THE PHYSICS OF COUNTING AND SAMPLING ON RANDOM INSTANCES

Lenka Zdeborová

(CEA Saclay and CNRS, France)

MAIN CONTRIBUTORS TO THE PHYSICS UNDERSTANDING OF RANDOM INSTANCES

- Braunstein, Franz, Kabashima, Kirkpatrick, Krzakala, Monasson, Montanari, Nishimori, Pagnani, Parisi, Ricci-Tersenghi, Saad, Semerjian, Sherrington, Sourlas, Sompolinsky, Tanaka, Weigt, Zdeborova, Zecchina, ...

WHY RANDOM?

also Andrea's talk

- First step towards “typical” instances.
- Intriguing mathematical properties
- Solvable (by theoretical physics standards, mean field ...). Using belief propagation, Bethe approximation and its extensions, cavity method, replica symmetry breaking.
- Ideas, inspiration and benchmarks for algorithms

THE PICTURE

- This talk's example: Graph coloring on random graphs.
- Resulting picture relevant for: satisfiability, CSP, vertex cover, independent sets, max-cut, ... error correcting codes, sparse estimation, regression, clustering, compressed sensing, feature learning, neural networks, ...

GRAPH COLORING

- How many proper colorings on a large random graph?
- Can they be sampled uniformly? MCMC properties?
- Variant 1: Finite temperature

$$\mu(\{s_i\}_{i=1,\dots,N}) = \frac{1}{Z_G(\beta)} e^{-\beta \sum_{(ij)\in E} \delta_{s_i,s_j}}$$

- Variant 2: Planted graphs
  Fix a random string of colors $\{s_i^*\}_{i=1,\dots,N}$

$$s_i^* = s_j^* \quad \Rightarrow \quad (ij) \notin E$$

also Andrea's talk

COUNTING COLORINGS

Averaging and the large N limit.

$$|V| = N,\ |E| = M, \quad c = 2M/N \qquad\qquad c \text{ fixed}, N \to \infty,$$

- Annealed entropy

$$s_{\mathrm{ann}} = \lim_{N\to\infty} \frac{1}{N}[\log \mathbb{E}(Z_G)] = \log q + \frac{c}{2} \log 1 - \frac{1}{q}$$

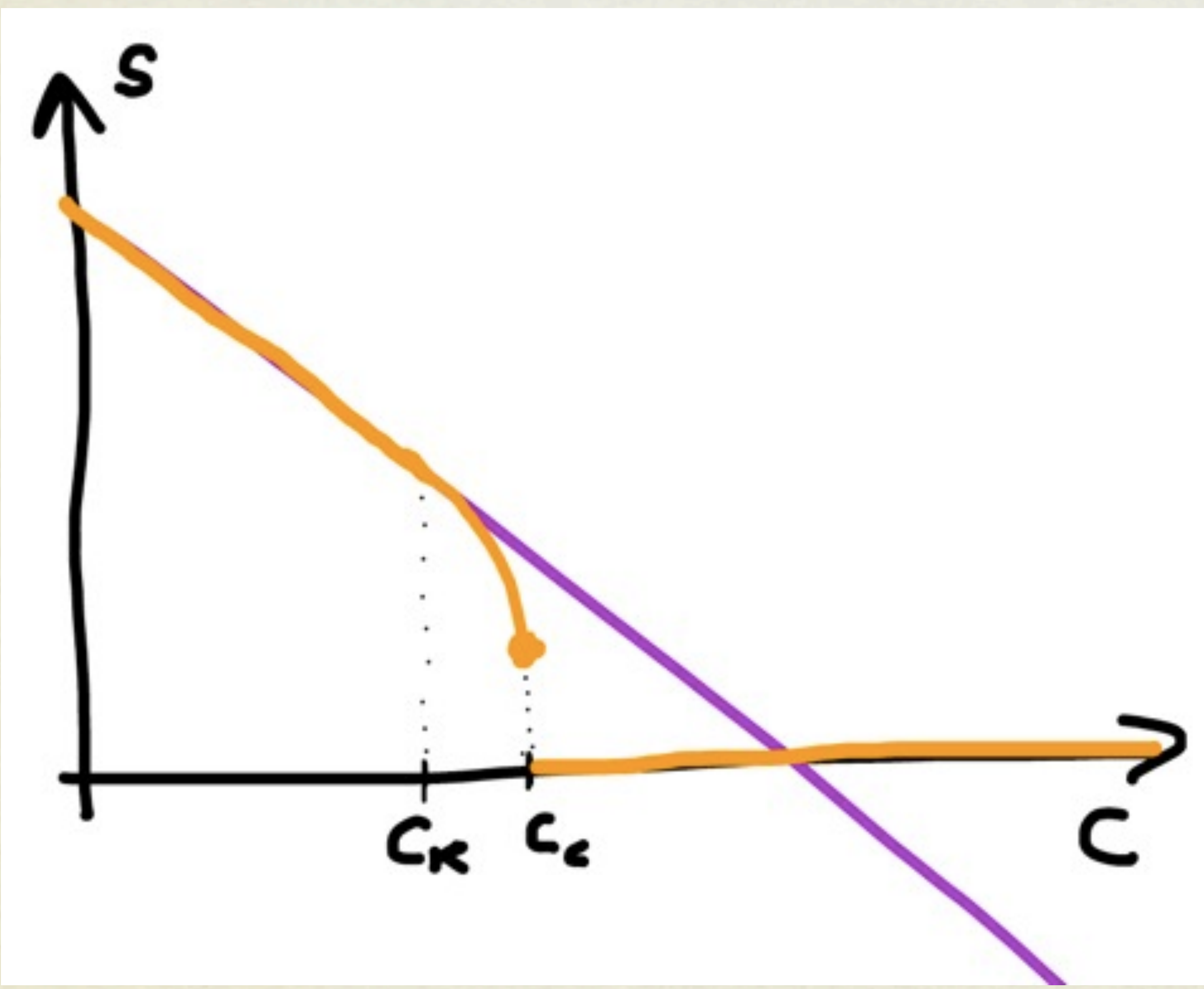
- Quenched entropy

$$s = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}[\log (Z_G + 1)]$$

COUNTING COLORINGS

$$|V| = N,\ |E| = M, \quad c = 2M/N \qquad\qquad c \text{ fixed}, N \to \infty,$$

$$s_{\text{ann}} = \lim_{N\to\infty} \frac{1}{N}[\log \mathbb{E}(Z_G)] = \log q + \frac{c}{2}\log{(1-\frac{1}{q})} \qquad s = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}[\log{(Z_G+1)}]$$



c_c: colorability threshold

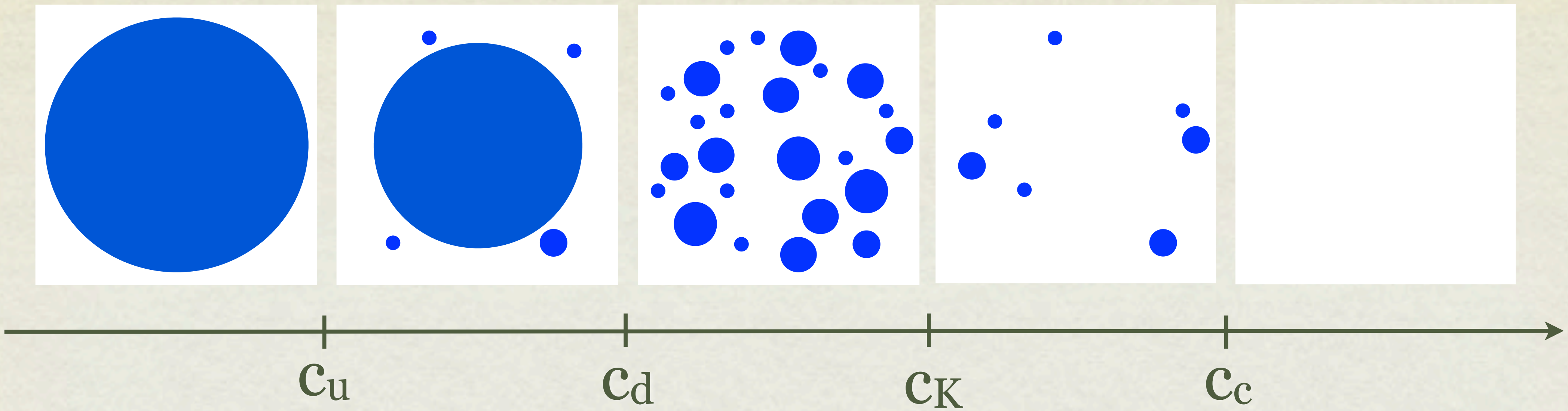
$$s = 0 \text{ for } c > c_c$$

c_K: Kauzmann/condensation

$$s < s_{\text{ann}} \text{ for } c > c_K$$

(Krzakala et al. PNAS'07)

PHASE TRANSITIONS



c_c: colorability threshold

c_K: Kauzmann/condensation transition

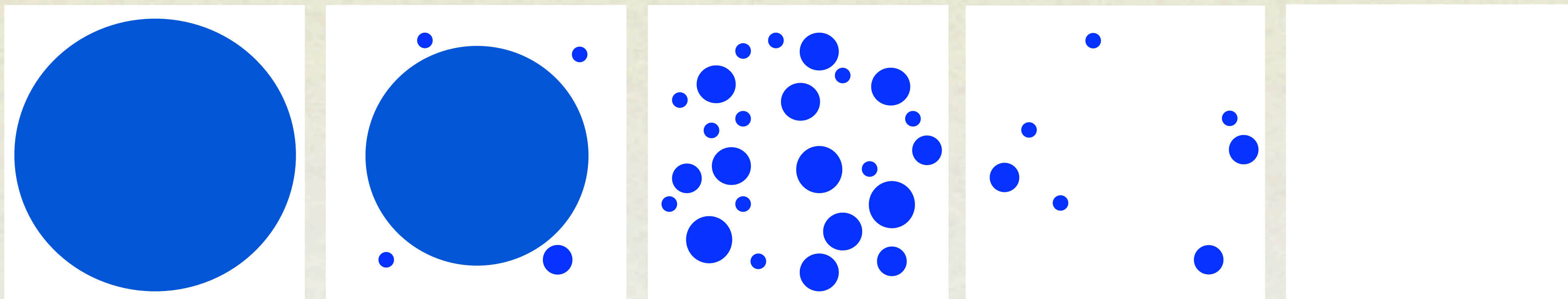
c_d: dynamical/clustering/reconstruction transition

c_u: unicity threshold

sometimes (e.g. 3-SAT or 3-coloring) $c_K = c_d$

SAMPLING

Recall: random graph, random or warm start, weaker convergence that total variation.



c_u c_d c_K c_c

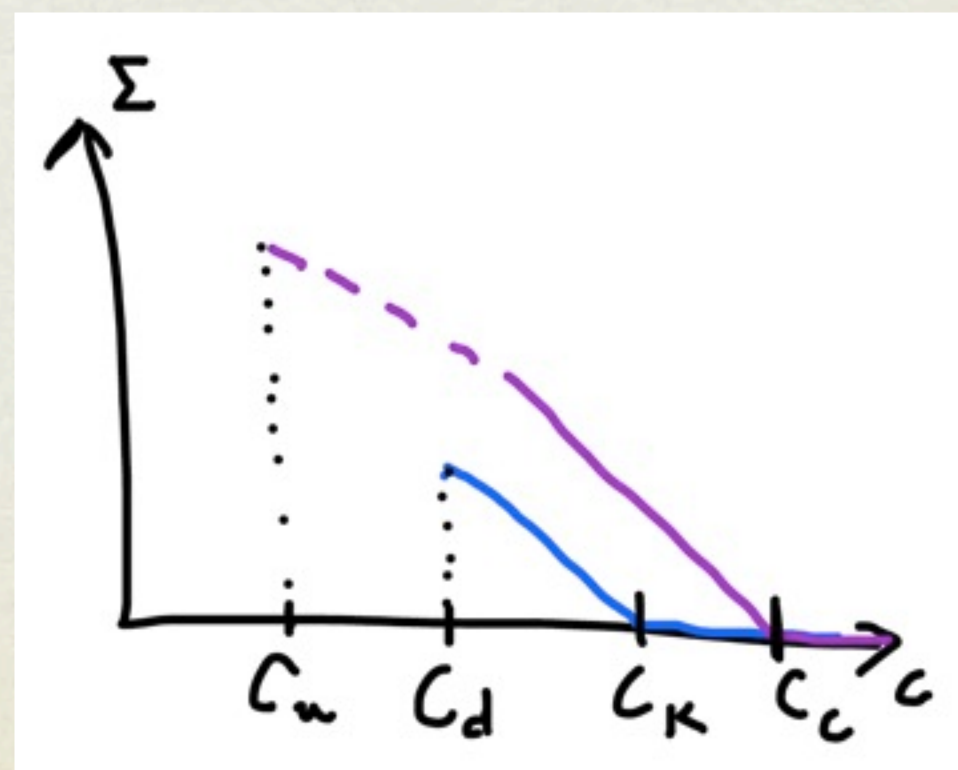
$q \to \infty$ $\approx q$ $\approx q \log q$ $\approx 2q \log q$

cluster = blue blob = subspace

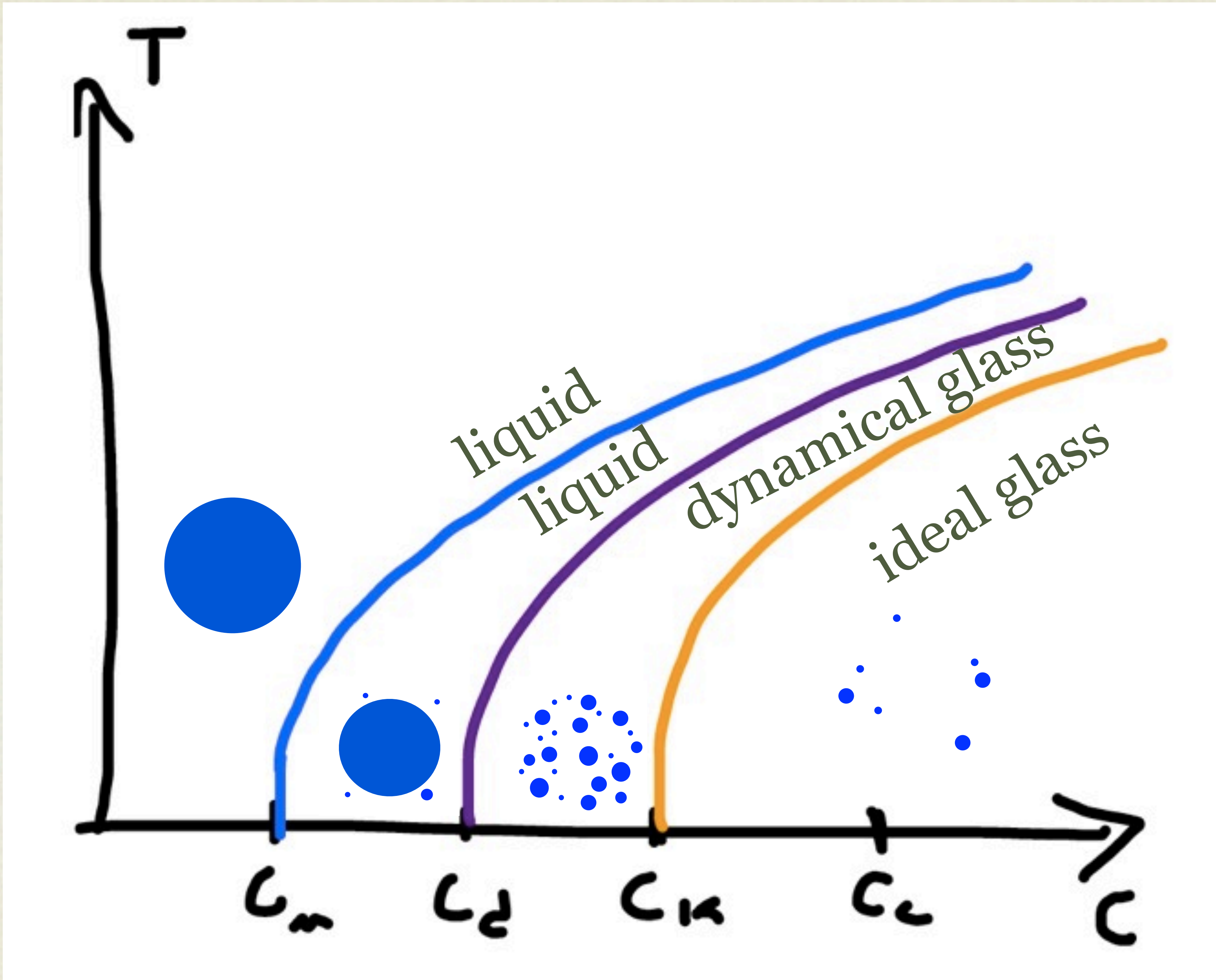
that MCMC samples

uniformly in linear time.

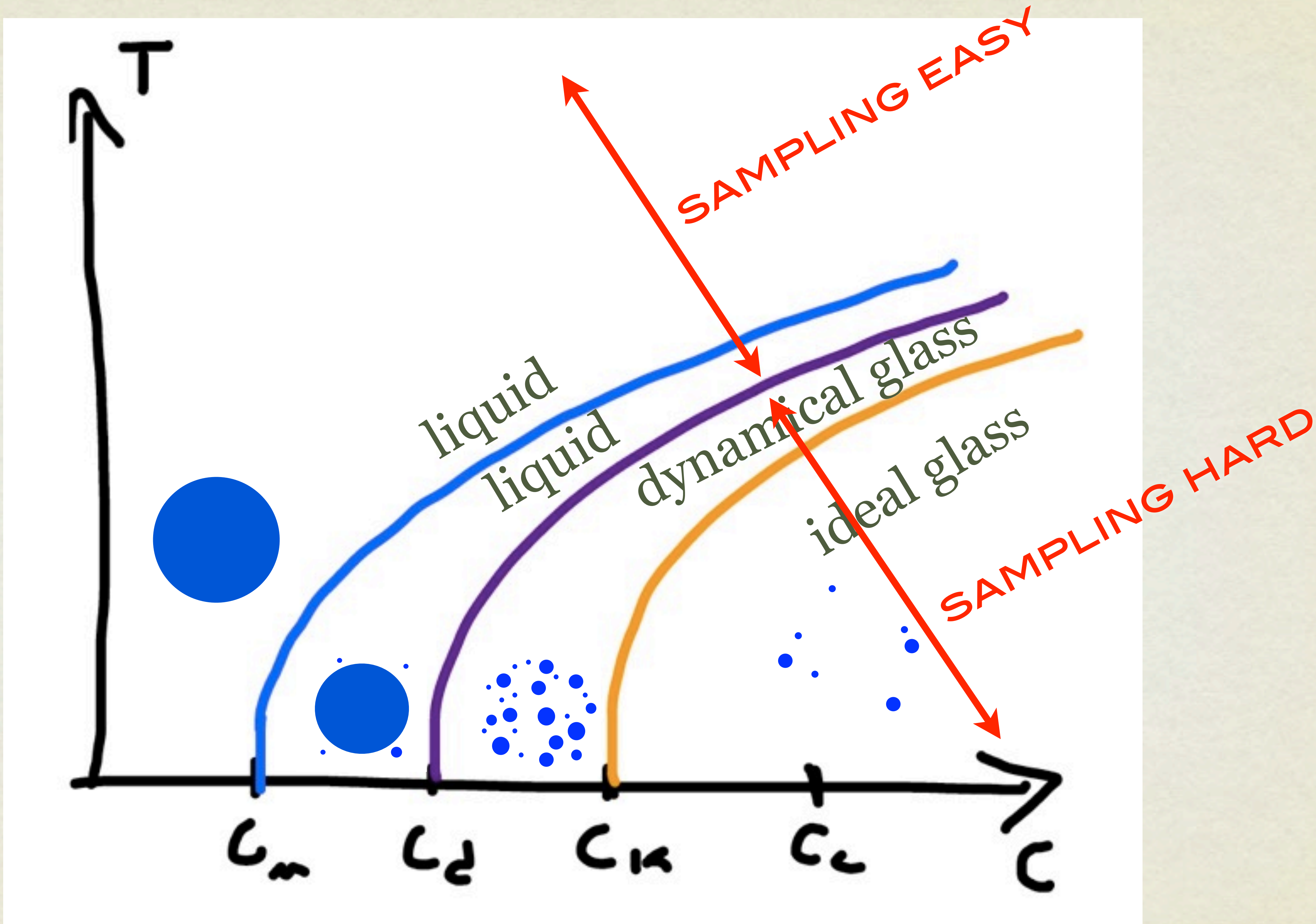
$$\Sigma = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}[\log(\#\text{clusters})]$$



TEMPERATURE DEPENDENCE



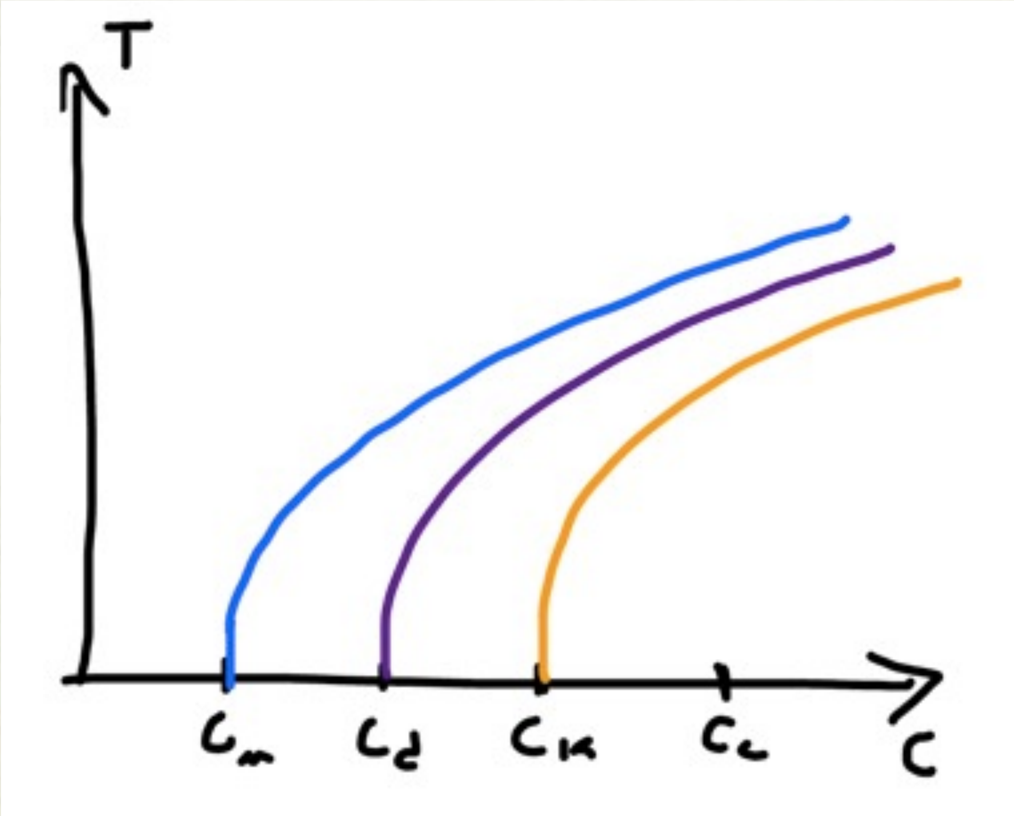
TEMPERATURE DEPENDENCE



WHAT IS A PHASE TRANSITION?

Phase transition always characterized by a divergence of a correlation length:

- point-to-set $T \searrow T_d$
- two-point $T \searrow T_K$

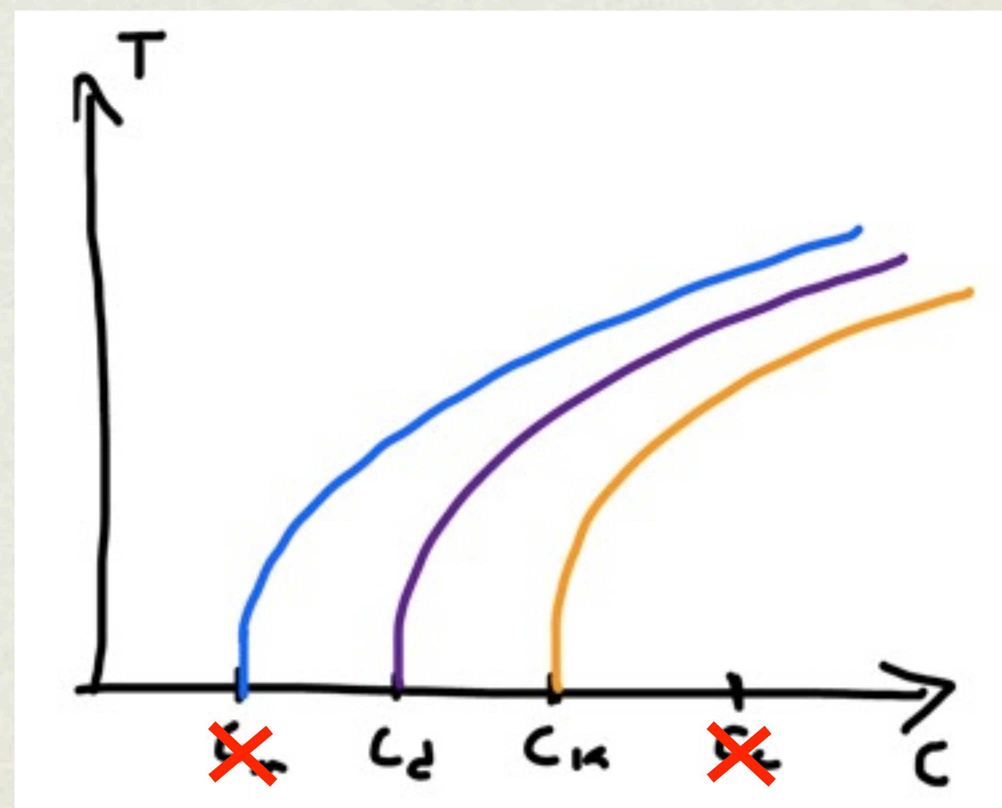


(Montanari, Semerjian'06)

WHAT IS A PHASE TRANSITION?

Phase transition always characterized by a divergence of a correlation length:

- point-to-set $T \searrow T_d$
- two-point $T \searrow T_K$



(Montanari, Semerjian'06)

PLANTED COLORING

also Andrea's talk

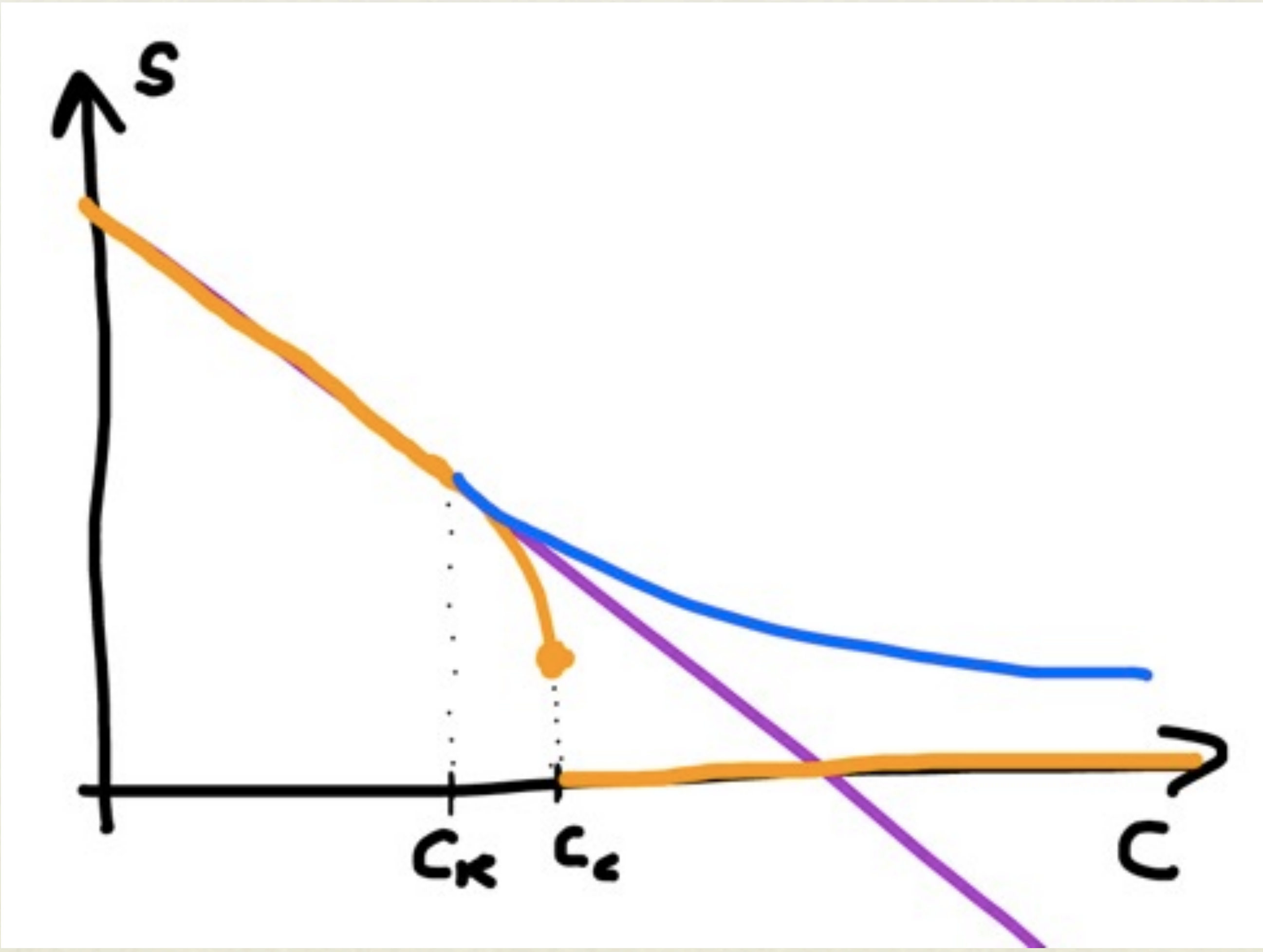
- Definition: Fix a random string $\{s_i^*\}_{i=1,\ldots,N}$ choose M edges at random such that if $s_i^* = s_j^* \quad \Rightarrow \quad (ij) \notin E$

- Interest n.1 = paradigm of statistical inference (95% of use of MCMC in computer science). Bayes optimal inference = computing marginals of the posterior distribution = approximate counting problem.

- Interest n.2 = Math simpler. “Warm start” for MCMC for free.

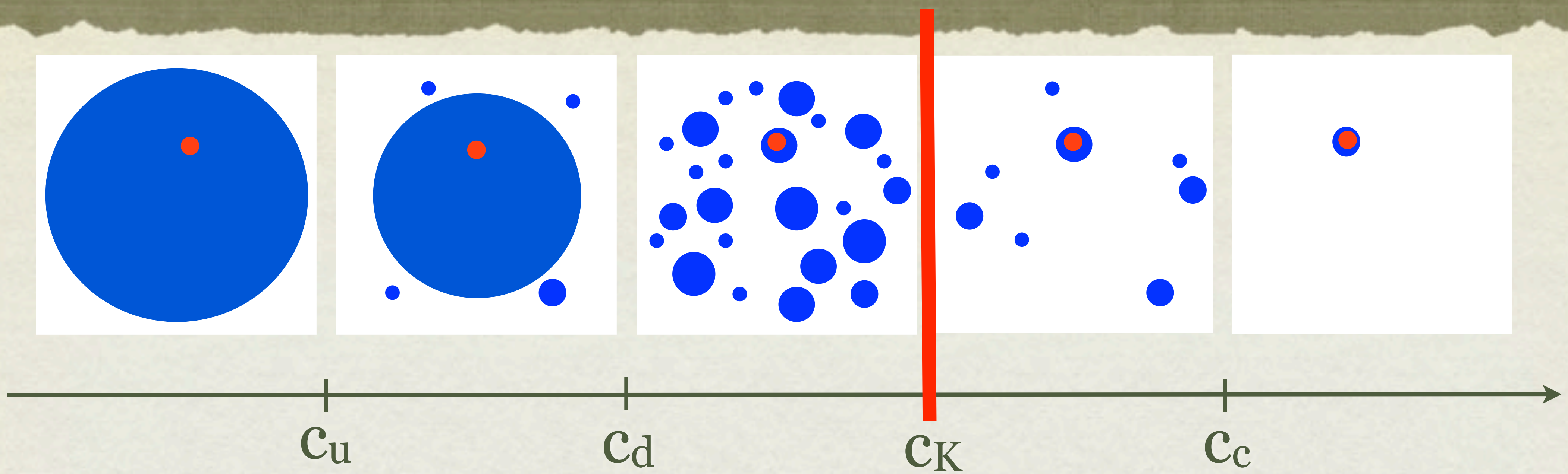
PLANTED COLORING

- Definition: Fix a random string $\{s_i^*\}_{i=1,\ldots,N}$ choose M edges at random such that if $s_i^* = s_j^* \quad \Rightarrow \quad (ij) \notin E$



(Krzakala, LZ'09)

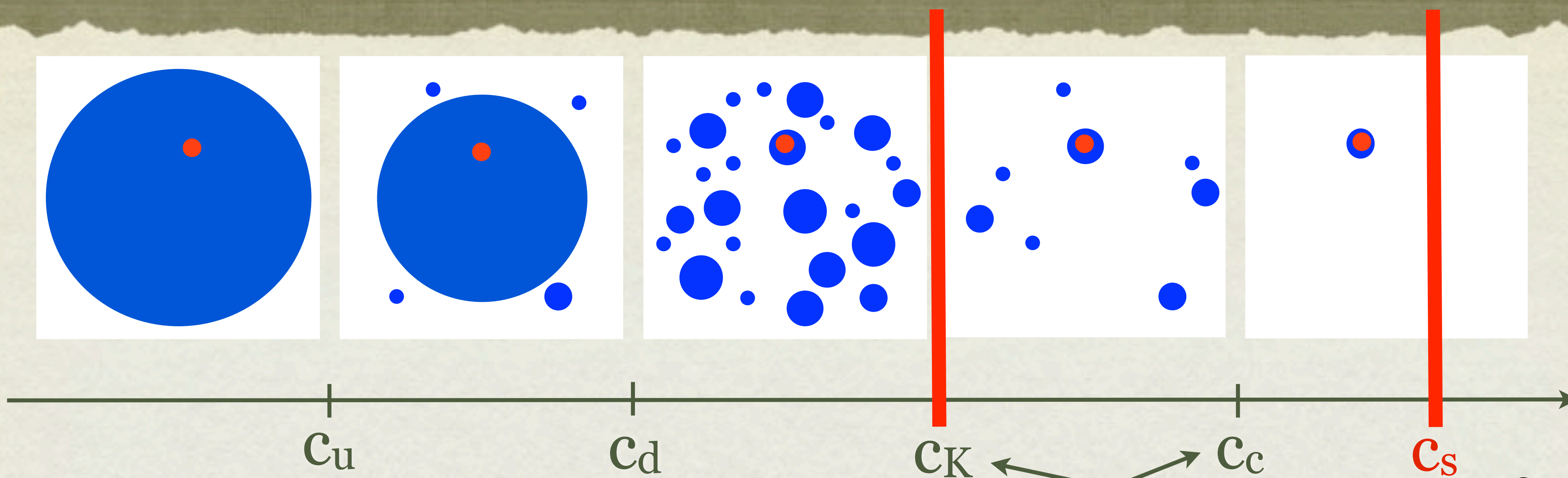
PLANTED PHASE TRANSITIONS



- For $c<c_K$ random and planted contiguous, statistical inference impossible, planted configuration is an equilibrium one.

- In cases for which $c_{K=} c_d$ (e.g. 3-col, 3-sat) we have that for $c>c_K$ statistical inference is easy. MCMC becomes fast correlated to the planted configuration.

PLANTED PHASE TRANSITIONS



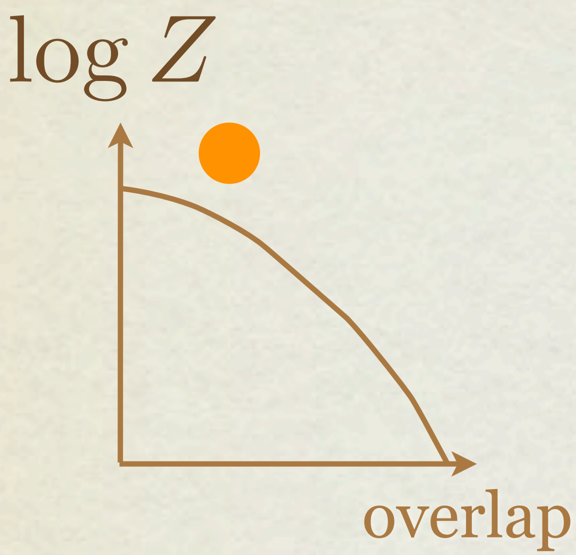
$q \to \infty \qquad \approx q \qquad \approx q \log q \qquad \approx 2q \log q \qquad \approx q^2$

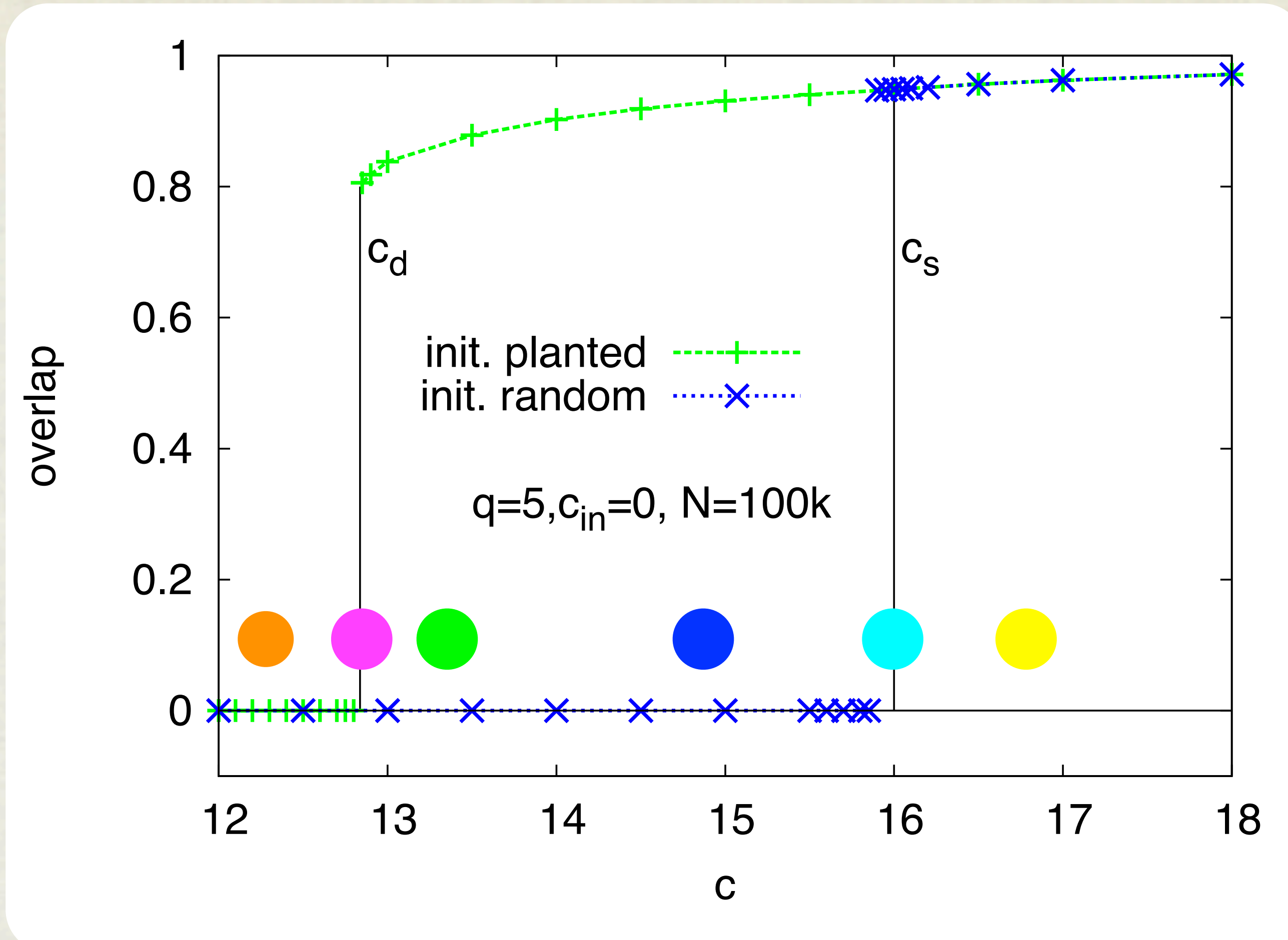
- For $c<c_K$ random and planted contiguous, statistical inference impossible, planted configuration is an equilibrium one.

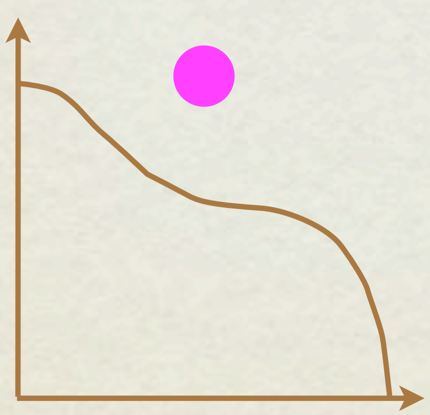
- "spinodal" phase transition at c_s
  - evaluation of marginals (inference, sampling) hard $c_K < c < c_s$
  - evaluation of marginals tractable $c > c_s$

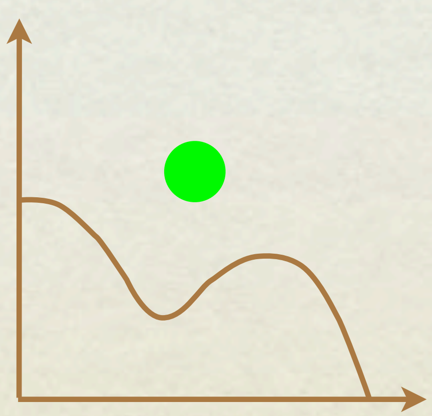
1ST ORDER TRANSITION

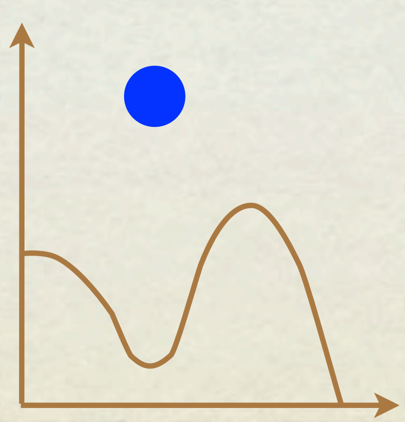
planted 5-coloring

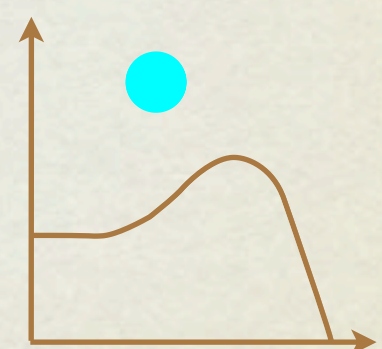


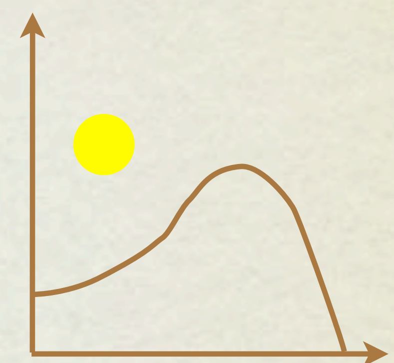






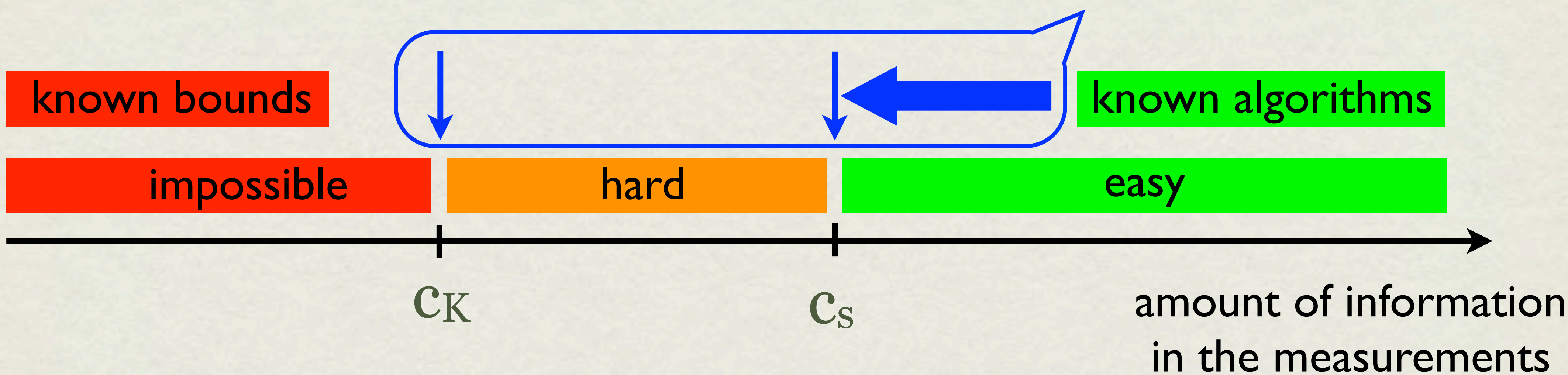






1ST ORDER TRANSITIONS

- Phase of possible but hard inference. Equilibrium state hidden by metastability.



- The hard phase quantified also in: planted constraint satisfaction, compressed sensing, stochastic block model, dictionary learning, blind source separation, sparse PCA, error correcting codes, others ....

THE BIG QUESTION

- Establish rigorous notions of algorithmic complexity (some kind of dichotomies) that are sensitive to the dynamical (c_d) and the spinodal (c_s) phase transition.