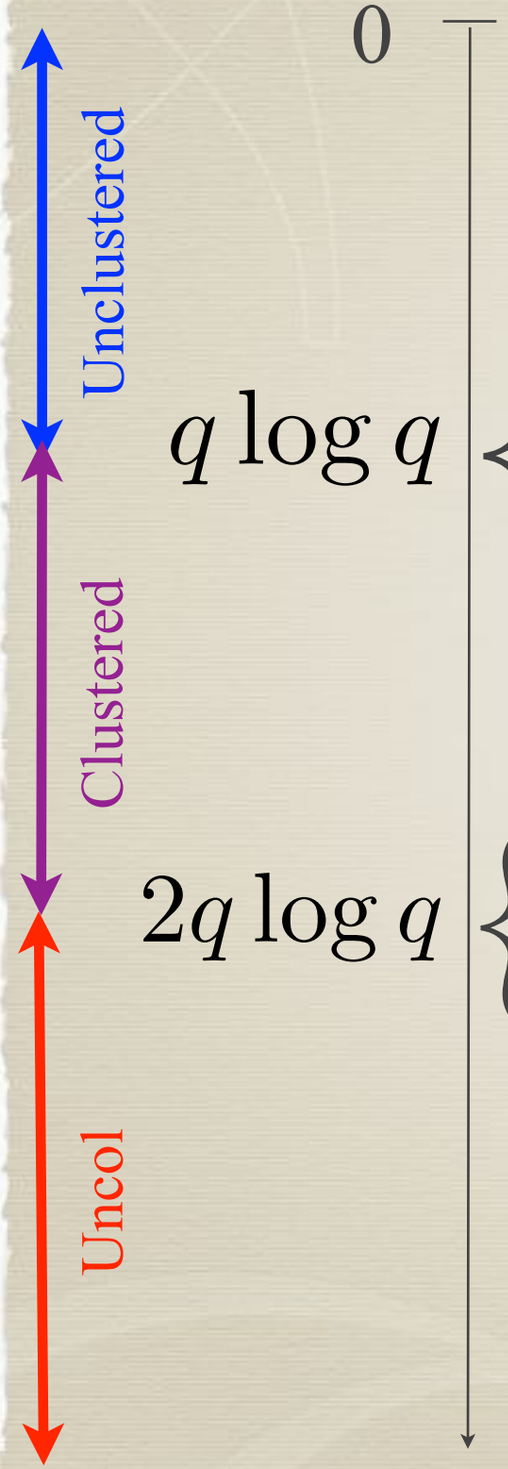
$c_s(3) = 4.69,\ c_s(4) = 8.90,\ c_s(5) = 13.67$

Rigidity/Jamming transition

- Frozen variables appears in the dominating states.

$c_r(3) = 4.66,\ c_r(4) = 8.83,\ c_r(5) = 13.55$

Asymptotic for large number of colors



$$c_d \approx q(\log q + \log\log q + 1 - O(1))$$

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

$$c_K \approx 2q\log q - \log q - 2\log 2$$

$$c_q \simeq 2q\log q - \log q - 1$$

years

PROOF OF THE SATISFIABILITY CONJECTURE FOR LARGE k

JIAN DING*, ALLAN SLY†, AND NIKE SUN

University of Chicago;
University of California–Berkeley and Australian National University;
Microsoft Research and Massachusetts Institute of Technology

ABSTRACT. We establish the satisfiability threshold for random k-SAT for all $k \geqslant k_0$. That is, there exists a limiting density $\alpha_s(k)$ such that a random k-SAT formula of clause density α is with high probability satisfiable for $\alpha < \alpha_s$, and unsatisfiable for $\alpha > \alpha_s$. The satisfiability threshold $\alpha_s(k)$ is given explicitly by the one-step replica symmetry breaking prediction from statistical physics. We believe that our methods may apply to a range of random CSPs in the 1RSB universality class.

1. INTRODUCTION

A long-standing open problem has been to understand the density of constraints at which a random constraint satisfaction problems become unsatisfied. In the random k-SAT model, while a series of results have given what are now impressively close upper and lower bounds [Coj13], the existence of a critical density remained a fundamental open problem. In this paper we establish the satisfiability threshold for all k sufficiently large, and further give an explicit description of the threshold.

Our proof relies heavily on insights from statistical physics, in particular the way in which solutions break into clusters. Based on so-called replica symmetry breaking arguments, physicists have made detailed predictions concerning the geometry of the space of solutions. These heuristics guide our proof, and the limiting density α_s that we establish is given by the one-step replica symmetry breaking prediction [MPZ02, MMZ06].

The sharp upper bound on α_s follows from the interpolative free energy bounds for diluted spin glass models established by Franz–Leone [FL03] and Panchenko–Talagrand [PT04]. The

d

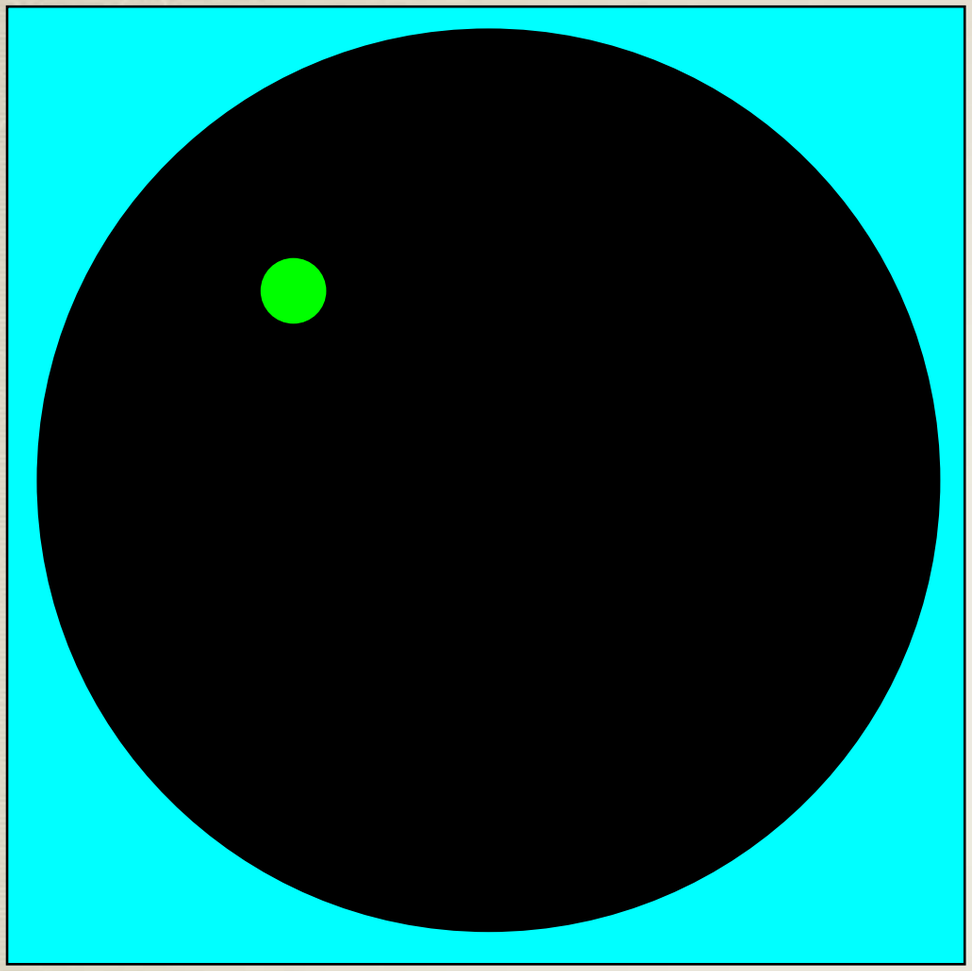
thers...

(Berkeley) (Berkeley) (Toronto)

2

Algorithmic consequences

Arkless strategy for flood victims



You are on a rugged landscape
that is being flooded
What to do?

Strategy 1: Build an ark !

Need for an ark-less strategy!

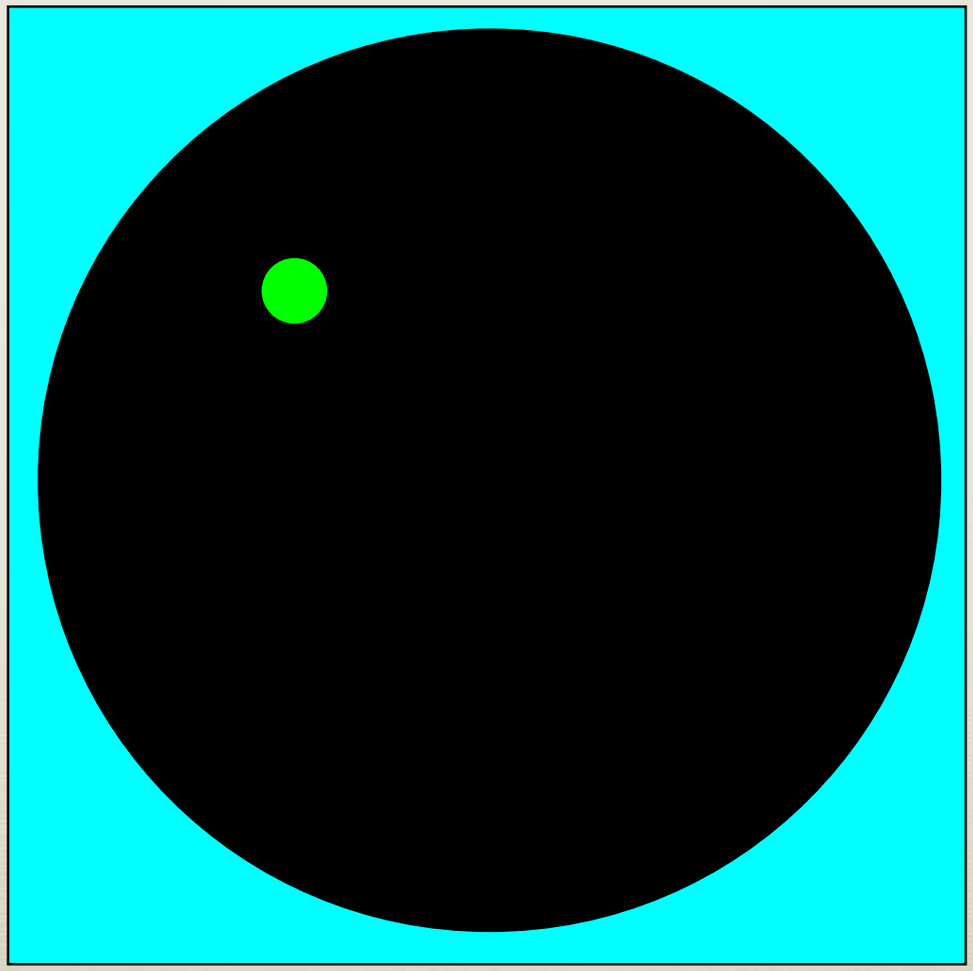
Too long !

*[**FK** , Kurchan '07]*

“Wet toes” algorithm

Arkless strategy for flood victims

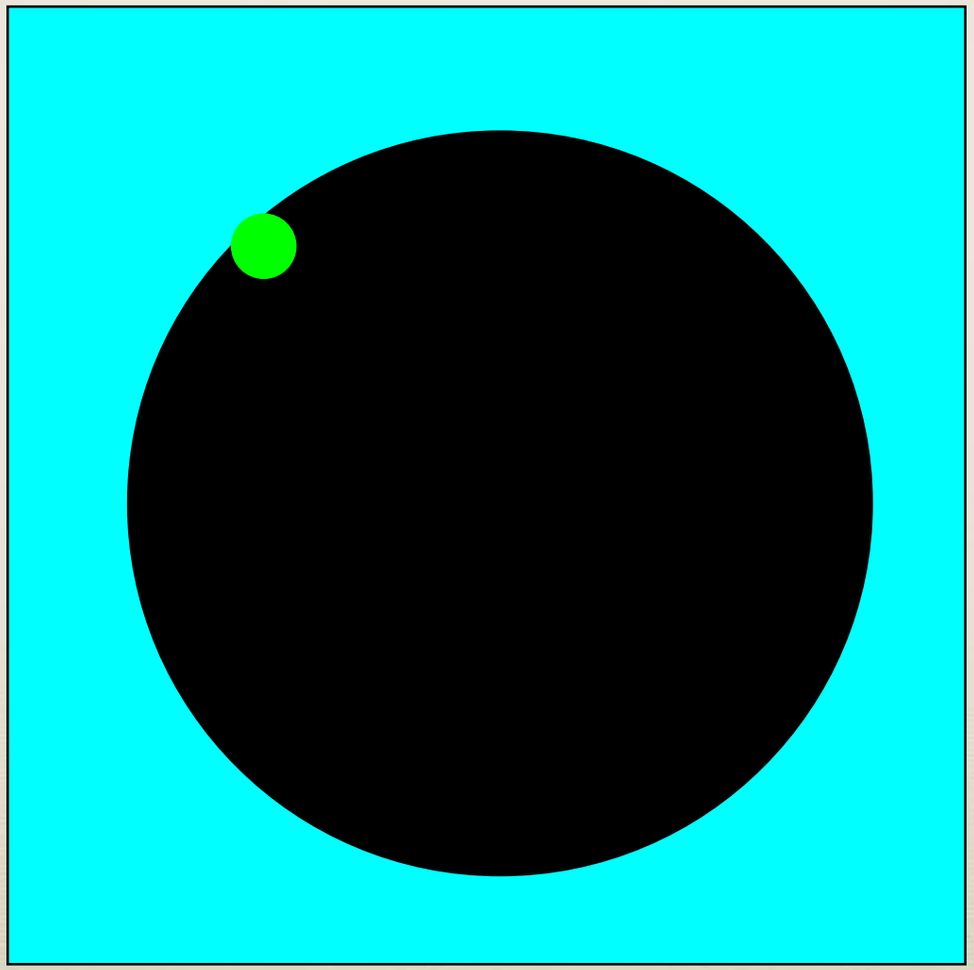
You are on a rugged landscape that is being flooded



“Wet toes” algorithm

Arkless strategy for flood victims

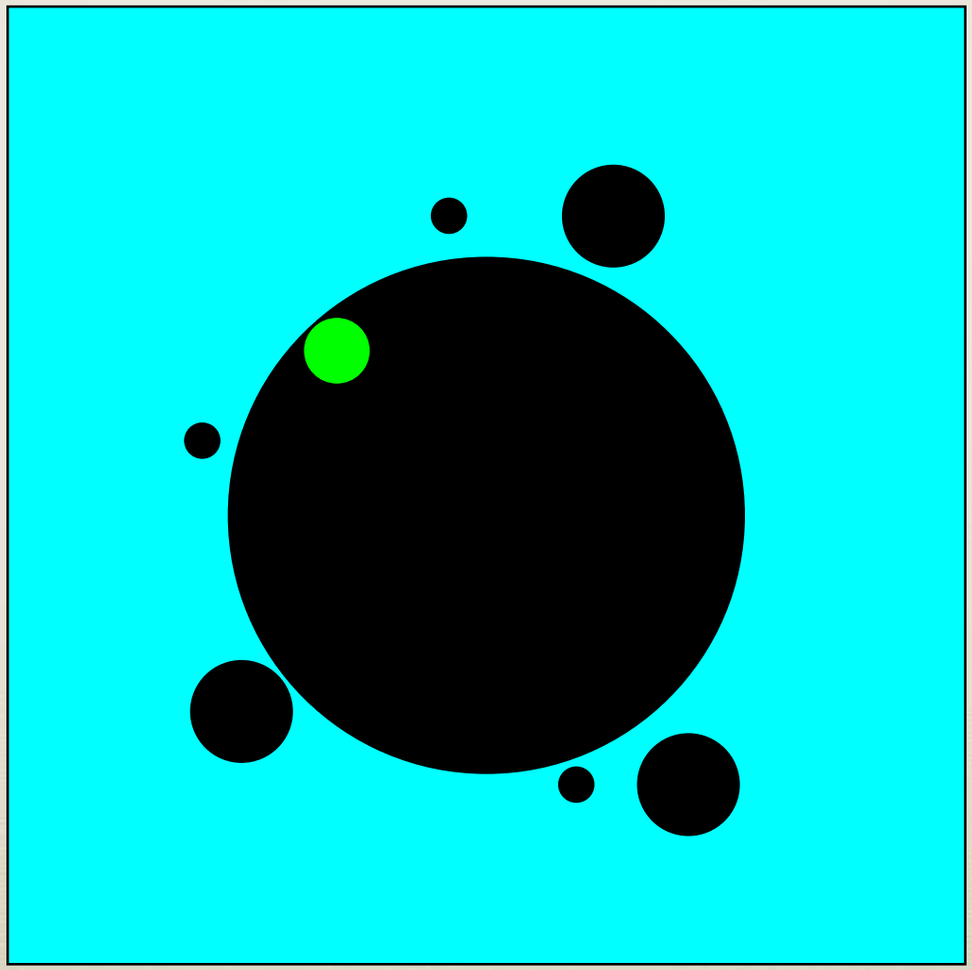
Water goes up. When your toes are wet
step back on the land!



“Wet toes” algorithm

Arkless strategy for flood victims

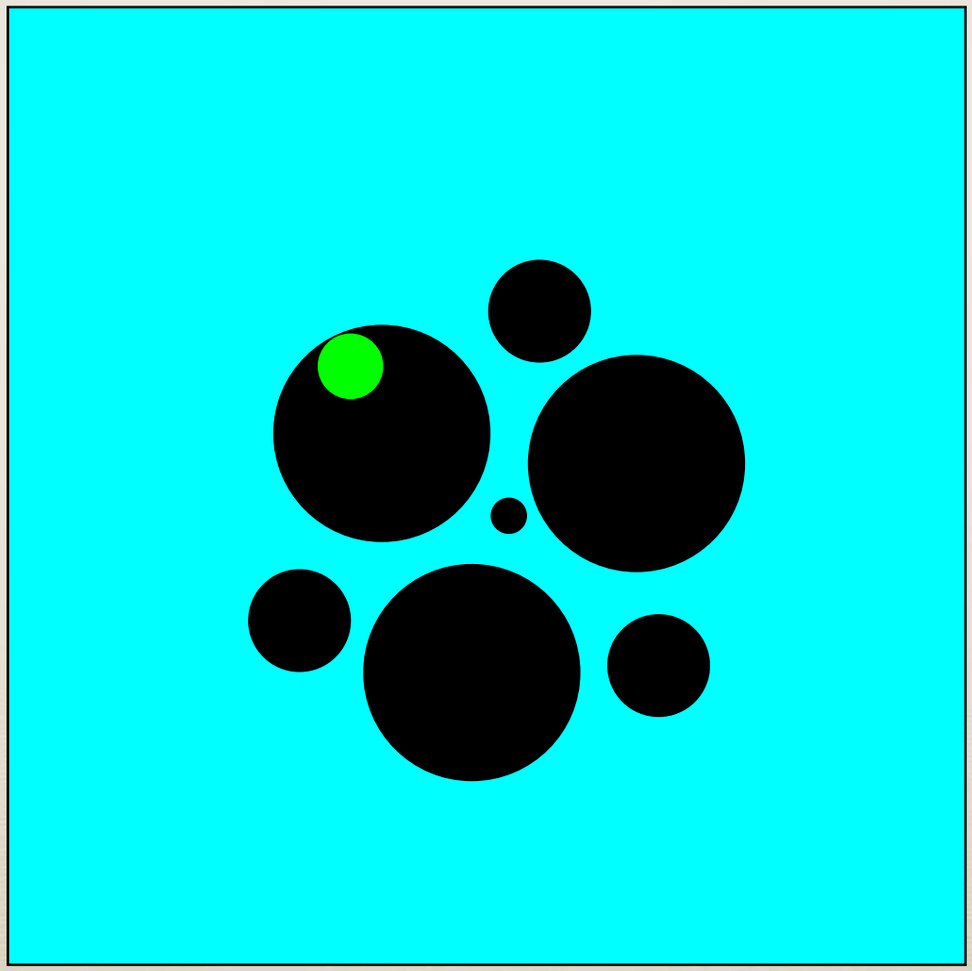
And wait until your toes get wet again...



“Wet toes” algorithm

Arkless strategy for flood victims

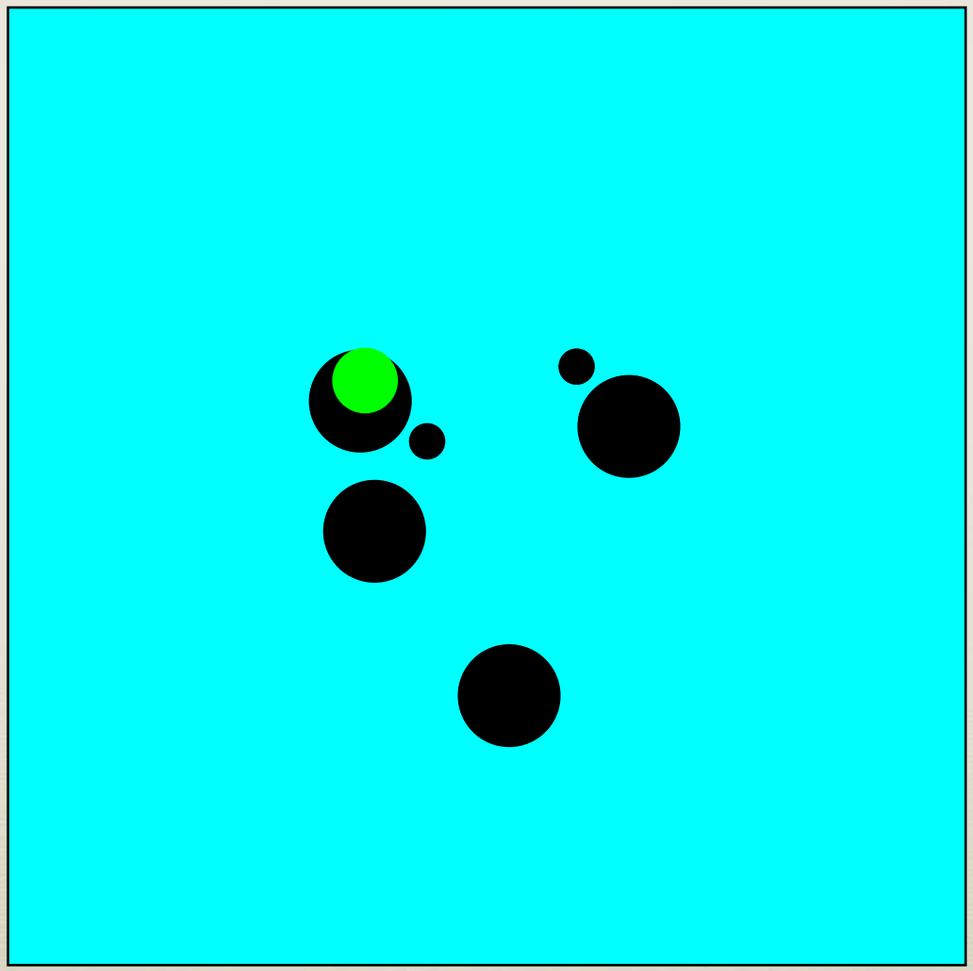
Sooner or later you’ll find yourself on a smaller island…



“Wet toes” algorithm

Arkless strategy for flood victims

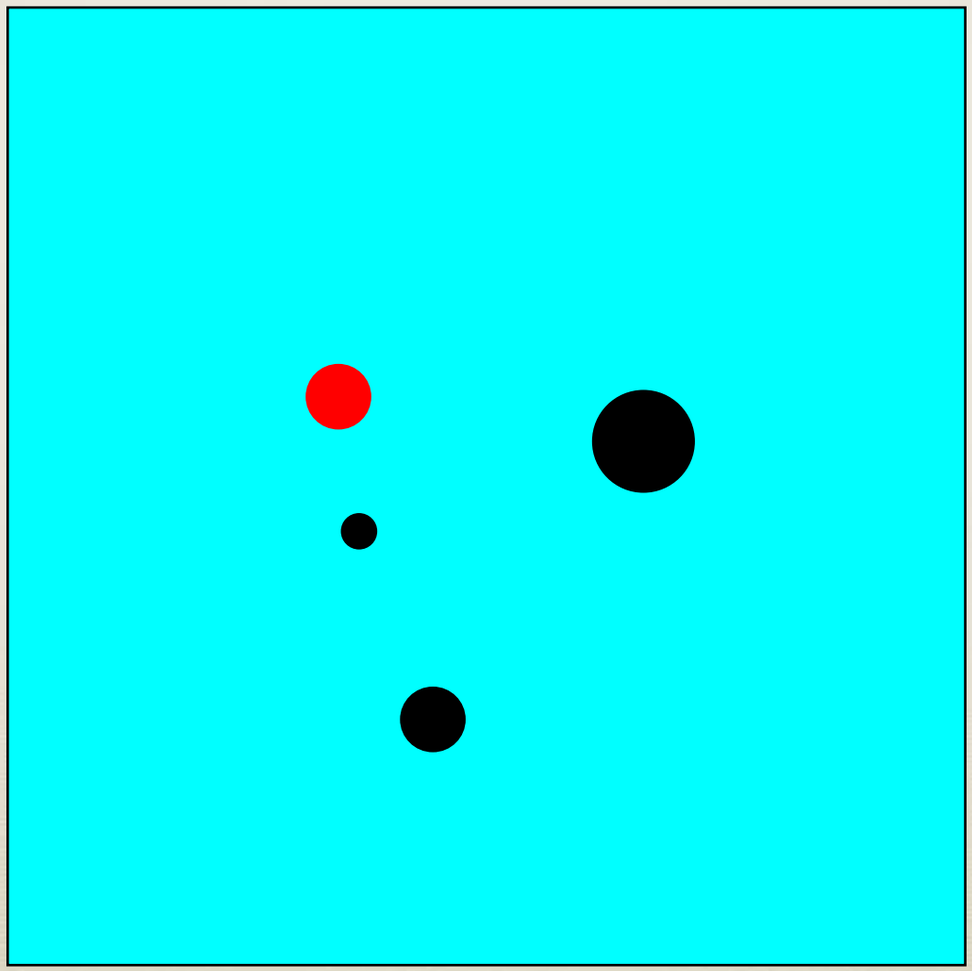
Then even a smaller one…



“Wet toes” algorithm

Arkless strategy for flood victims

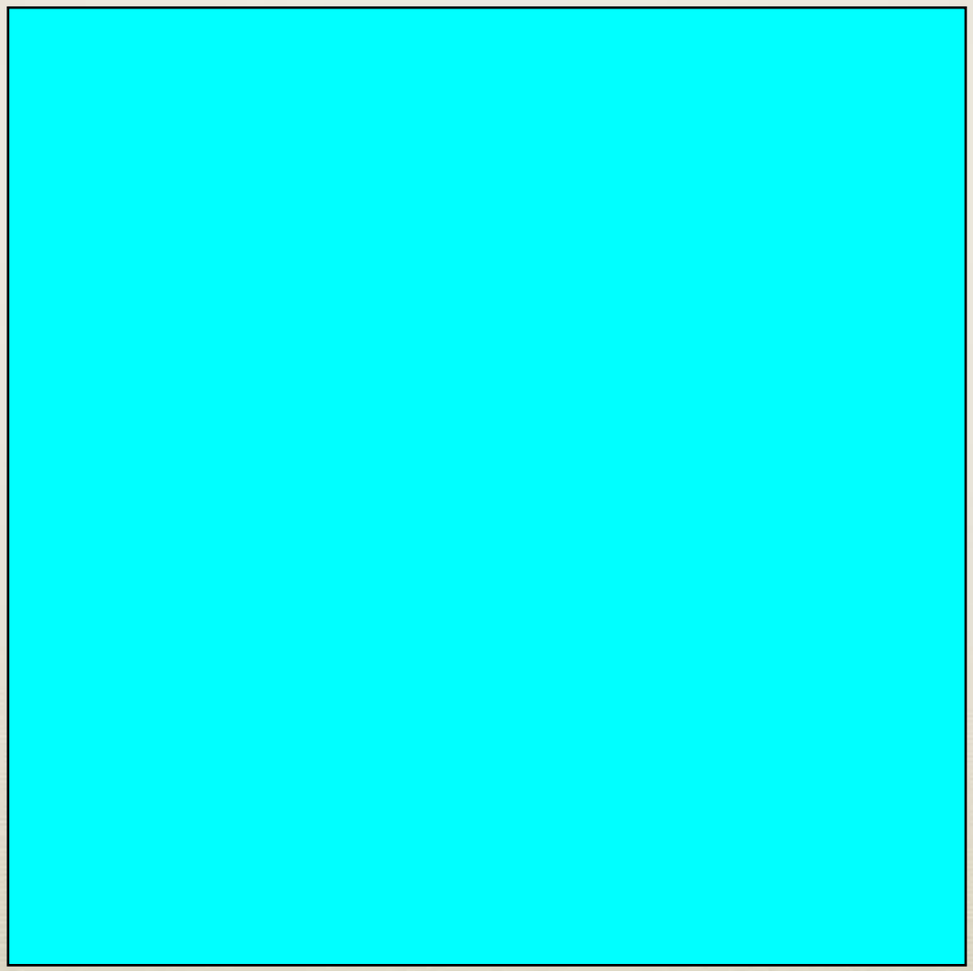
Until eventually you’ll drown (if you can’t swim!)



“Wet toes” algorithm

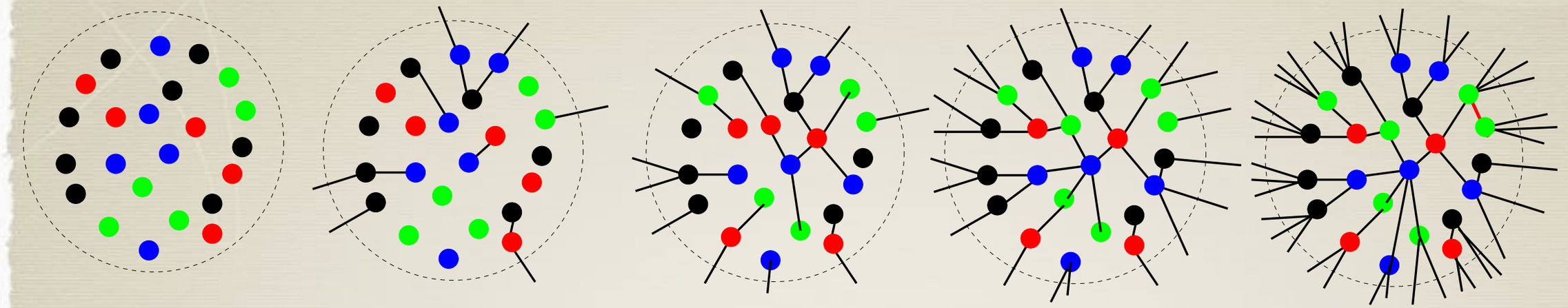
Arkless strategy for flood victims

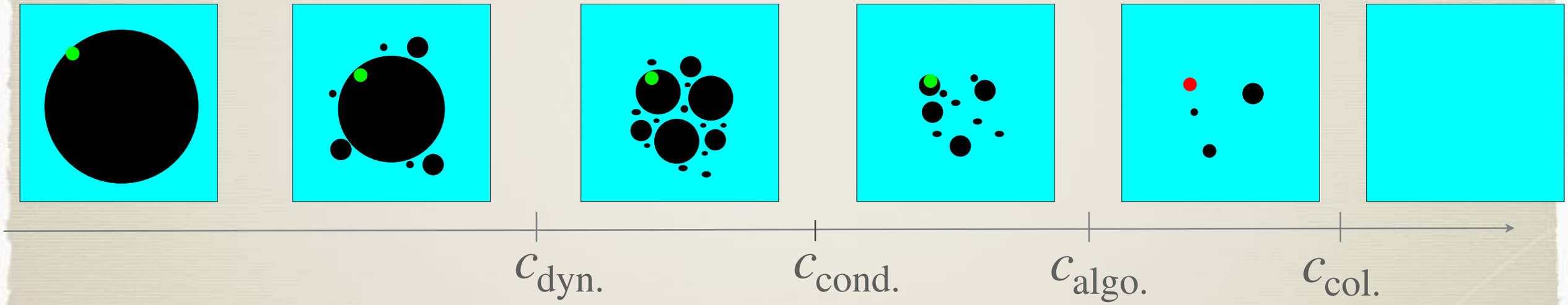
Finally, all land will be flooded!



“Wet toes” algorithm and coloring

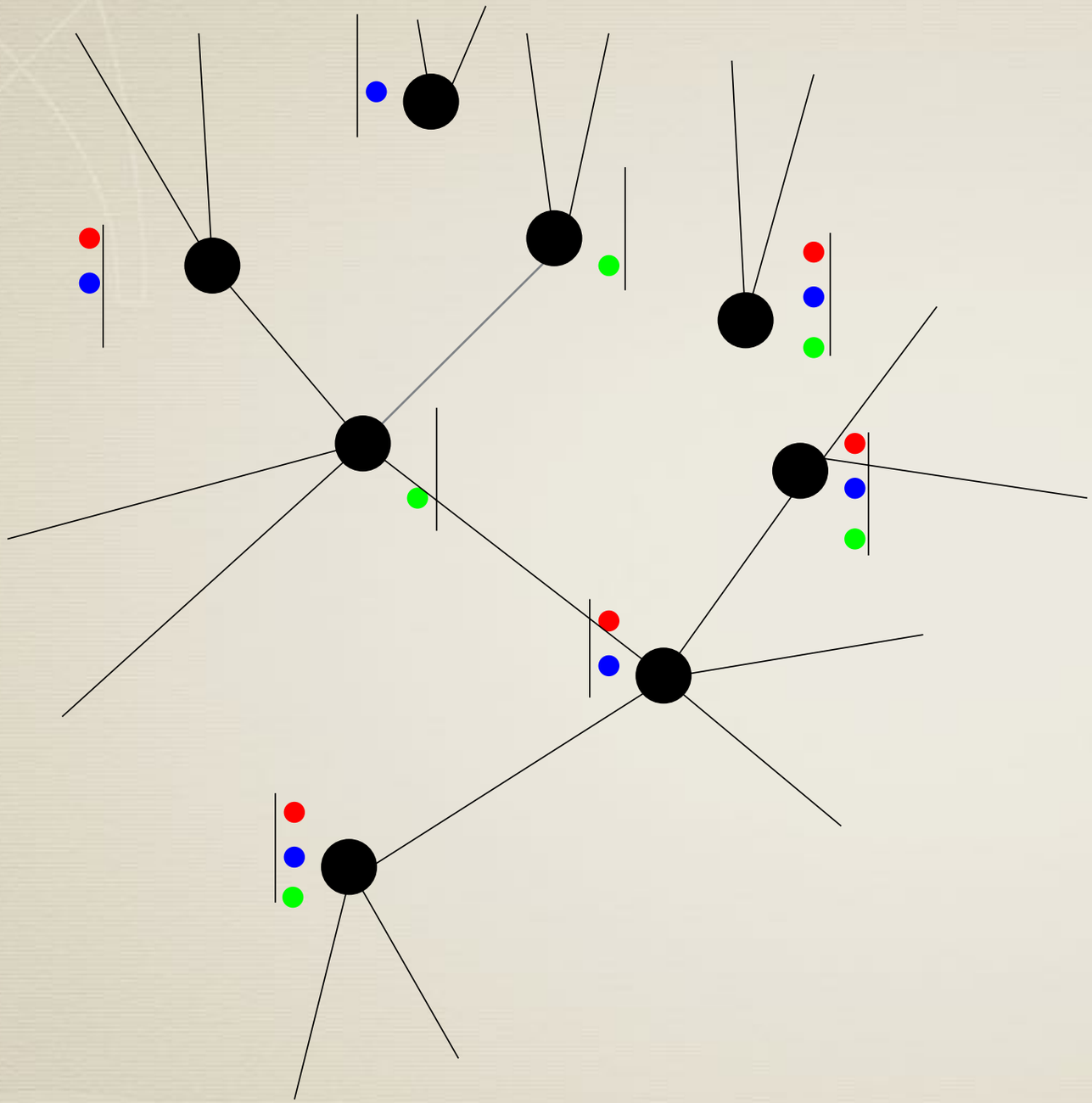
Add links one by one and use a local algorithm to solve contradictions





The algorithm works until the cluster disappears

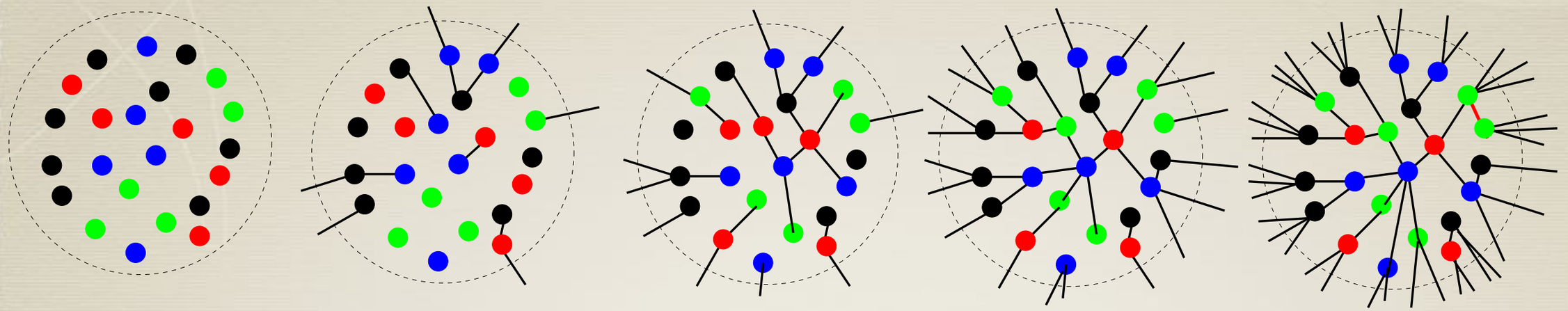
The hard-field catastrophe

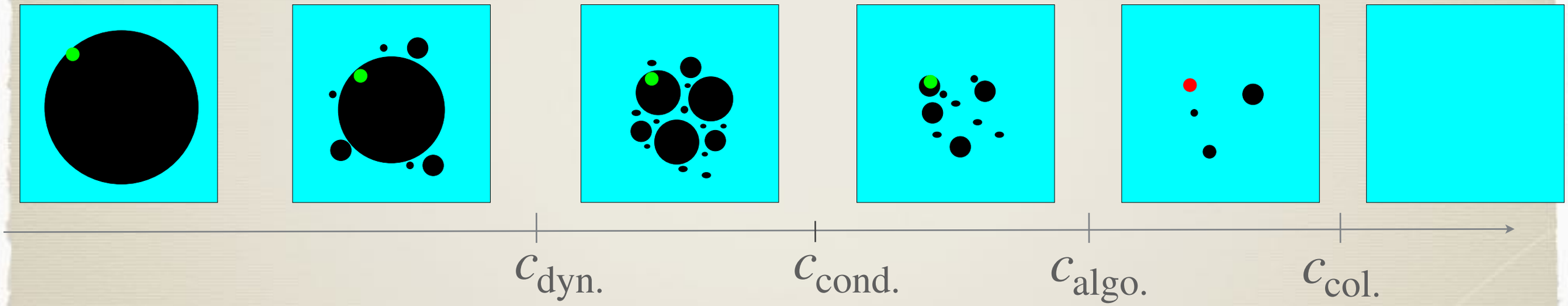


- *Consider a cluster with (a finite fraction of) frozen variables*
- *Add a link at random*
- *With a finite probability, the whole cluster is killed!*

“Wet toes” algorithm and coloring

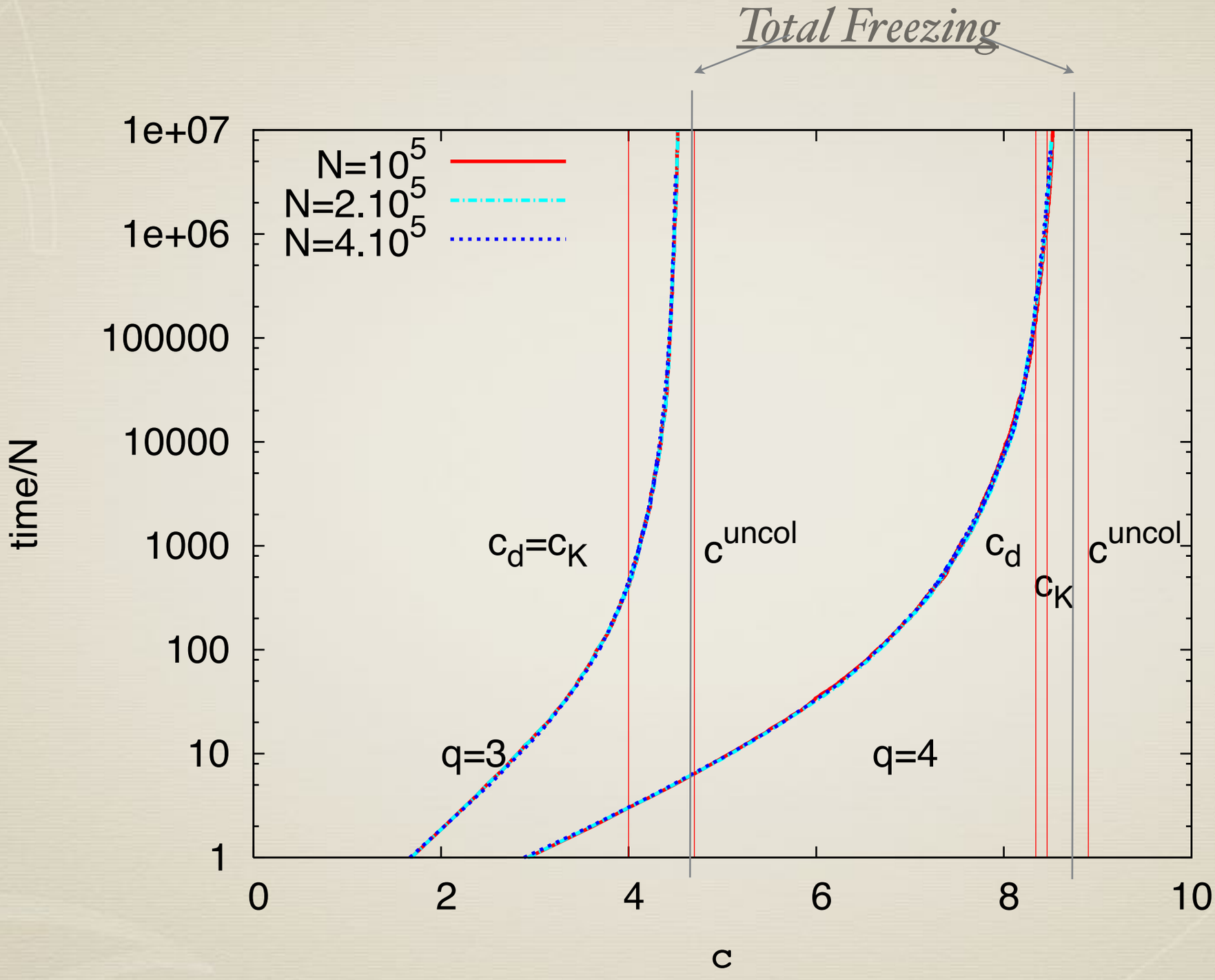
Add links one by one and use a local algorithm to solve contradictions





The algorithm works until the cluster disappears... and this happens when variables freeze!

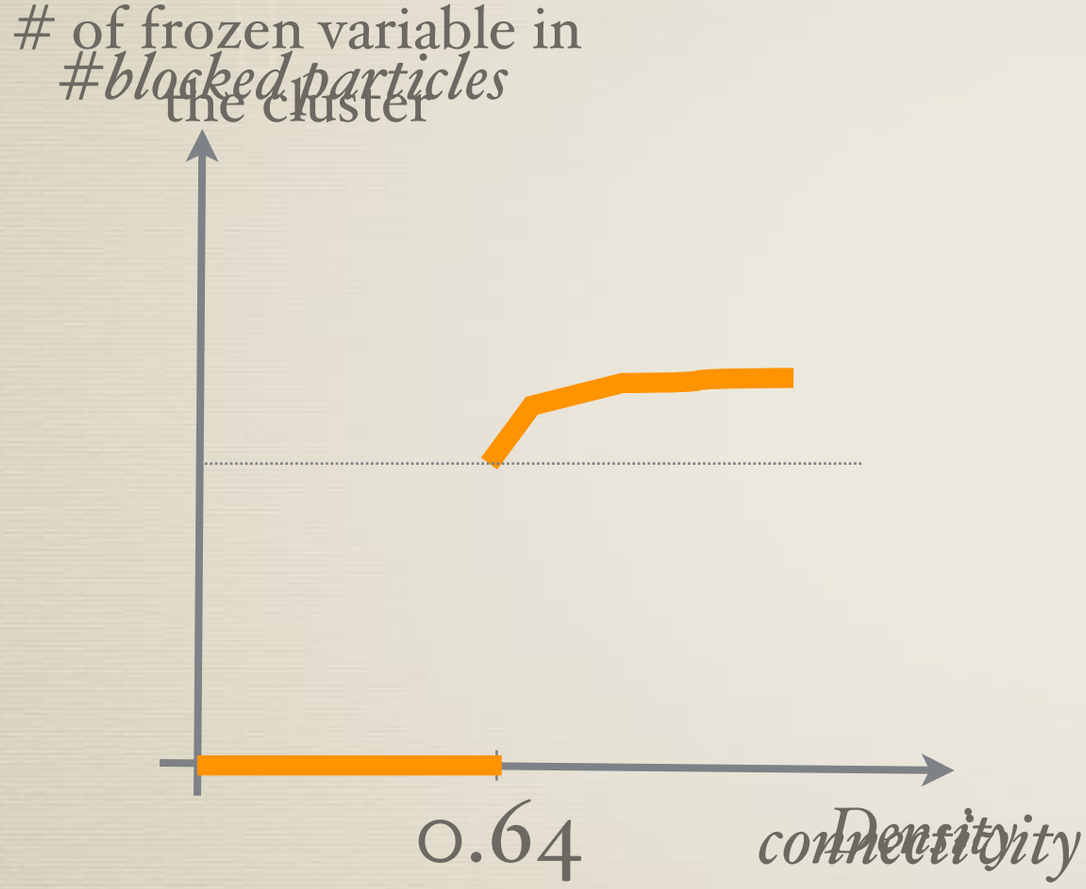
Performance of the "Wet toes" algorithm



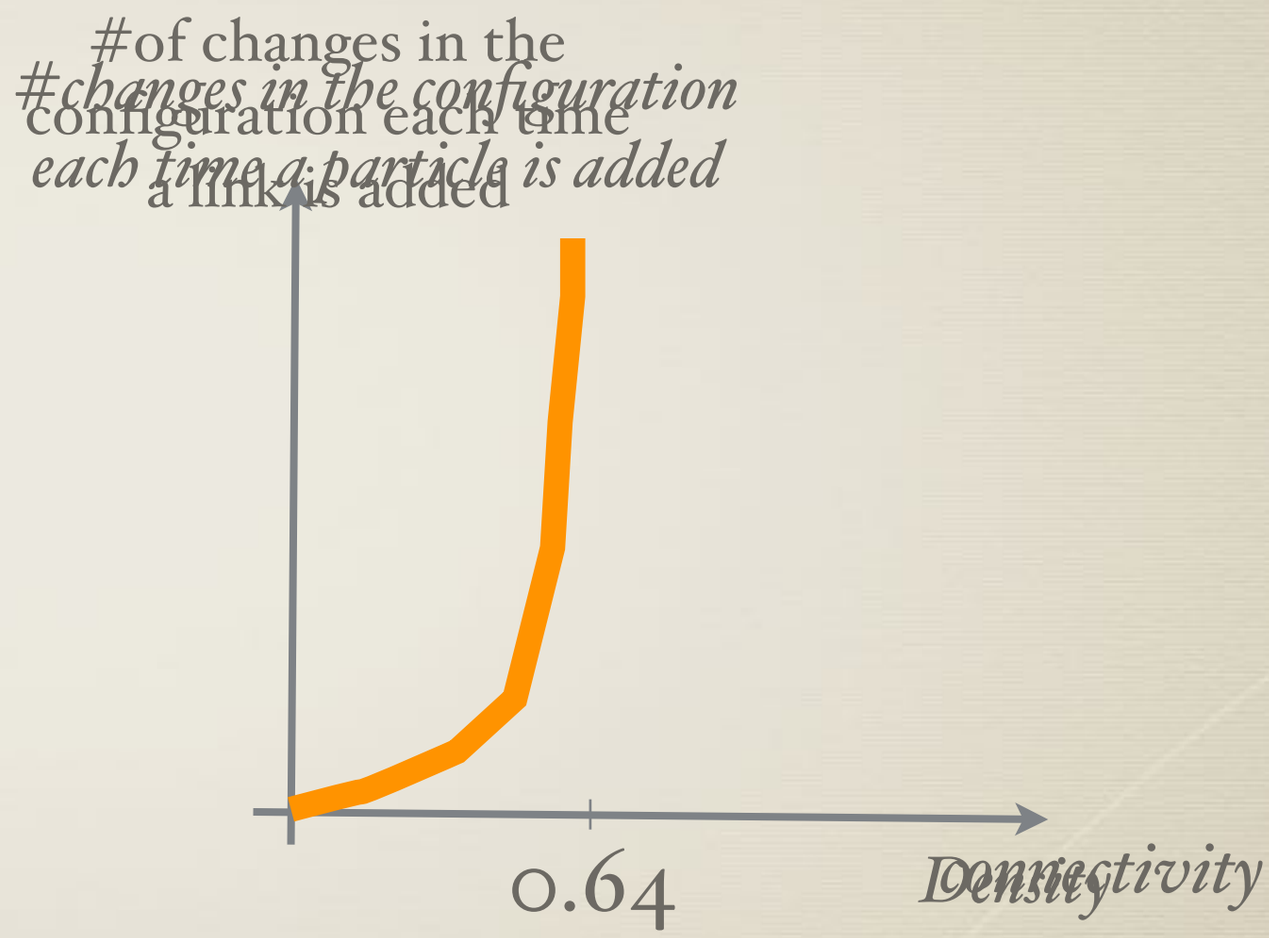
Goes beyond *the dynamical and the condensation transitions* for q=3 & 4

*[**FK** , Kurchan '07]*

Jamming = freezing



First order Transition

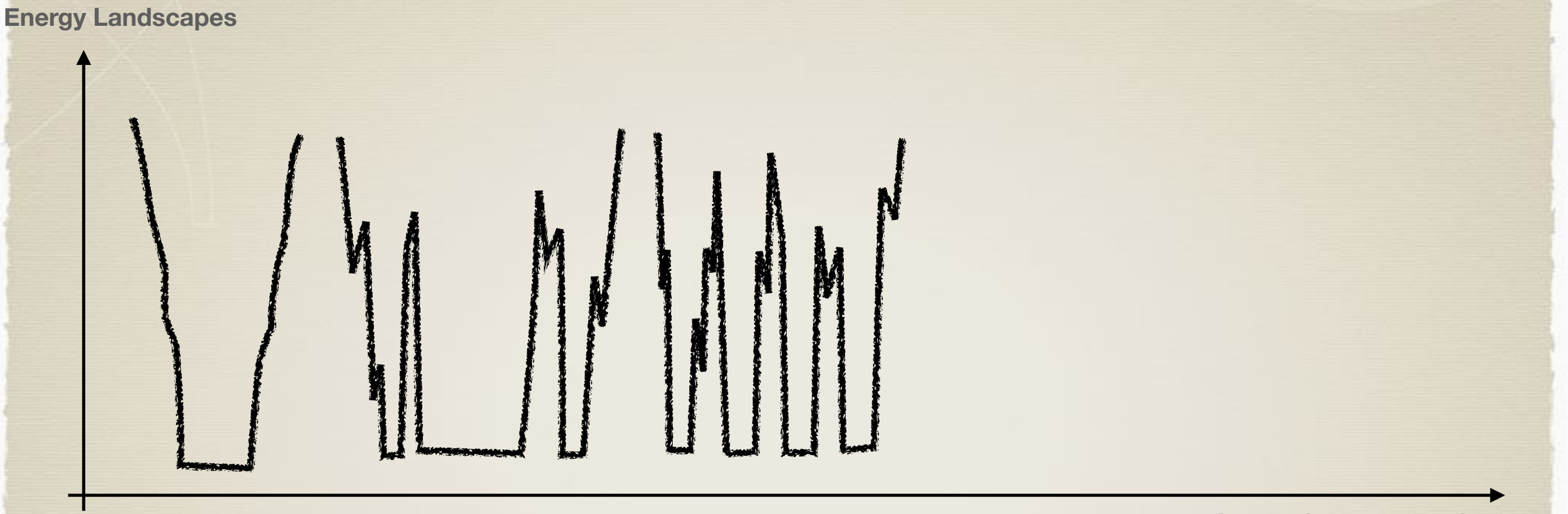


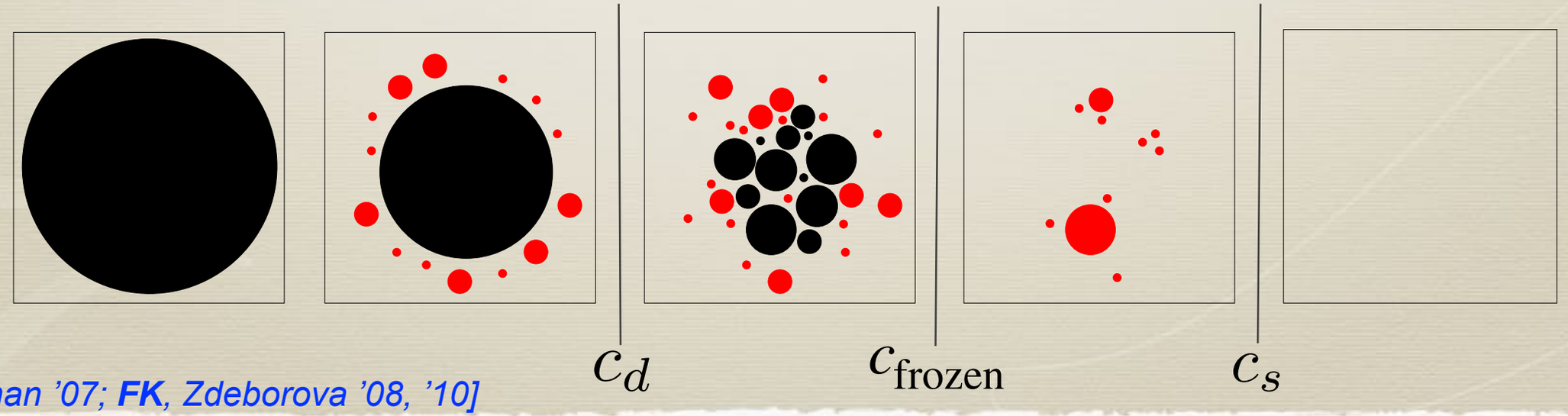
Second order Transition

3

Energy landscapes

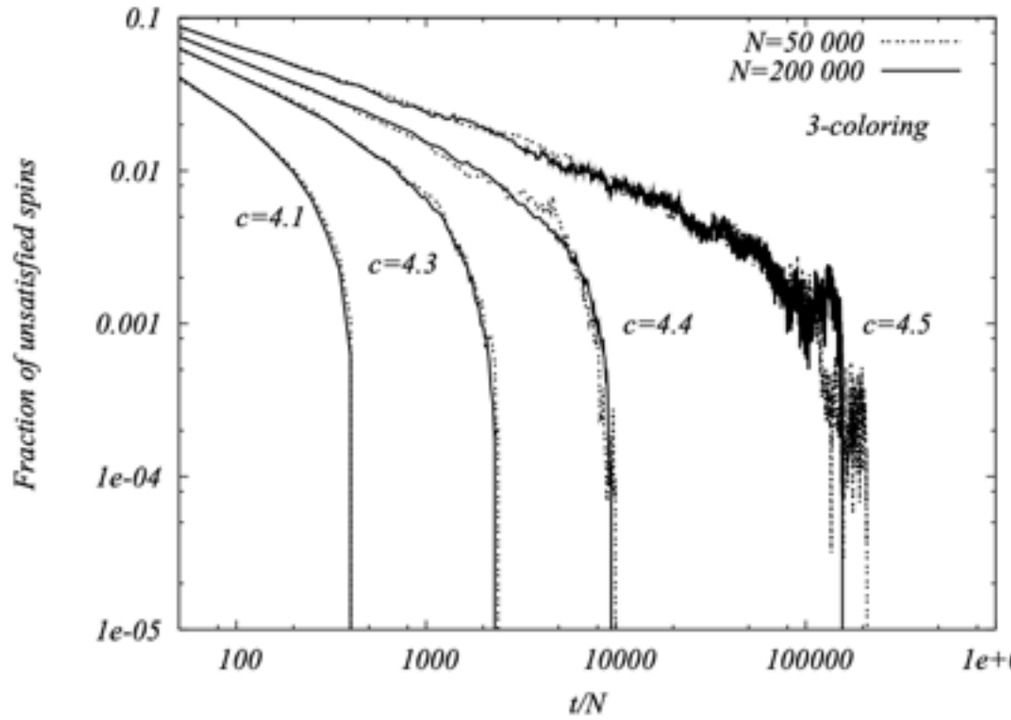
Energy landscapes





*[**FK** , Kurchan ’07; **FK**, Zdeborova ’08, ’10]*

Descent algorithms work just fine so far!



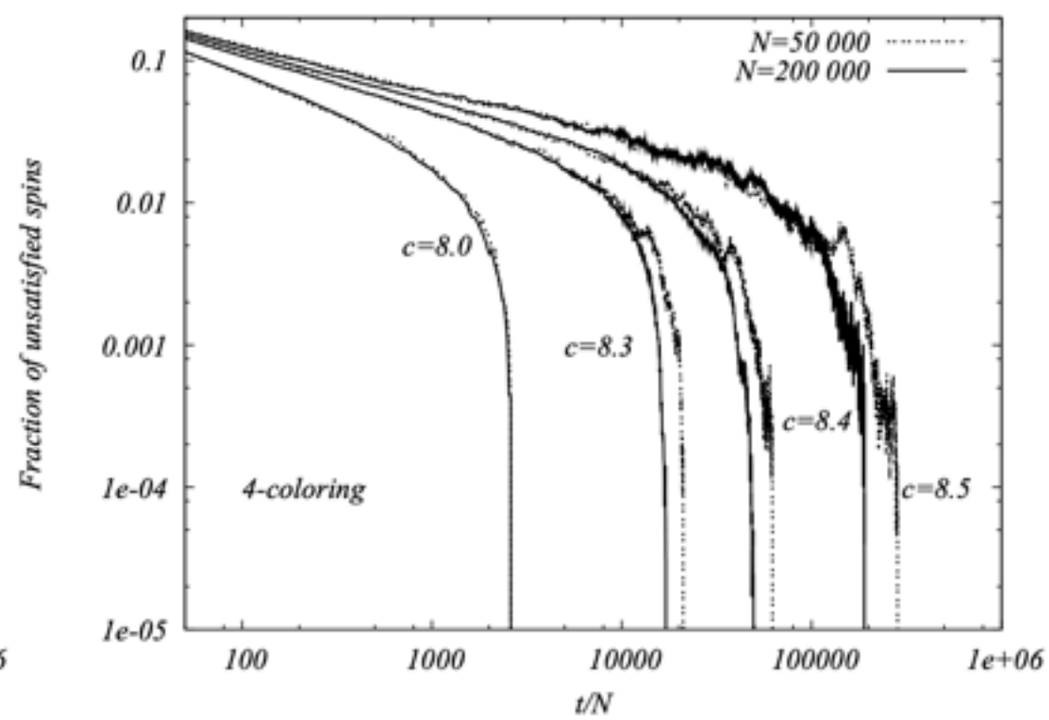
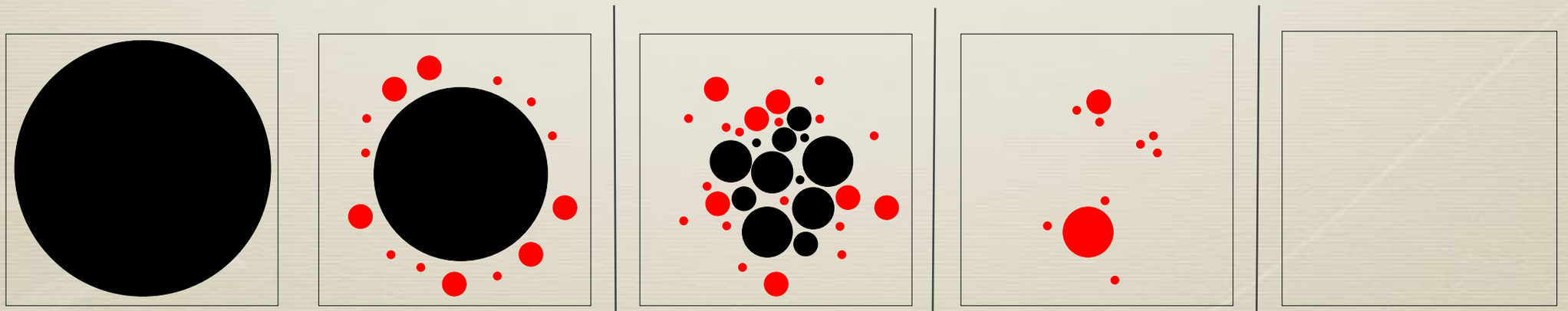
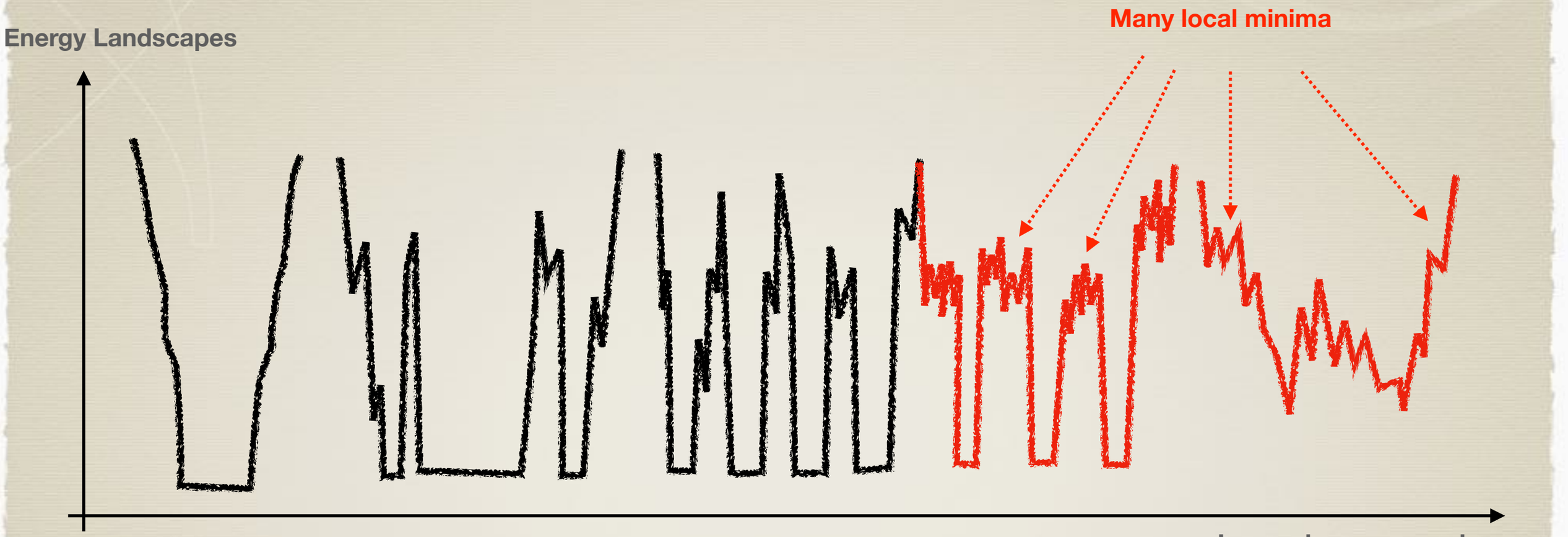


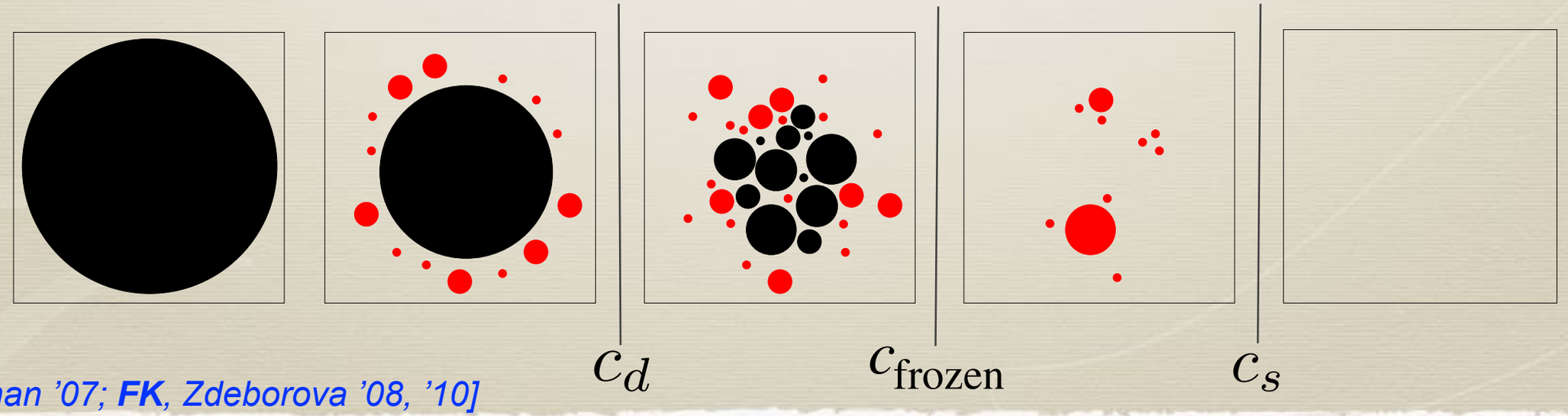
FIG. 10: Performance of the Walk-COL algorithm in coloring random graphs for 3−coloring (left) and 4−coloring (right). We plot the rescaled time (averaged over 5 instances) needed to color a graph of connectivity c. The strategy allows one to go beyond the clustering transition ($c_d = 4$ for 3-coloring and $c_d = 8.35$ for 4-coloring) in linear time with respect to the size of the graph.



*[**FK** , Kurchan '07; **FK**, Zdeborova '08, '10]*

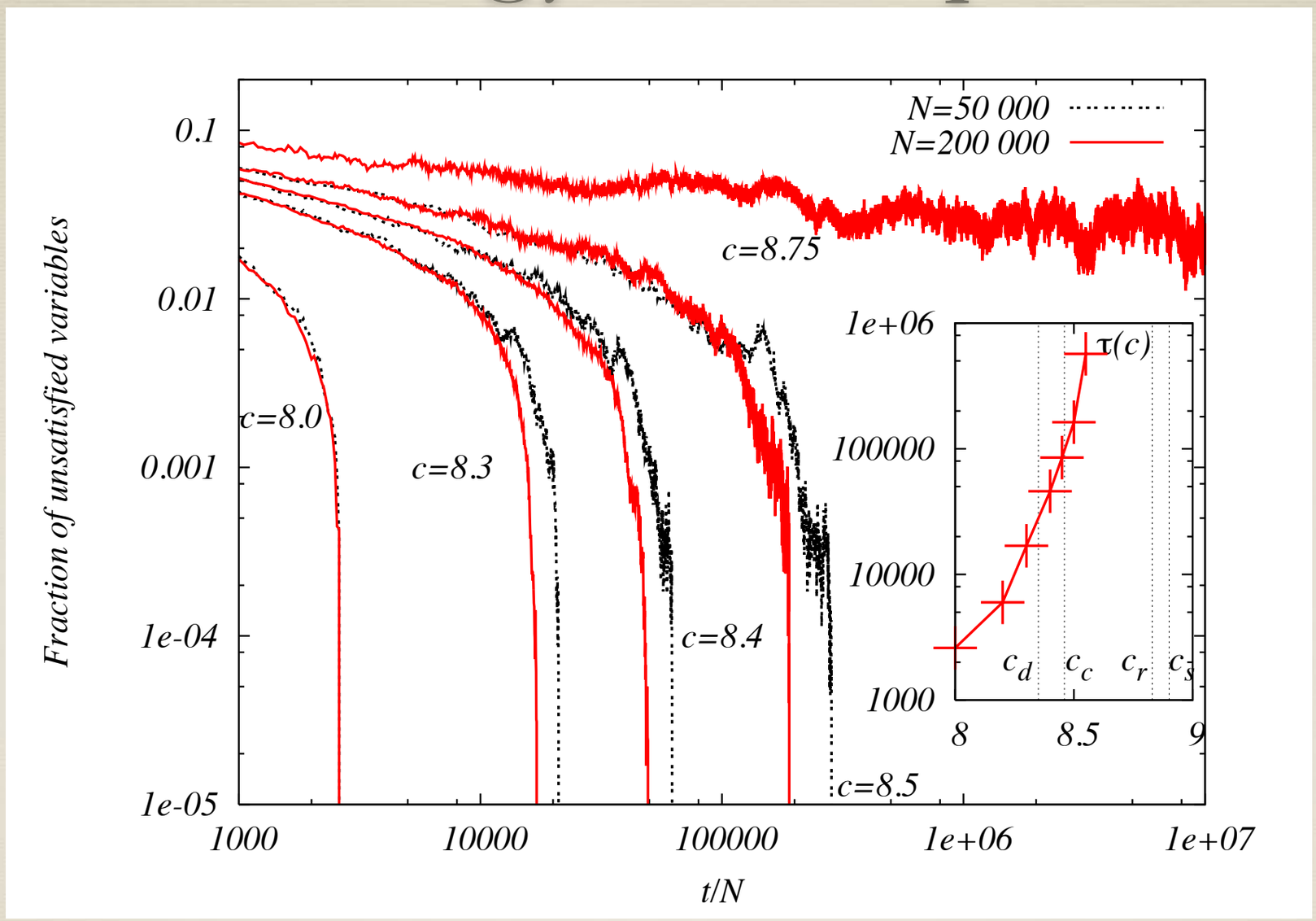
Energy landscapes

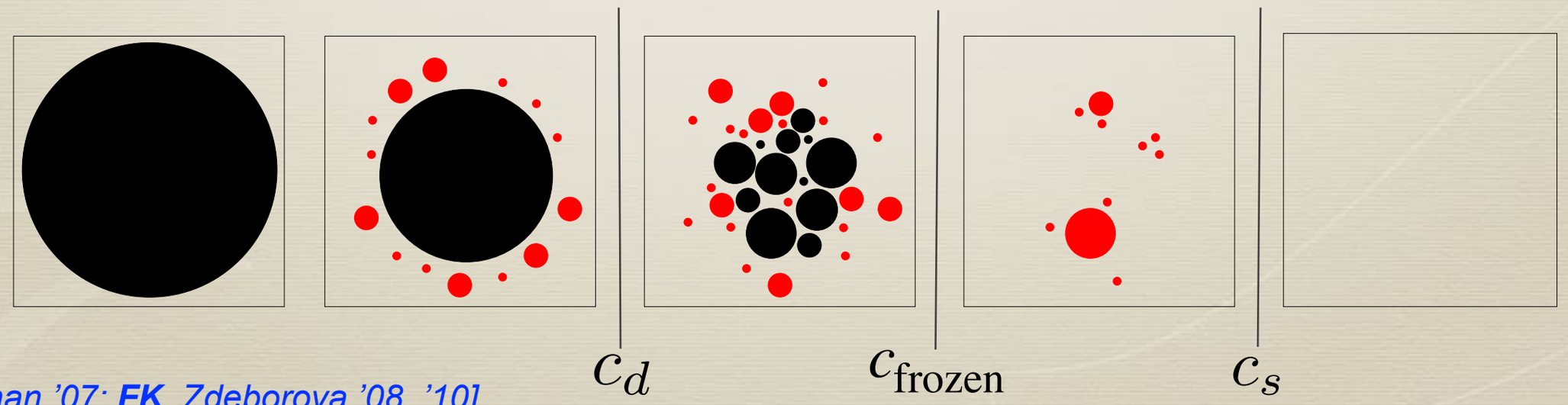




*[**FK** , Kurchan '07; **FK**, Zdeborova '08, '10]*

Energy landscapes





*[**FK** , Kurchan '07; **FK**, Zdeborova '08, '10]*

Rigid solutions are hard to find

Many polynomial algorithm works

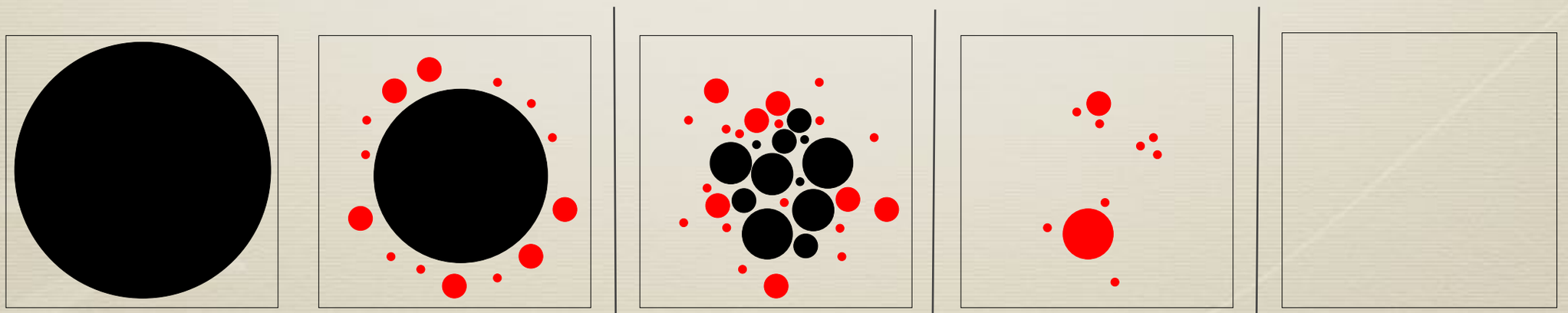
Nothing works

(We tried rigorous algorithms, empirical ones, survey propagation, many clever and many stupid ideas & even quantum computing….)

Two major open problems:

*** Can one design an algorithm that find "rigid","frozen" or "jammed" solution?**

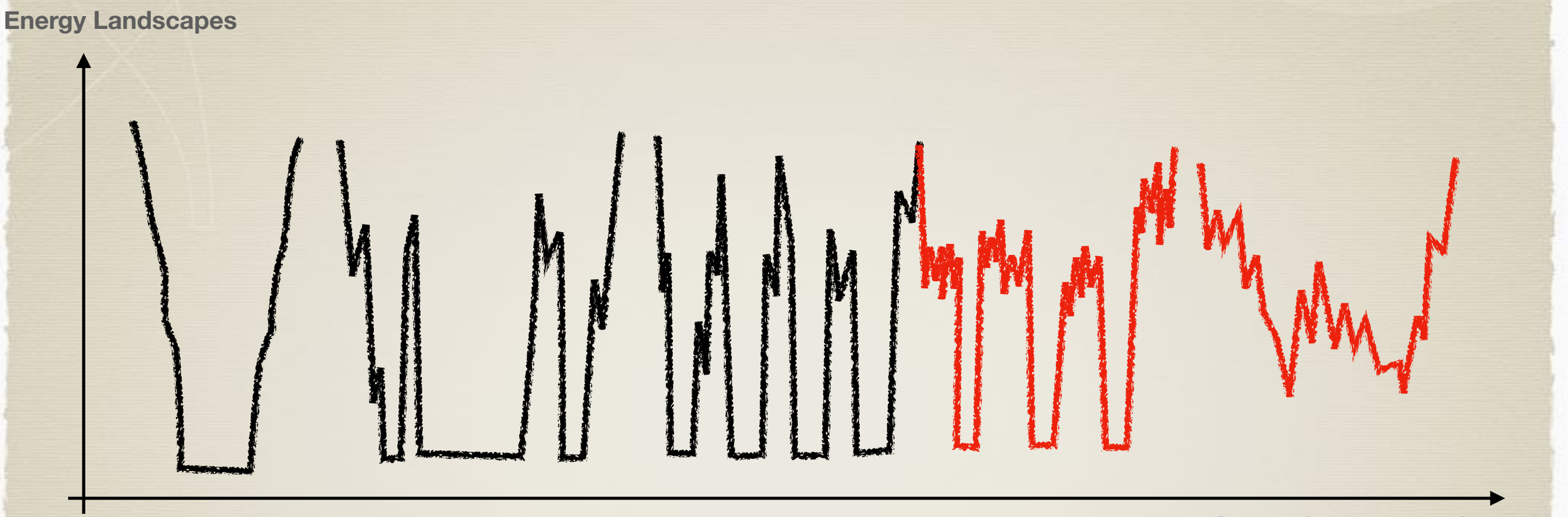
*** Can one find a solution to q colouring for connectivity larger than q log(q)?**



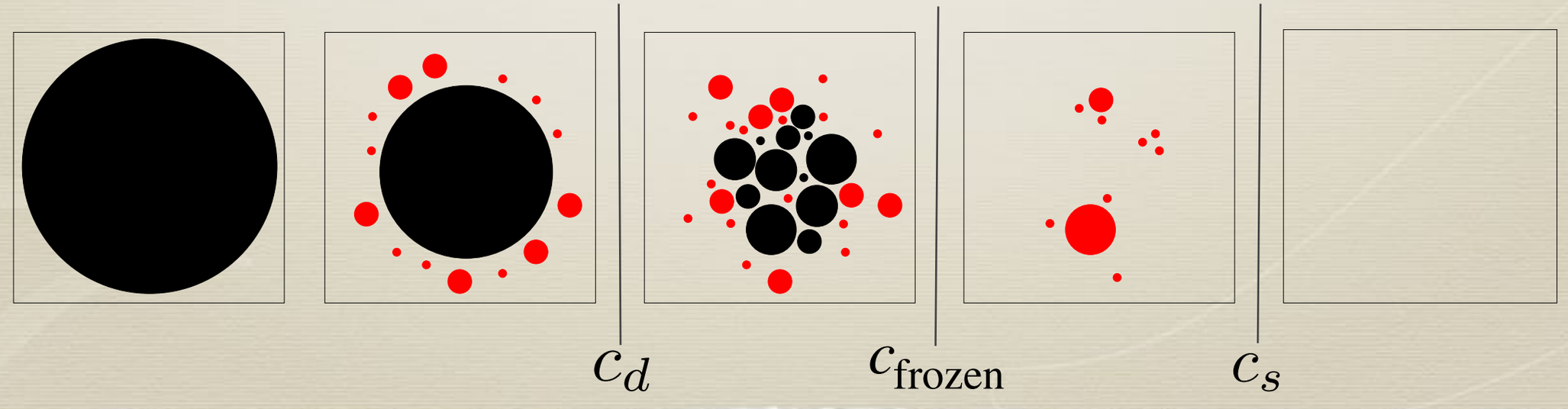
*[**FK** , Kurchan '07; **FK**, Zdeborova '08, '10]*

Back to machine learning

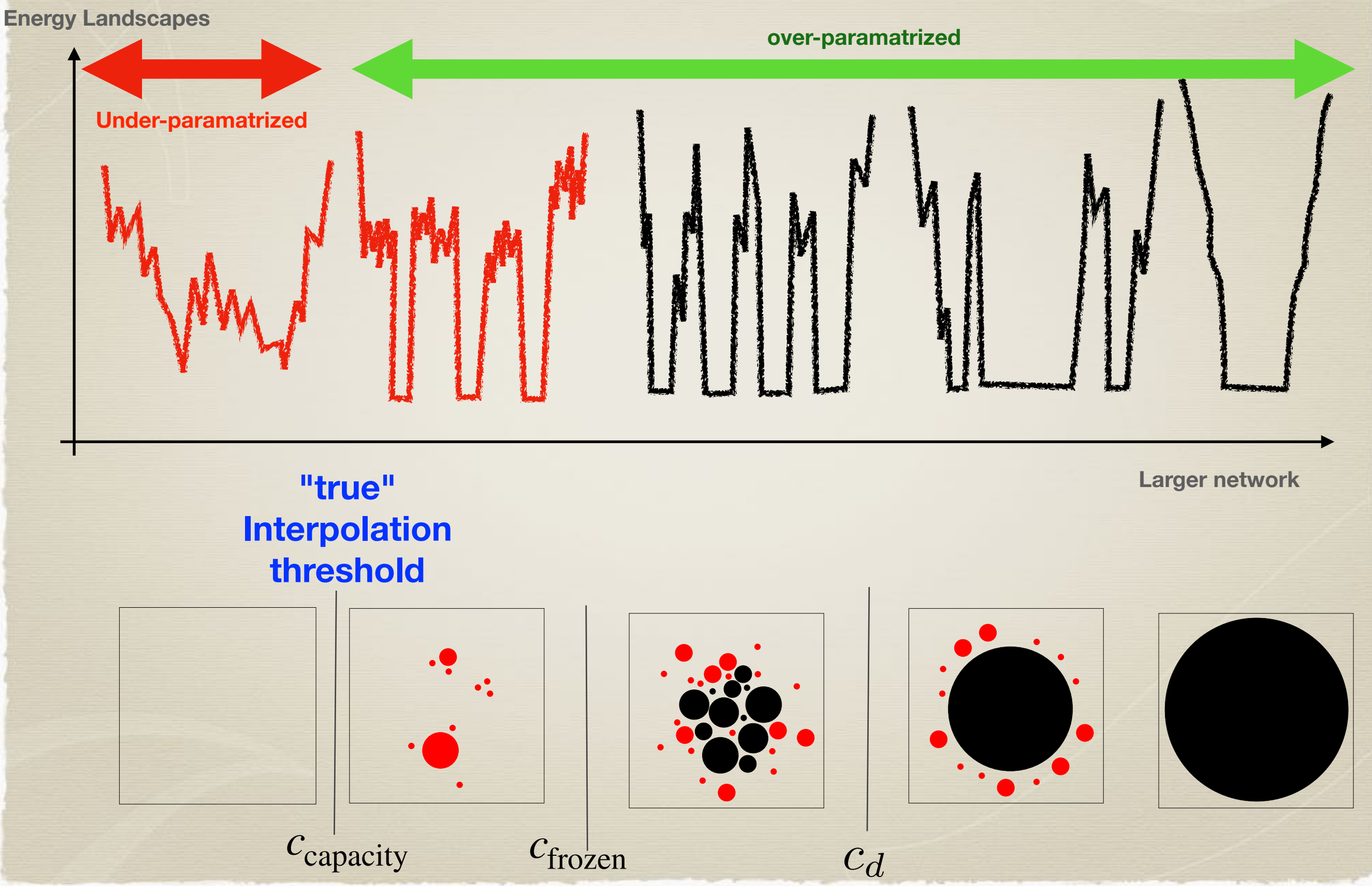
Energy landscapes



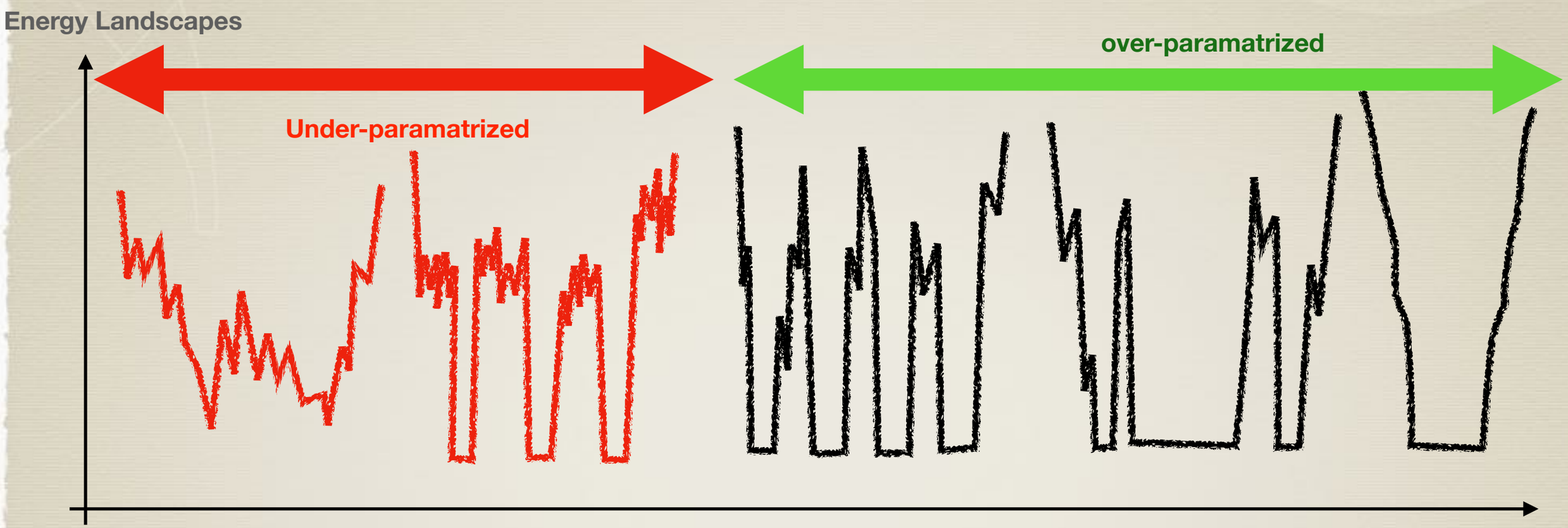
Smaller network



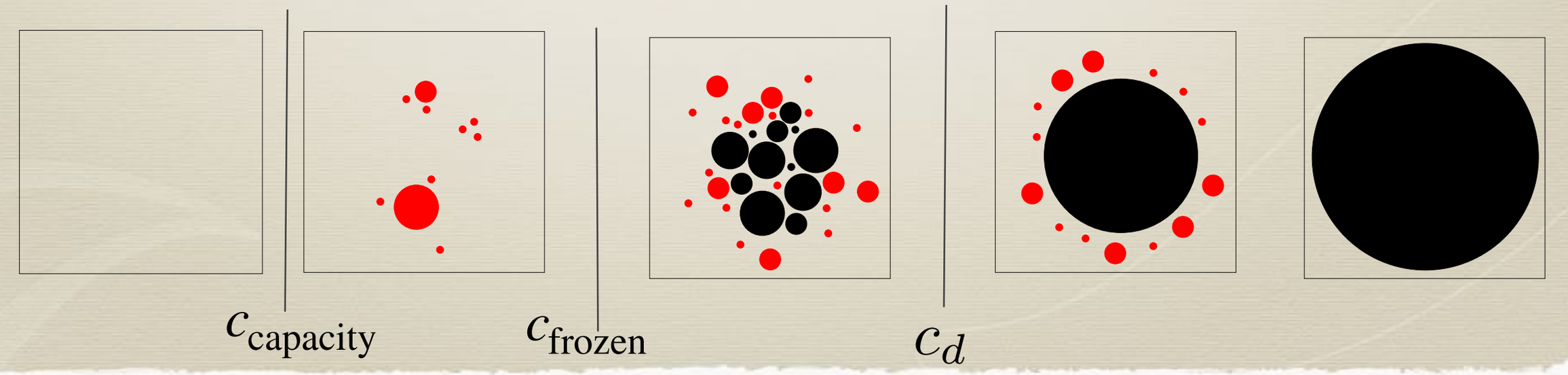
Energy landscapes



Energy landscapes



"effective"
Interpolation
threshold



Energy landscapes

Comparing Dynamics: Deep Neural Networks versus Glassy Systems

Marco Baity-Jesi [1] Levent Sagun [2 3] Mario Geiger [3] Stefano Spigler [3 2] Gérard Ben Arous [4]
Chiara Cammarota [5] Yann LeCun [4 6 7] Matthieu Wyart [3] Giulio Biroli [2 8]

PAPER

A jamming transition from under- to over-parametrization affects generalization in deep learning

S Spigler[1,3,4], M Geiger[1,3], S d'Ascoli[2], L Sagun[1], G Biroli[2] and M Wyart[1]

Published 28 October 2019 • © 2019 IOP Publishing Ltd

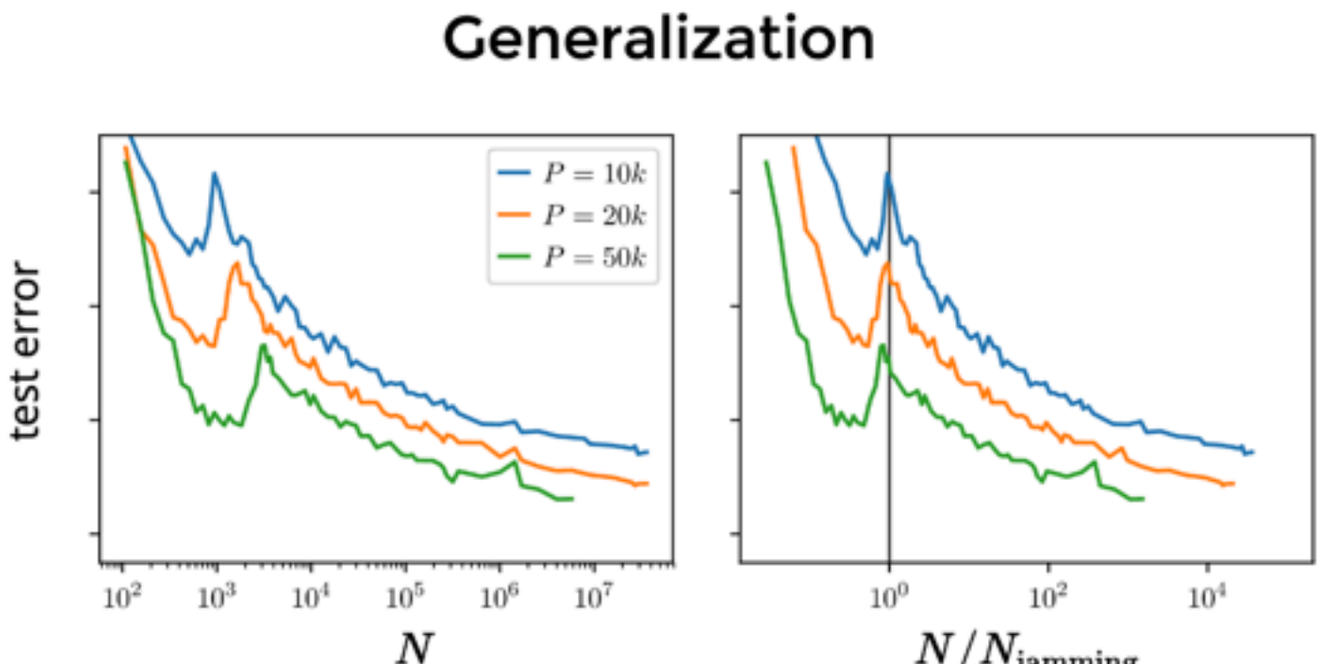
Journal of Physics A: Mathematical and Theoretical, Volume 52, Number 47

Machine Learning and Statistical Physics: Theory, Inspiration, Application

The jamming transition as a paradigm to understand the loss landscape of de

Fig

"Double Descent" behaviour @ the algorithmic interpolation threshold



Other interesting ideas?

- All solutions are connected in super-dupper-over-parameterized networks

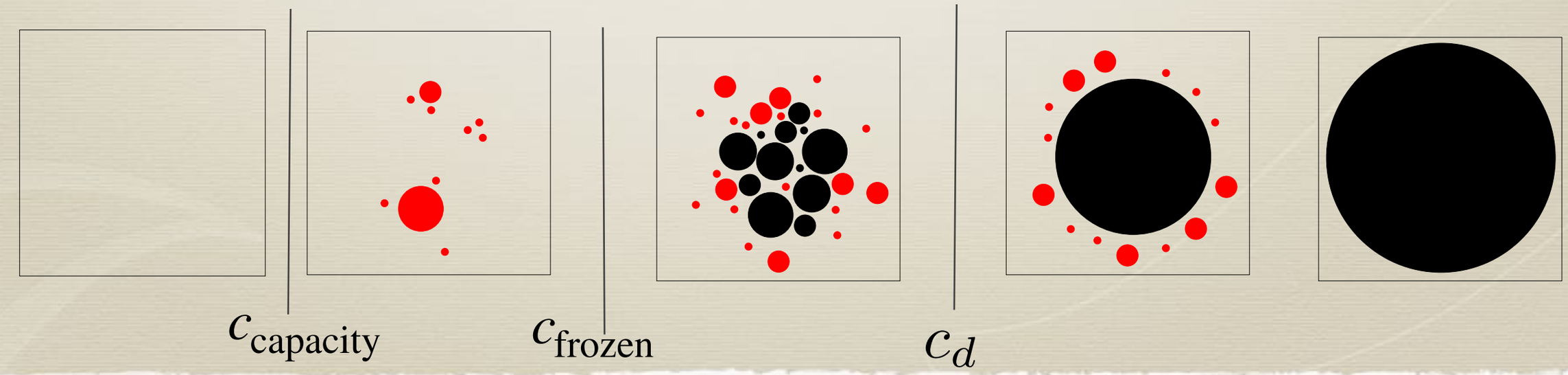
Freeman, C. D. & Bruna, J. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540* (2016).

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P. & Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, 8789–8798 (2018).

Draxler, F., Veschgini, K., Salmhofer, M. & Hamprecht, F. A. Essentially no barriers in neural network energy landscape. *arXiv preprint arXiv:1803.00885* (2018).

Kuditipudi, R. *et al.* Explaining landscape connectivity of low-cost solutions for multilayer nets. In *Advances in Neural Information Processing Systems*, 14574–14583 (2019).

Larger network



Other interesting ideas?

C ...over-parameterized networks

C ...olutions should generalise well!

Bad Global Minima Exist and SGD Can Reach Them

Shengchao Liu, Dimitris Papailiopoulos
University of Wisconsin–Madison

Dimitris Achlioptas
University of California, Santa Cruz

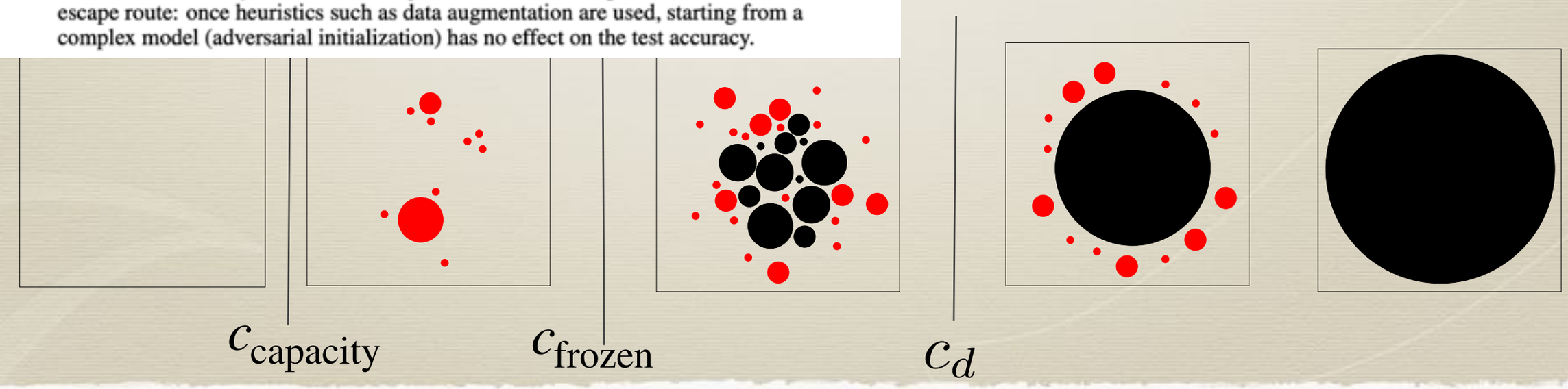
Abstract

Several recent works have aimed to explain why severely overparameterized models, generalize well when trained by Stochastic Gradient Descent (SGD). The emergent consensus explanation has two parts: the first is that there are "no bad local minima", while the second is that SGD performs implicit regularization by having a bias towards low complexity models. We revisit both of th... of image classification with common deep neural network... finding is that there exist bad *global* minima, *i.e.*, models ... perfectly, yet have poor generalization. Our second find... *unlabeled* training data, we can easily construct initializatio... to quickly converge to such bad global minima. For example... and (Restricted) ImageNet, this can be achieved by starting S... by fitting random labels on the training data: while subsequ... the correct labels) will reach zero training error, the result... a test accuracy degradation of up to 40% compared to training from a random initialization. Finally, we show that regularization seems to provide SGD with an escape route: once heuristics such as data augmentation are used, starting from a complex model (adversarial initialization) has no effect on the test accuracy.

Recipe for bad solutions:

1) Train your data with random labels until zero errors
2) Slowly put back the correct label, train to zero error

This yield rare, very bad solutions!



Other interesting ideas?

- All solutions are connect
- The set of solutions can

Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes

Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina

Check for updates

lished November 15, 2016

?

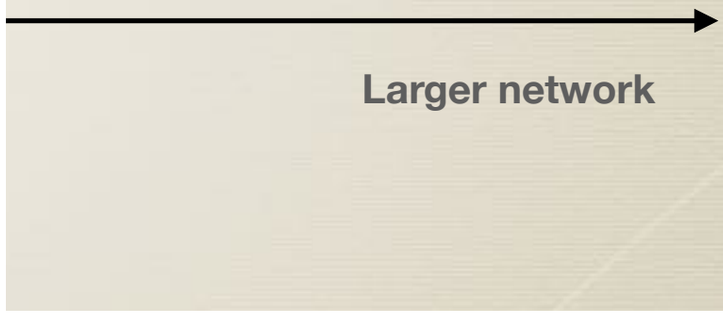
Shaping the learning landscape in neural networks around wide flat minima

Carlo Baldassi, Fabrizio Pittorino, and Riccardo Zecchina

PNAS January 7, 2020 117 (1) 161-170; first published December 23, 2019 https://doi.org/10.1073/pnas.1908636117

Edited by Yuhai Tu, IBM, Yorktown Heights, NY, and accepted by Editorial Board Member Herbert Levine November 20, 2019 (received for review May 20, 2019)

Article | Figures & SI | Info & Metrics | PDF



ML 2019

Entropy-SGD: biasing gradient descent into wide valleys*

Pratik Chaudhari[1,2], Anna Choromanska[3], Stefano Soatto[1], Yann LeCun[4,5], Carlo Baldassi[6,7], Christian Borgs[8], Jennifer Chayes[8], Levent Sagun[4] and Riccardo Zecchina[6,7]

Published 20 December 2019 • © 2019 IOP Publishing Ltd and SISSA Medialab srl

Journal of Statistical Mechanics: Theory and Experiment, Volume 2019, December 2019

Machine Learning 2019

A word of conclusion

Physics:

- glasses
- electron glasses
- hard spheres...

Computer science:

- constraint satisfaction problems

Neural networks:

- Hopfield model
- Energy-based models
- Deep learning

Information theory:

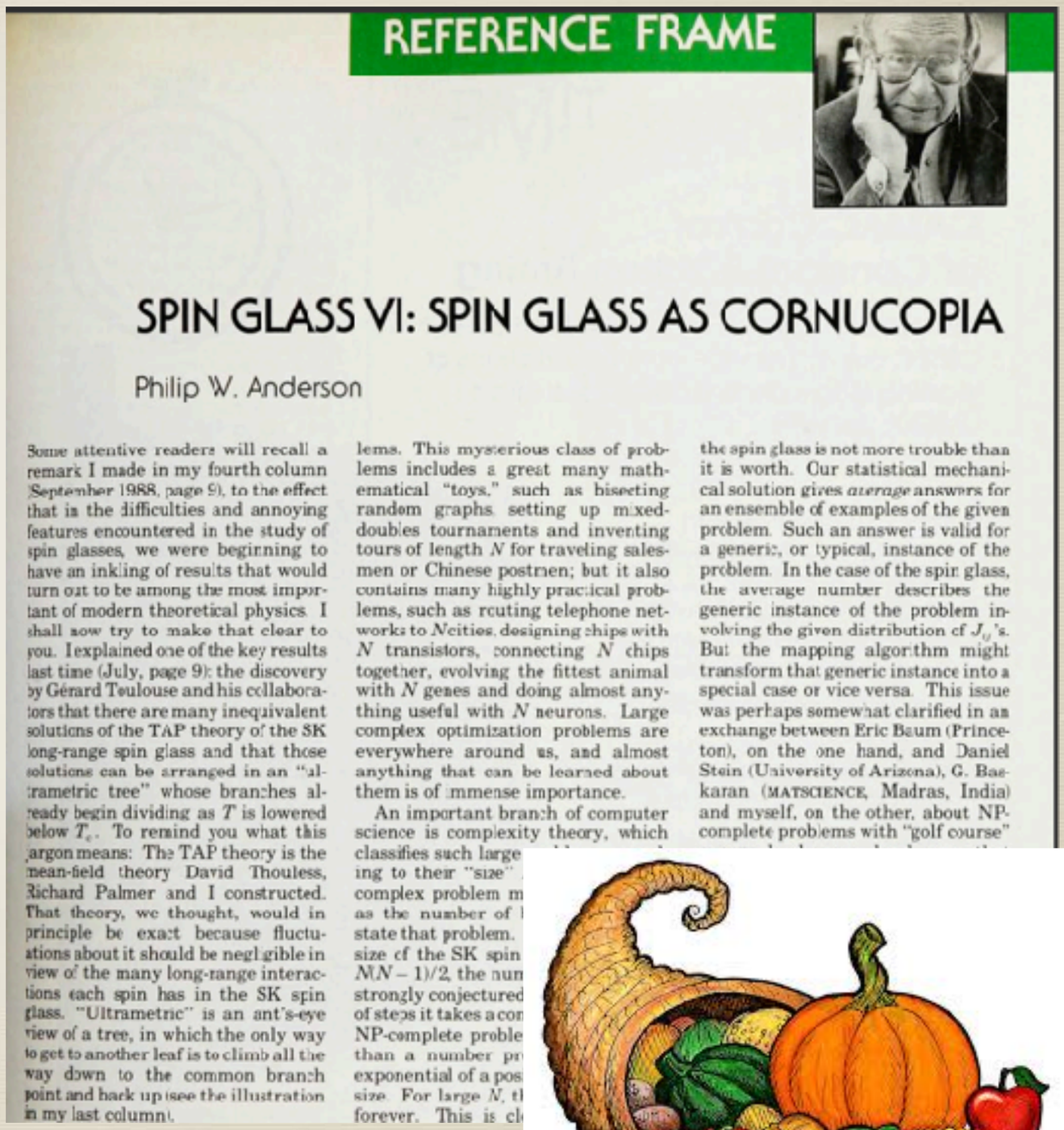
- Error correcting codes

Signal processing:

- Compressed sensing

Finance

Statistics, inference and machine learning



REFERENCE FRAME

SPIN GLASS VI: SPIN GLASS AS CORNUCOPIA

Philip W. Anderson

Some attentive readers will recall a remark I made in my fourth column (September 1988, page 9), to the effect that is the difficulties and annoying features encountered in the study of spin glasses, we were beginning to have an inkling of results that would turn out to be among the most important of modern theoretical physics. I shall now try to make that clear to you. I explained one of the key results last time (July, page 9): the discovery by Gérard Toulouse and his collaborators that there are many inequivalent solutions of the TAP theory of the SK long-range spin glass and that those solutions can be arranged in an "ultrametric tree" whose branches already begin dividing as T is lowered below T_c. To remind you what this jargon means: The TAP theory is the mean-field theory David Thouless, Richard Palmer and I constructed. That theory, we thought, would in principle be exact because fluctuations about it should be negligible in view of the many long-range interactions each spin has in the SK spin glass. "Ultrametric" is an ant's-eye view of a tree, in which the only way to get to another leaf is to climb all the way down to the common branch point and back up (see the illustration in my last column).

lems. This mysterious class of problems includes a great many mathematical "toys," such as bisecting random graphs, setting up mixed-doubles tournaments and inventing tours of length N for traveling salesmen or Chinese postmen; but it also contains many highly practical problems, such as routing telephone networks to N cities, designing chips with N transistors, connecting N chips together, evolving the fittest animal with N genes and doing almost anything useful with N neurons. Large complex optimization problems are everywhere around us, and almost anything that can be learned about them is of immense importance.

An important branch of computer science is complexity theory, which classifies such large
ing to their "size"
complex problem m
as the number of
state that problem.
size of the SK spin
$MN-1)/2$, the nun
strongly conjectured
of steps it takes a cor
NP-complete proble
than a number pr
exponential of a pos
size. For large N, t
forever. This is cl

the spin glass is not more trouble than it is worth. Our statistical mechanical solution gives *average* answers for an ensemble of examples of the given problem. Such an answer is valid for a generic, or typical, instance of the problem. In the case of the spin glass, the average number describes the generic instance of the problem involving the given distribution of J_{ij}'s. But the mapping algorithm might transform that generic instance into a special case or vice versa. This issue was perhaps somewhat clarified in an exchange between Eric Baum (Princeton), on the one hand, and Daniel Stein (University of Arizona), G. Baskaran (MATSCIENCE, Madras, India) and myself, on the other, about NP-complete problems with "golf course"

We want you in Switzerland!

Lenka and I are looking for talented postdocs & students in EPFL