The computer science and physics of community detection

Kaito Takanami

December 26, 2023

School of Science, The University of Tokyo

Table of Contents

Introduction

Introduction: Why physics?

Approach to the community detection problem

- Min Bisection
  - Optimizing the objective function (e.g. modularity) for a given network
  - Maximizing is NP-hard but it performs well in real-world networks
  - However, model sometimes overfits to the data

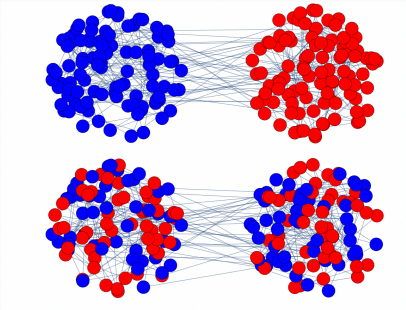


Figure 1: Partition of a random graph

- (Top) 38 edges crossing, (Bottom) 39 edges crossing
- The score is almost the same, but the community is completely different. $\Rightarrow$ Finding ground state is not enough!!

Introduction: Why physics?

Physics is useful for understanding high-dimensional statistics, not only the ground state.
(e.g. The typical number of molecules is $\sim 10^{23}$!!)

Also,

- In computer science, we think worst-case instances for evaluating algorithms.
- However, the real world networks are rarely worst-case instances.
- In physics, we think typical instances for evaluating models.
  (e.g. thermodynamics)
  $\Rightarrow$ It is natural to use physical perspective to evaluate the community detection models.

Problem setting

Problem setting

- Sparse and symmetric stochastic block model (SBM)
  - n: # nodes
  - q: # communities
  - each node i belongs to a community $\sigma_i \in \{1, \ldots, q\}$
  - if i and j belong to the same community, $A_{ij} \sim \text{Bern}(p_{in})$
  - if i and j belong to different communities, $A_{ij} \sim \text{Bern}(p_{out})$
  - $p_{in} = c_{in}/n$ and $p_{out} = c_{out}/n$ where c_{in} and c_{out} are constants.
- In this model, the expected degree of each vertex is

$$c = \frac{n}{q} \times p_{in} + \left(1 - \frac{n}{q}\right) \times p_{out} = \frac{c_{in} + (q-1)c_{out}}{q} \tag{1}$$

- Note that Eronds-Renyi model is a special case of SBM where $G(n, p = c/n)$.

Goal

There are several types of goal in the community detection problem when G generated from SBM is given.

- Exact reconstruction: Finding the planted assignment exactly up to a global permutation.
- Reconstruction (weak recovery): Finding the planted assignment whose accuracy is better than random guess ($1/q$).
- Detection: Hypothesis testing whether G is generated from SBM or Eronds-Renyi model.

In this talk, we assume c_{in} and c_{out} are known and focus on the reconstruction and detection problems, which are strongly analogous to physics.

Introduction to statistical physics

Preparation: Statistical physics and Ising model

- Statistical physics is a study of macroscopic properties of a system from microscopic properties.
- Ising model: a simple but fundamental model of magnetism and phase transition.
  - There are n atoms on a lattice $G = (V, E)$ and each atom has a spin $\sigma_i \in \{-1, 1\}$.
  - Neighboring atoms tend to align their spins, so the energy of the system is

$$H(\sigma) = -J \sum_{(i,j)\in E} \delta_{\sigma_i,\sigma_j}.$$

  - The probability of the system in equilibrium is given by the Boltzmann distribution

$$P(\sigma) \propto \exp(-H(\sigma)/T) = \exp\left(\frac{J}{T} \sum_{(i,j)\in E} \delta_{\sigma_i,\sigma_j}\right),$$

    where T is the temperature.

Preparation: Statistical physics and Ising model

There is a phase transition in the Ising model, which is a sudden change of the macroscopic properties of the system.

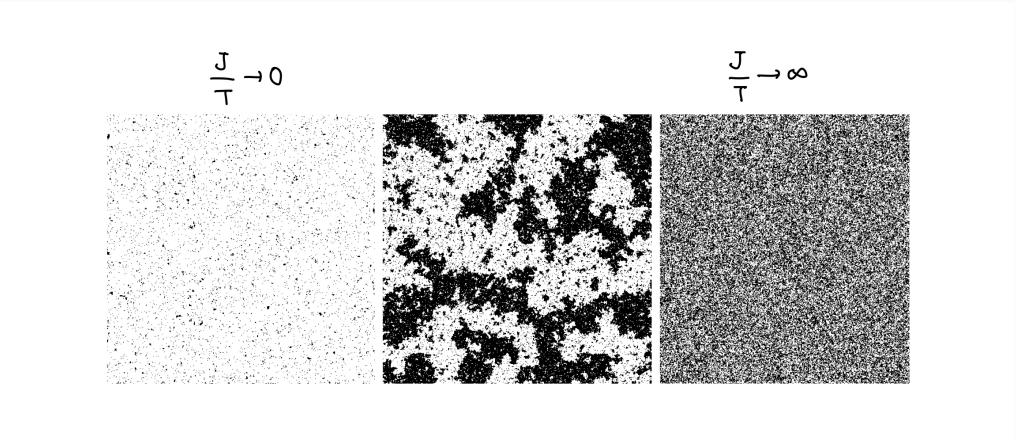


Figure 2: A typical states of the Ising model

- When J is large, the system is in the ferromagnetic phase where the spins are aligned.
- When J is small, the system is in the paramagnetic phase where the spins are randomly distributed.
- The change between these phases is sudden: phase transition.

Community detection as a statistical physics problem

In our model, the probability distribution of σ is given by

$$
\begin{aligned}
& P(\sigma \mid G) \\
\propto\ & P(G \mid \sigma) \\
=\ & \prod_{(i,j)\in E} p_{\sigma_i,\sigma_j} \prod_{(i,j)\notin E} (1 - p_{\sigma_i,\sigma_j}) \\
=\ & \prod_{(i,j)\in E} \left(\frac{p_{in}}{p_{out}}\right)^{\delta_{\sigma_i,\sigma_j}} \prod_{(i,j)\notin E} \left(1 - \frac{p_{in}}{p_{out}}\right)^{1-\delta_{\sigma_i,\sigma_j}} \\
=\ & \exp\left(\log\left(\frac{p_{in}}{p_{out}}\right) \sum_{(i,j)\in E} \delta_{\sigma_i,\sigma_j} + \log\left(1 - \frac{p_{in}}{p_{out}}\right) \sum_{(i,j)\notin E} (1 - \delta_{\sigma_i,\sigma_j})\right) \\
\sim\ & \exp\left(\log\left(\frac{p_{in}}{p_{out}}\right) \sum_{(i,j)\in E} \delta_{\sigma_i,\sigma_j}\right)
\end{aligned}
$$

Community detection as a statistical physics problem

- In conclusion, the community detection problem is equivalent to the Ising model with $J/T = \log\left(\frac{p_{in}}{p_{out}}\right)$ in a simple approximation.
- From an analogy of the phase transition in the Ising model, we can expect the success-to-failure phase transition in the community detection.
- Considering only the MLE is not enough to understand the properties of magnetism, which is the same for the community detection.

Technical tools: Belief propagation

- Belief propagation (BP): method in statistical physics for computing marginal distributions $p(\sigma_i \mid G)$ of a graphical model.
- Sketch of BP procedure
  - message: $m_{i\to j}(\sigma_j)$ is the probability of σ_i if j does not exist.
  - If we know the messages $m_{1\to 4}$ and $m_{2\to 4}$, we can compute the message $m_{4\to 3}$ by

$$m_{4\to 3}(\sigma_3) \propto \left(\sum_{\sigma_1} m_{1\to 3}(\sigma_1) f(\sigma_1, \sigma_3)\right)\left(\sum_{\sigma_2} m_{2\to 3}(\sigma_2) f(\sigma_2, \sigma_3)\right)$$

    where $f(\sigma_1, \sigma_3)$ is the relative probability of σ_1 and σ_3.
  - If we can compute all the messages, we can compute the marginal distribution $p(\sigma_i \mid G)$ by

$$p(\sigma_3 \mid G) \propto \prod_{j\in\{1,2,4\}} \sum_{\sigma_j} m_{j\to 3}(\sigma_3) f(\sigma_3, \sigma_j)$$

Easy regime

Technical tools: Belief propagation

- If the model is tree, BP is exact.
- In general, we need to propagate the messages for several times until convergence.

Question

When does BP converge to the "correct answer"?

- $m_{i\to j}(\sigma_j = r) = 1/q$ is a trivial fixed point, where $p(\sigma_i \mid G)$ $= 1/\mathrm{q}$.
- If BP is stacked at this fixed point, BP is no better than random guess.
- One can calculate the stability of fixed points by linearizing the BP equation.

Sketch of stability analysis

- Suppose the messages are almost uniform:

$$m_{i\to j}(\sigma_j) = \frac{1}{q} + \epsilon_{i\to j}(\sigma_j)$$

- Substituting this into the BP equation, we obtain in the first order of $\epsilon_{i\to j}$,

$$\epsilon^{(l+1)} \propto M\epsilon^{(l)}$$

If M has an eigenvalue whose absolute value is larger than 1, $\epsilon^{(l)}$ diverges.
$\Rightarrow$ trivial fixed point is unstable ☺.

Sketch of stability analysis

- From some technical calculation, we obtain

$$M = B \bigotimes T$$

where

$$T = \frac{1}{qc} \begin{pmatrix} c_{in} & \cdots & c_{out} \\ \vdots & \ddots & \vdots \\ c_{out} & \cdots & c_{in} \end{pmatrix}$$

and

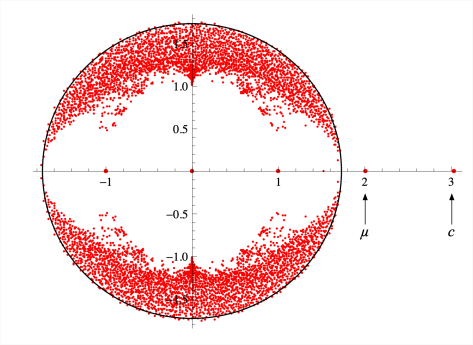
$$B_{(i,j),(k,\ell)} = \begin{cases} 1 & \text{if } \ell = i \text{ and } k \neq j \\ 0 & \text{otherwise.} \end{cases}$$

Sketch of stability analysis

- The relevant first eigenvalue of T is given by

$$\lambda = \frac{c_{in} - c_{out}}{qc}$$

and the relevant first eigenvalue of B is given by $\max(c\lambda, \sqrt{c})$.



which leads to the following theorem.

Kesten-Stigum threshold

If $c\lambda^2 > 1$, the trivial fixed point is unstable.

It is proofed that in this regime, an efficient algorithm for weak reconstruction exist ☺.

Impossible regime

Contiguity

So far, we found "easy regime" of weak reconstruction.

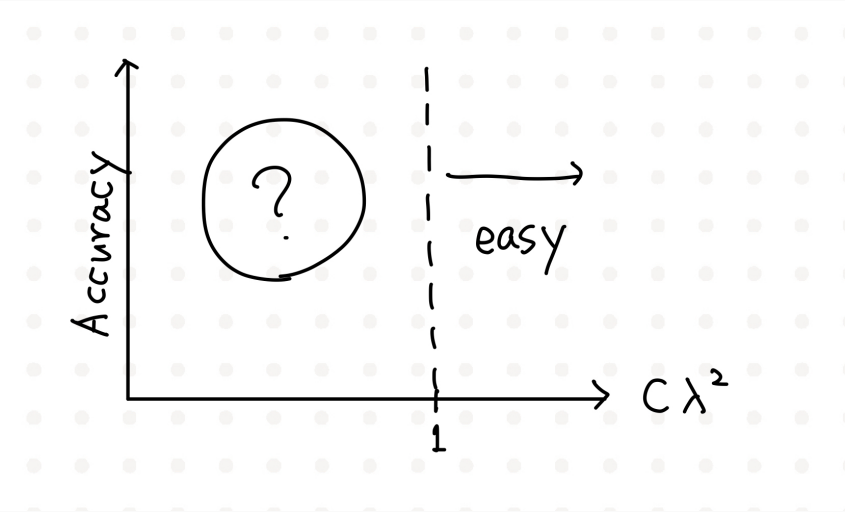


Figure 3: Phase diagram of the community detection

Question

Is there any regime where the weak reconstruction is impossible by any algorithm?

Contiguity

P: SBM, Q: Eronds-Renyi model

- $P \trianglelefteq Q$: $\forall E$ such that $\lim_{n\to 0} Q(E) = 0$, $\lim_{n\to 0} P(E) = 0$.
- If $P \trianglelefteq Q$ and $Q \trianglelefteq P$, we say P and Q are contiguous.
- Interestingly, if P and Q are contiguous, the weak reconstruction is impossible by any algorithm.

Intuitive Proof:

Suppose there is an algorithm for weak reconstruction, which means that we can say "there is a community" to G in SBM with high probability.

Then by the same algorithm, we can say "there is community" to G in Eronds-Renyi model with high probability, which is incorrect.

Contiguity

The derivation of the condition of contiguity is complex ...

The result is presented as follows,

- $q = 2$: the same as the Kesten-Stigum threshold
- $q \geq 3$: $c\lambda^2 < \frac{2\ln(q-1)}{q-1}$

a lower bound on information-theoritic threhold

When $q \geq 3$, the weak reconstruction is impossible if $c\lambda^2 < \frac{2\ln(q-1)}{q-1}$.

Hard regime

Hard regime

So far, we found "easy regime" and "impossible regime" of weak reconstruction.

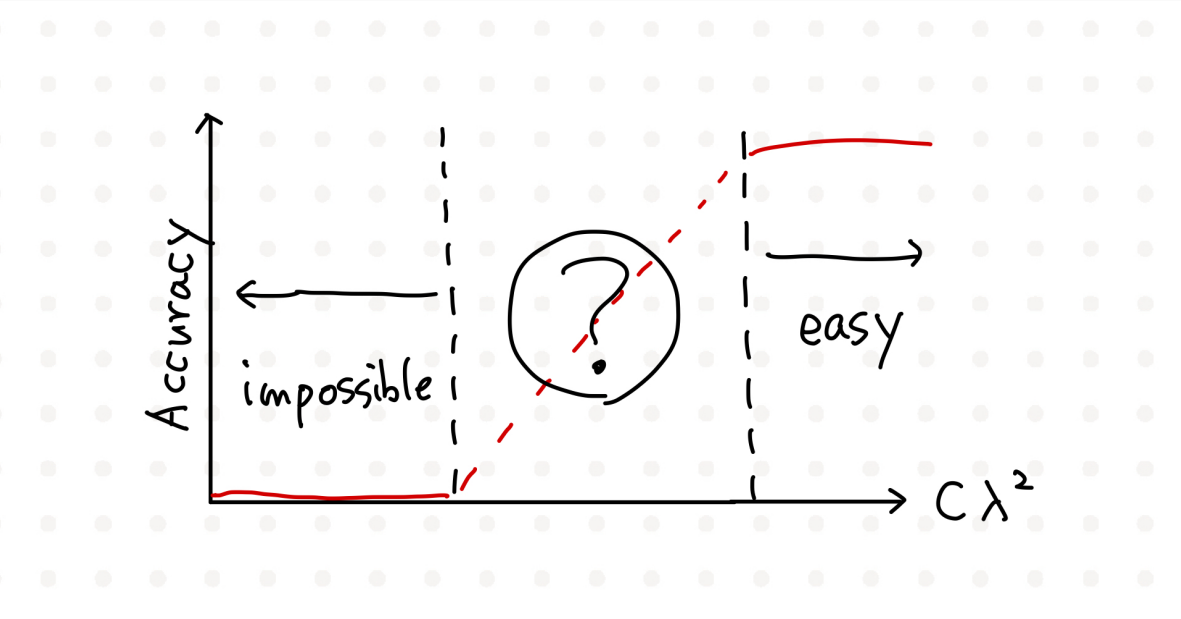


Figure 4: Phase diagram of the community detection

Question

What happens in the middle regime?

Hard regime

Suppose we can explore the whole space of spin configurations.

- In physics, the realized state is the one in which the free energy is minimized.
- Similarly, in Bayesian statistics, the typical state is the one in which the free energy

$$F(\sigma) = -\frac{1}{n} \log P(G)$$

(negative log-likelihood) is minimized.
- It is reasonable to adopt the estimation result from the two,
  (1) a correct fixed point
  (2) a uniform fixed point

  which yields the lower free energy.

Complete phase diagram

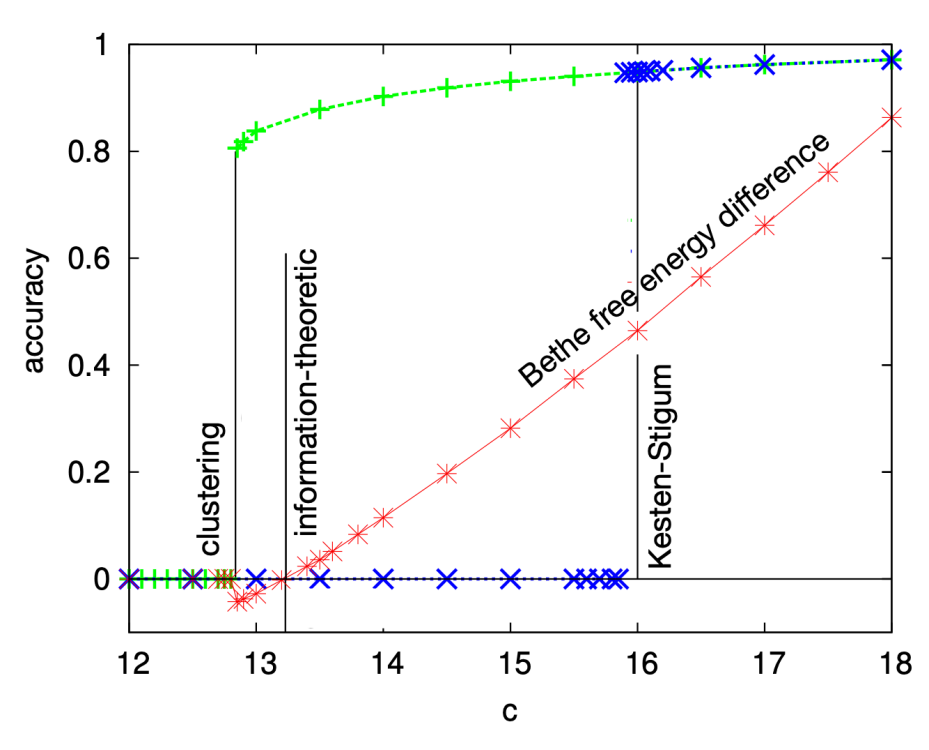


Figure 5: Full phase diagram of the community detection ($q = 5$)

Two hard regimes:

(1) accurate point $<$ uniform point

(2) accurate point $>$ uniform point

Two hard regimes

(1) accurate point $<$ uniform point
- Accurate fixed point exists, but the uniform one is better.
- This corresponds to the overfitting: many local minimum exist.

(2) accurate point $>$ uniform point
- Uniform fixed point is better than the accurate one.
- However, exponentially small fraction of the initial messages are attracted to the uniform fixed point. $\Rightarrow$ One must reselect initial conditions an exponential number of times.

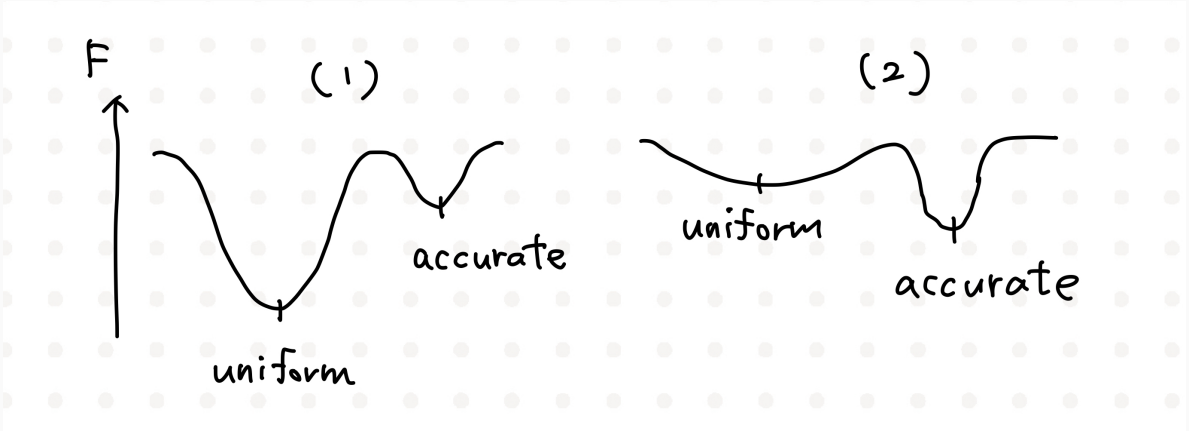


Figure 6: Landscapes of the free energy in two hard regimes

Summary

Summary (1)

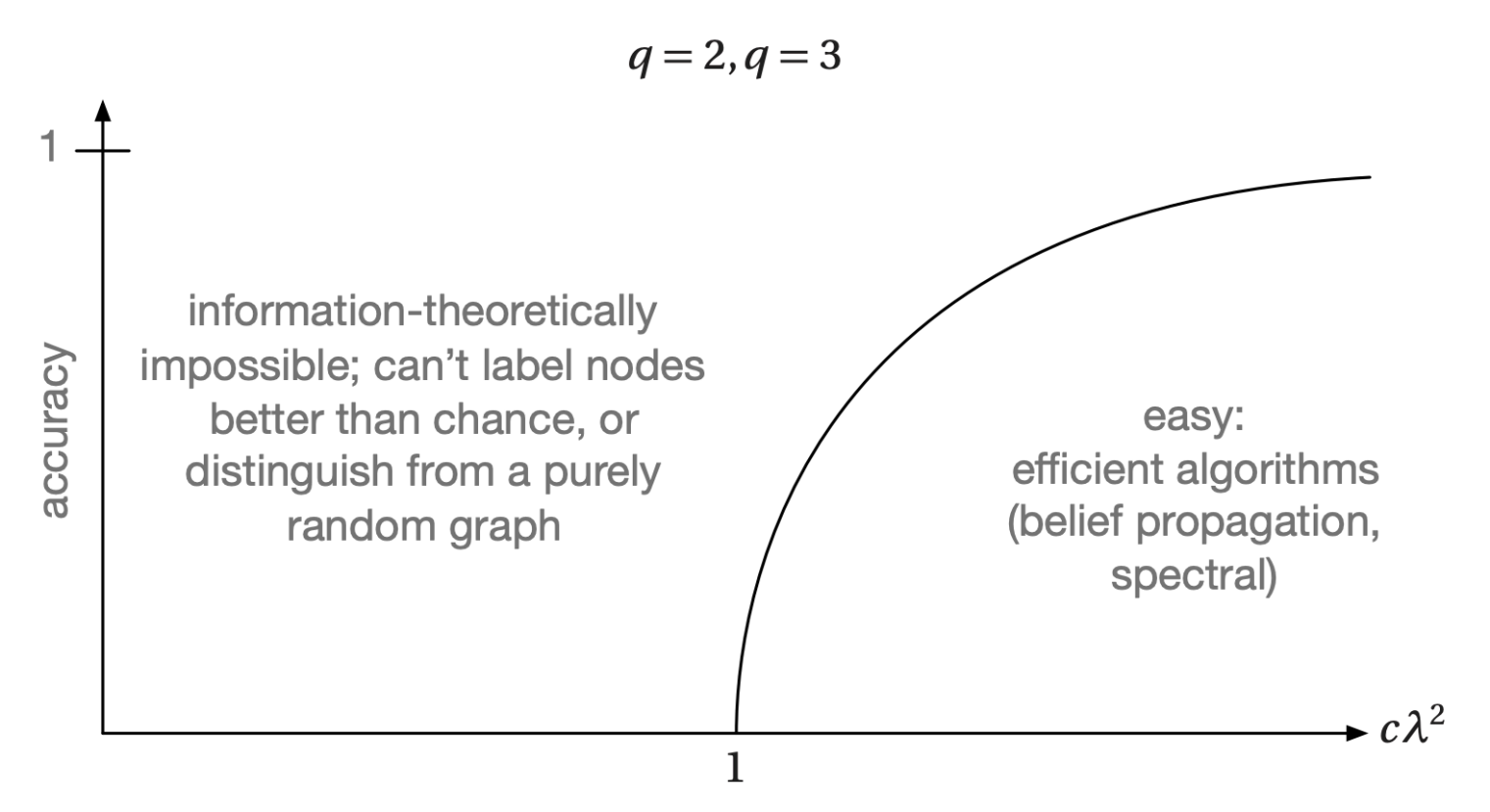


Figure 7: Full phase diagram of the community detection ($q = 2, 3$)

Summary (2)

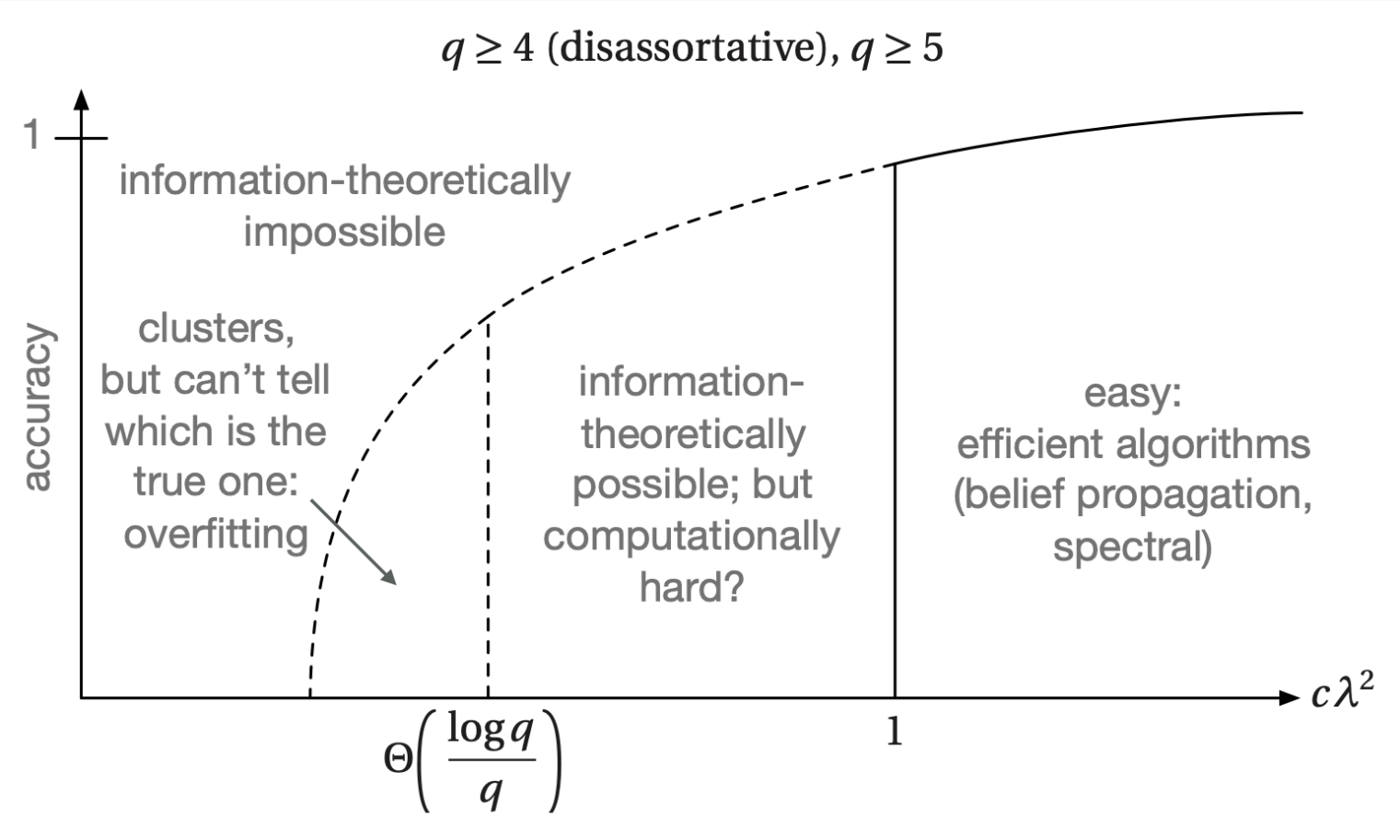


Figure 8: Full phase diagram of the community detection ($q > 3$)

Open problems

Open problems as of 2017

- The rigorous information-theoretical and computational threshold in assortative SBM (i.e. $c_{in} > c_{out}$).
- The rigorous proof that reconstruction is hard for specific algorithms in the hard regimes.
  (e.g. Do MCMC or BP with random initial conditions really take exponential time?)
- In hard regimes, Is it possible to solve efficiently by providing some correct labels (side information)?