Community Detection via Semidefinite Programming

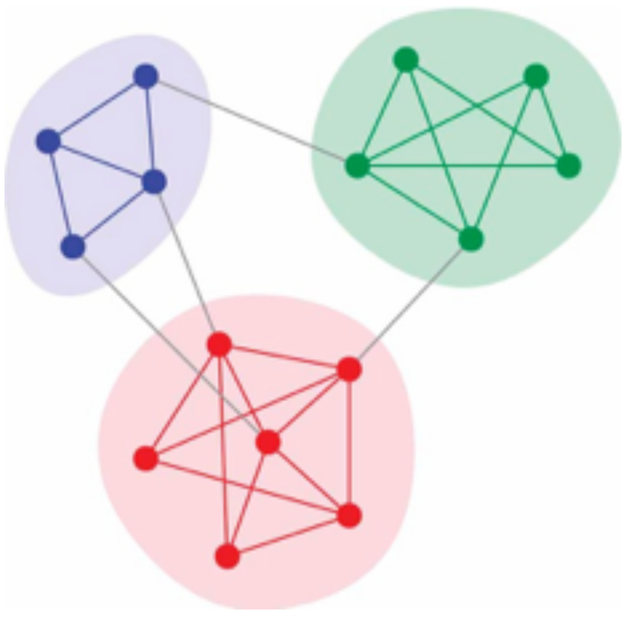
Federico Ricci-Tersenghi
(Sapienza University)

in collaboration with
Adel Javanmard and Andrea Montanari

PNAS 113, E2218 (2016)
J. Phys.: Conf. Ser. 699, 012015 (2016)

Communities detection problem

- Detecting communities/partitions/clusters in graphs is a widespread problem in many different disciplines



- We need fast (linear and scalable) algorithms
  - robust (real datasets are very noisy and not random)
  - close to optimal (on random ensemble benchmarks)

Benchmark for community detection

Hidden partition model or stochastic block model (SBM)

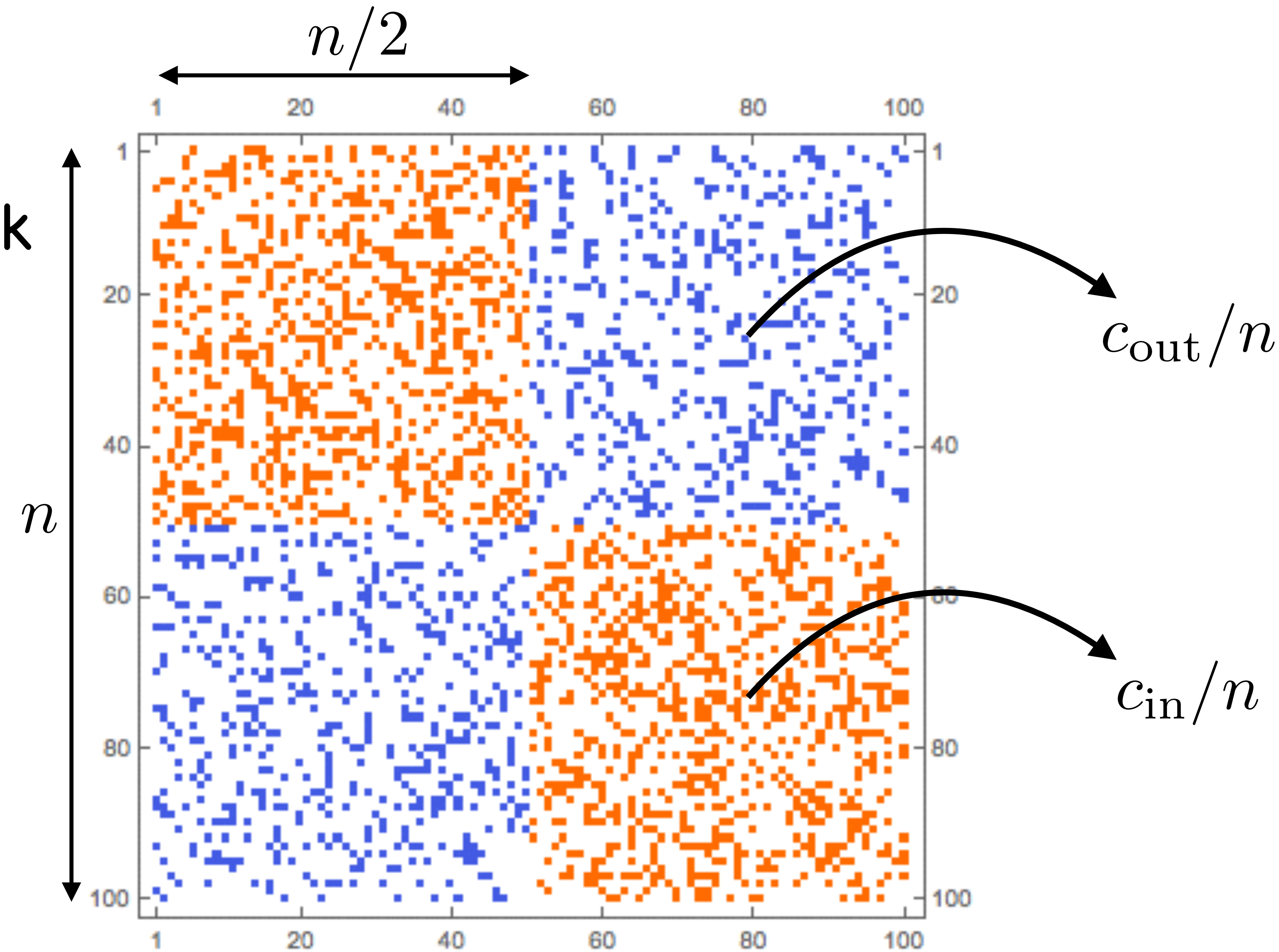
- Generate a partition of n nodes: e.g. q groups of size n/q
- Add independently edges between any pair of nodes according to the following probability

$$\mathbb{P}[(ij) \in E] = \begin{cases} c_{\text{in}}/n & \text{same group} \\ c_{\text{out}}/n & \text{different groups} \end{cases}$$

- Assortative model $c_{\text{in}} > c_{\text{out}}$
  Disassortative model $c_{\text{in}} < c_{\text{out}}$

The hidden partition model

Stochastic block model (SBM) with $q = 2$

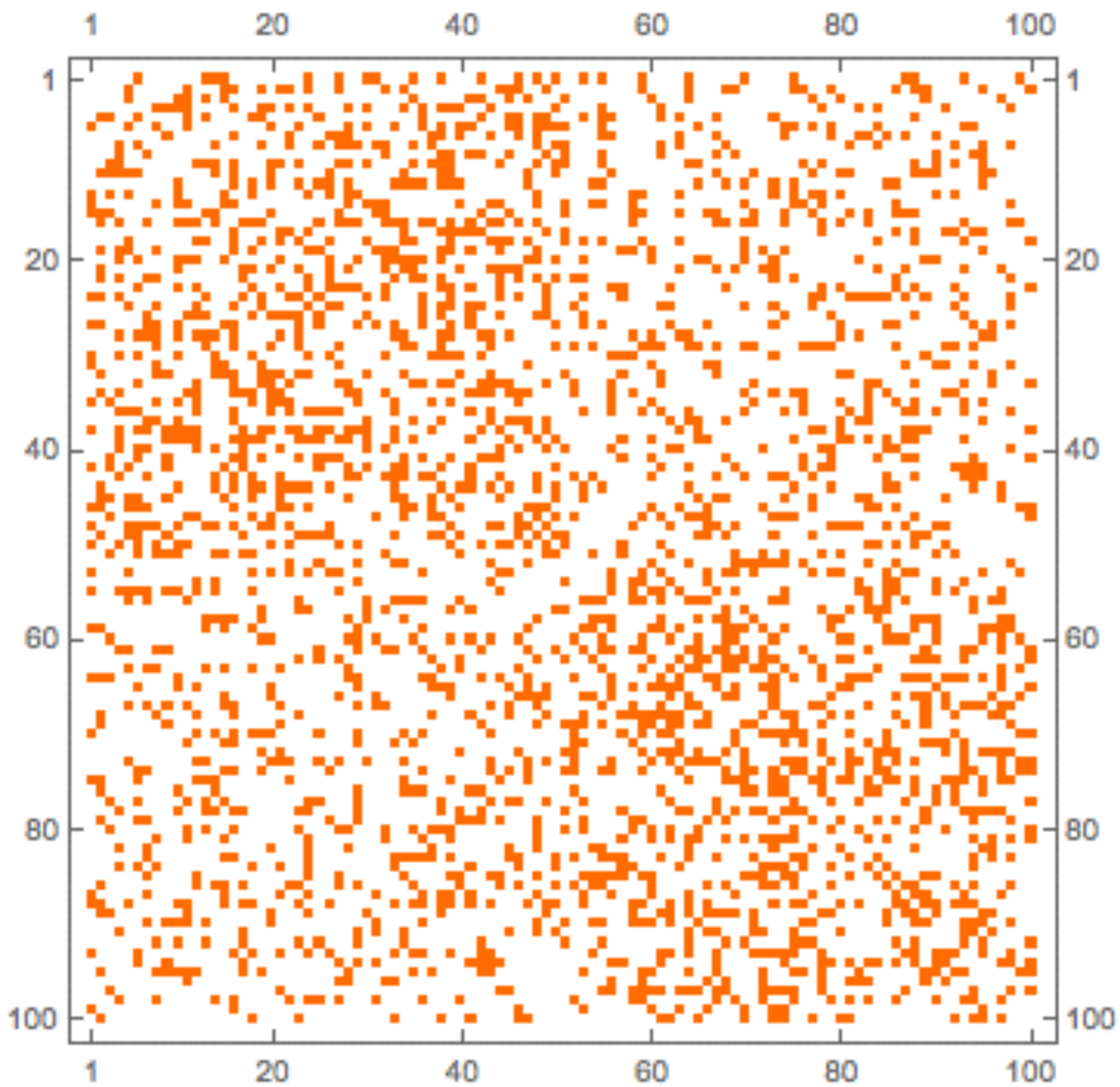


$$\mathbb{P}[(ij) \in E] = \begin{cases} c_{\mathrm{in}}/n & \text{same group} \\ c_{\mathrm{out}}/n & \text{different groups} \end{cases}$$

Assortative model:

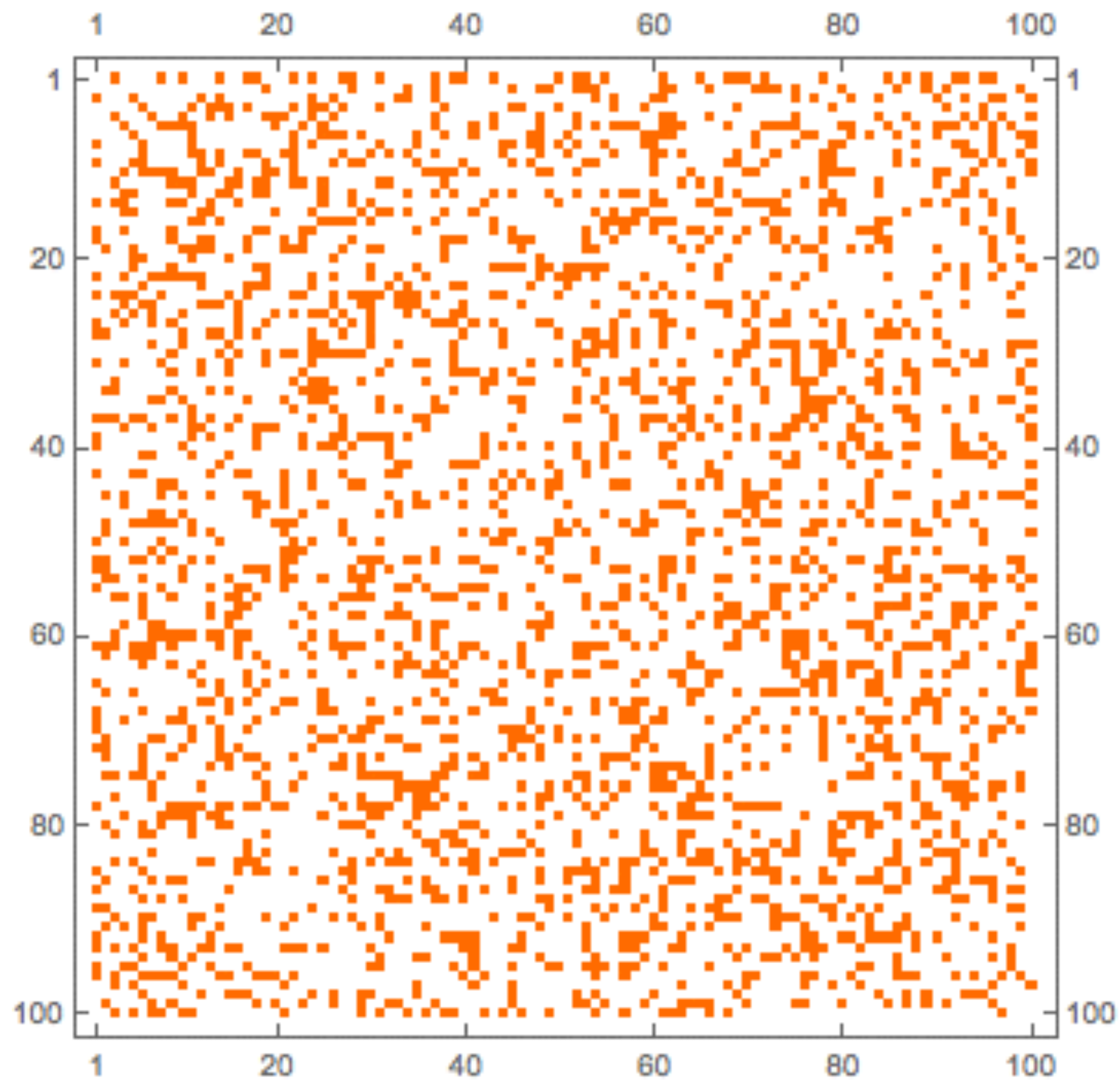
$$c_{\mathrm{in}} > c_{\mathrm{out}}$$

The hidden partition model



Colors are not provided !

The hidden partition model

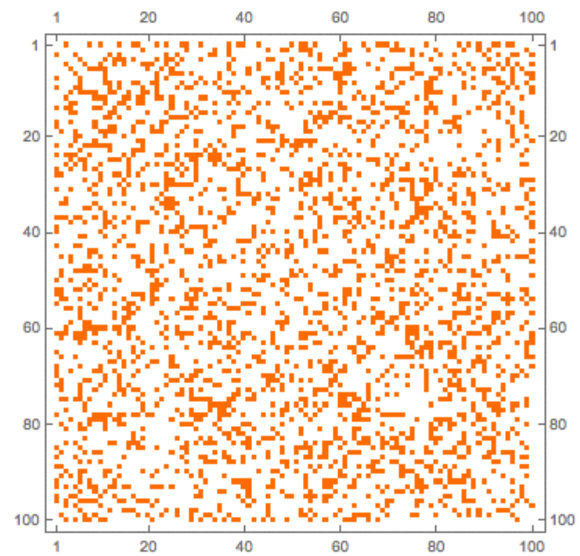


The right ordering neither !!

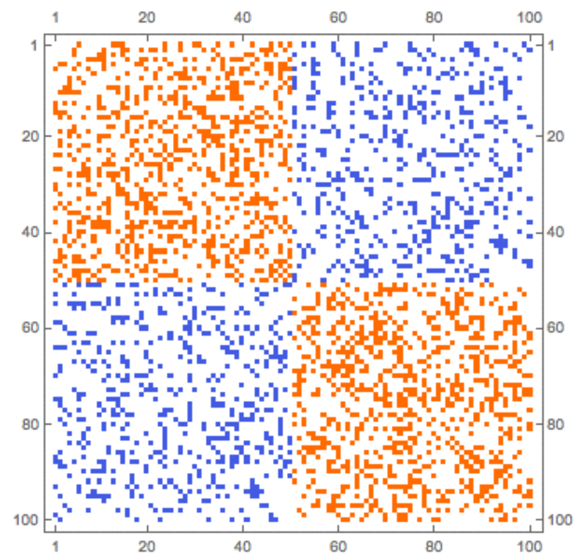
The hidden partition model

Given only the adjacency matrix

$$A_{ij} = A_{ji} = \mathbb{I}[(ij) \in E]$$



Infer the hidden partition



Hidden (true) partition -> $\boldsymbol{x}_0 \in \{+1, -1\}^n$

Estimated partition -> $\hat{\boldsymbol{x}}(G) \in \{+1, -1\}^n$

Quality of inference via the overlap -> $Q = \frac{1}{n} |\langle \hat{\boldsymbol{x}}(G), \boldsymbol{x}_0 \rangle|$

Assortative SBM with 2 equal-size groups

Relevant parameters and threshold

- Mean degree $d = \frac{c_{\text{in}} + c_{\text{out}}}{2}$
- Signal-to-noise ratio $\lambda = \frac{c_{\text{in}} - c_{\text{out}}}{2\sqrt{d}}$
- Bayes optimal threshold $\lambda_c = 1$
  - Impossible detection for $\lambda < \lambda_c$
  - BP algorithm with $Q > 0$ for $\lambda > \lambda_c$

[Decelle, Krzakala, Moore, Zdeborova, 2011]
[Massoulie, 2013] [Mossel, Neeman, Sly, 2013]

Maximum Likelihood (ML)

- If no information on the generative model is given (apart being assortative and with 2 equal-size groups) a good choice is to maximize the likelihood

$$\text{maximize} \quad \sum_{i,j} A_{i,j} x_i x_j$$

$$\text{subject to} \quad x_i \in \{+1, -1\} \quad \text{and} \quad \sum_i x_i = 0$$

- NP-hard problem

Spectral relaxation

- Relaxes the constraint $x \in \{+1, -1\}^n$
- Compute largest/smallest eigenvalues of a combination of adjacency (A) and degrees (D) matrices
  Project the corresponding eigenvector to $\{+1, -1\}^n$
  - Laplacian $L = D - A$
  - Normalized Laplacian $D^{-1/2} L D^{-1/2}$

  Eigenvector localization on high or low degree nodes

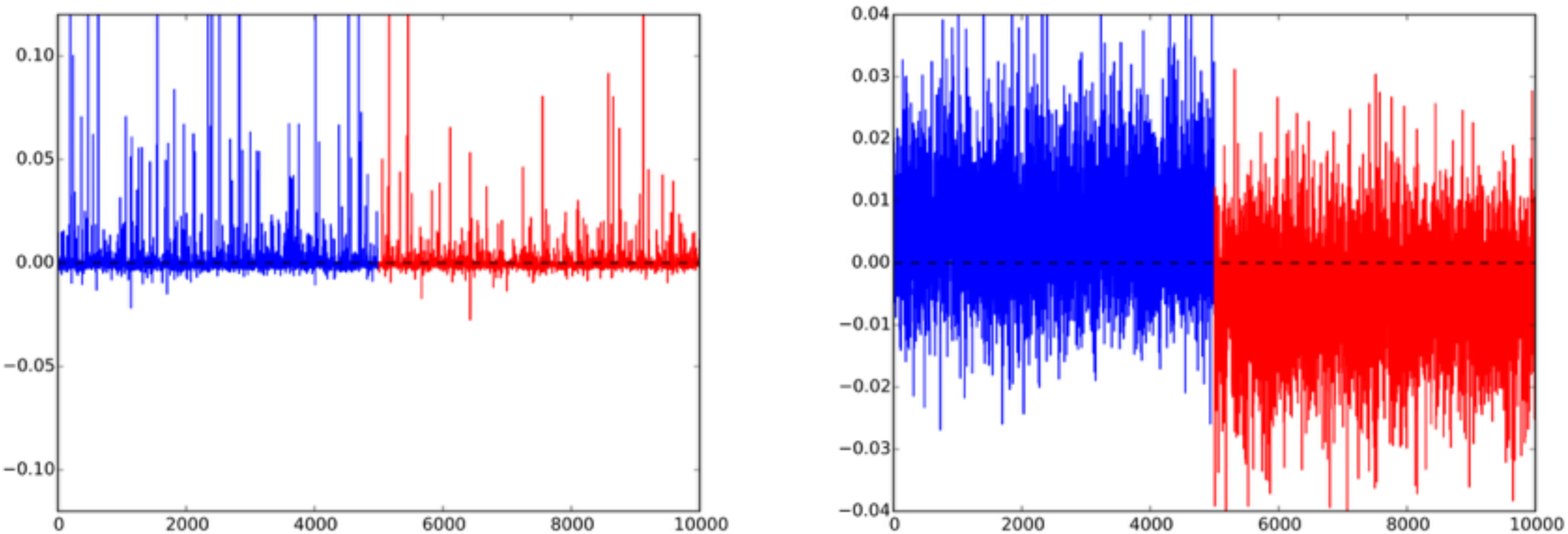
  - Bethe Hessian $H(\lambda) = (\lambda^2 - 1)\mathbb{I} + D - \lambda A$
    [Saade, Krzakala, Zdeborova, 2014]
  - z-Laplacian $L_z = zA - D$
    [Banks, Moore, Newman, Zhang, 2014]

  Eigenvector localization on cliques

Spectral relaxation fails on sparse graphs

$$n = 10^4 \quad \lambda = 1.2$$

$$d = 3 \qquad\qquad d = 20$$



$$\boldsymbol{v}_1(\boldsymbol{A}^{\mathbf{cen}})$$

Quasi-random graphs (SBM + random cliques)

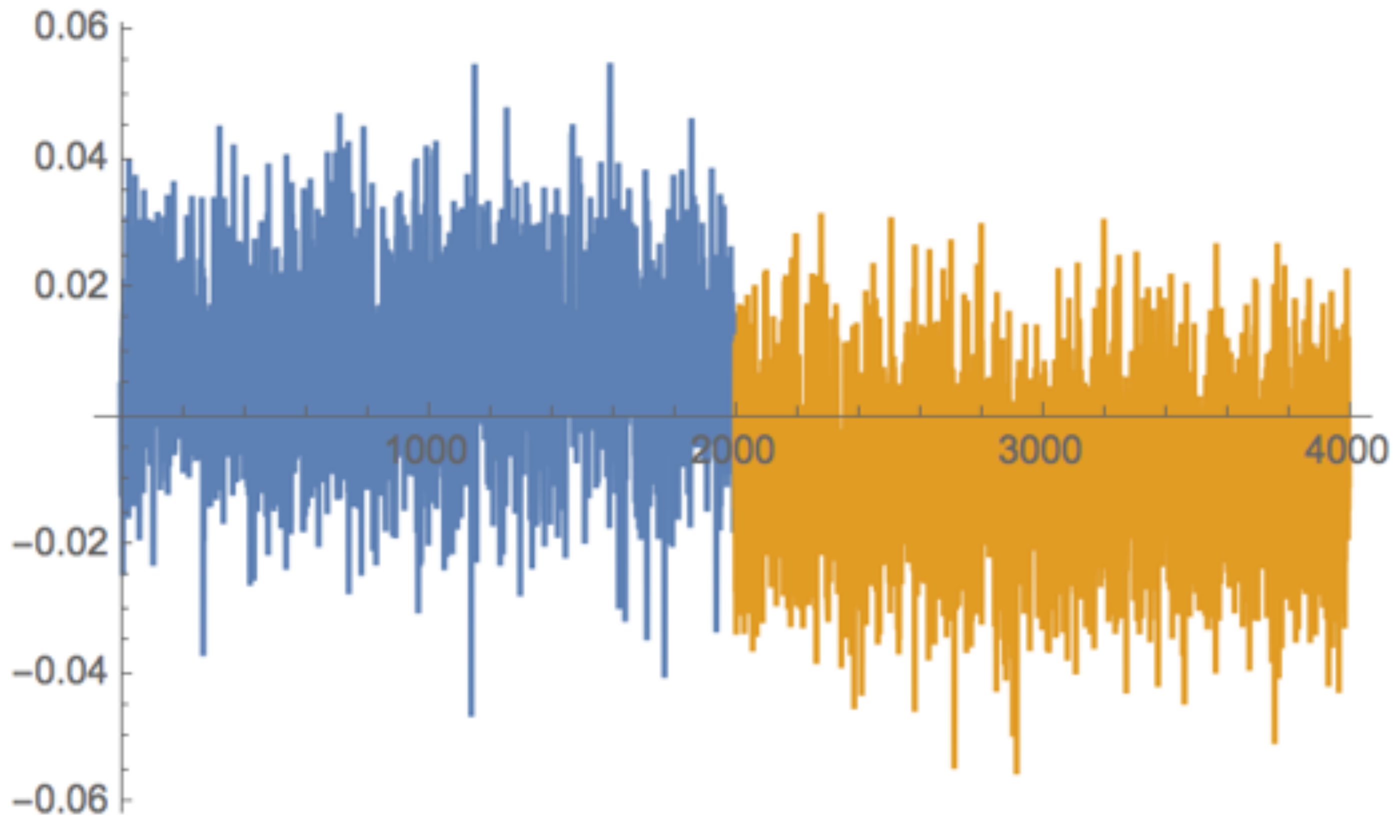
- Generate a graph according to the SBM
- Choose a subset S of vertices of size $|S| = \alpha n$
- For each vertex in S connect all its neighbours
- The number of edges increases by $\sim \alpha d^2 n/2$ i.e. by a fraction $\sim \alpha d$
- A robust inference method should work also for $\alpha > 0$ at least in the regime $\alpha \ll 1/d$

Improved spectral methods fail on quasi-random graphs

$$n = 4000 \quad d = 4 \quad \lambda = 1.1$$

$\alpha = 0$

$\boldsymbol{v}_2(H(\sqrt{d}))$

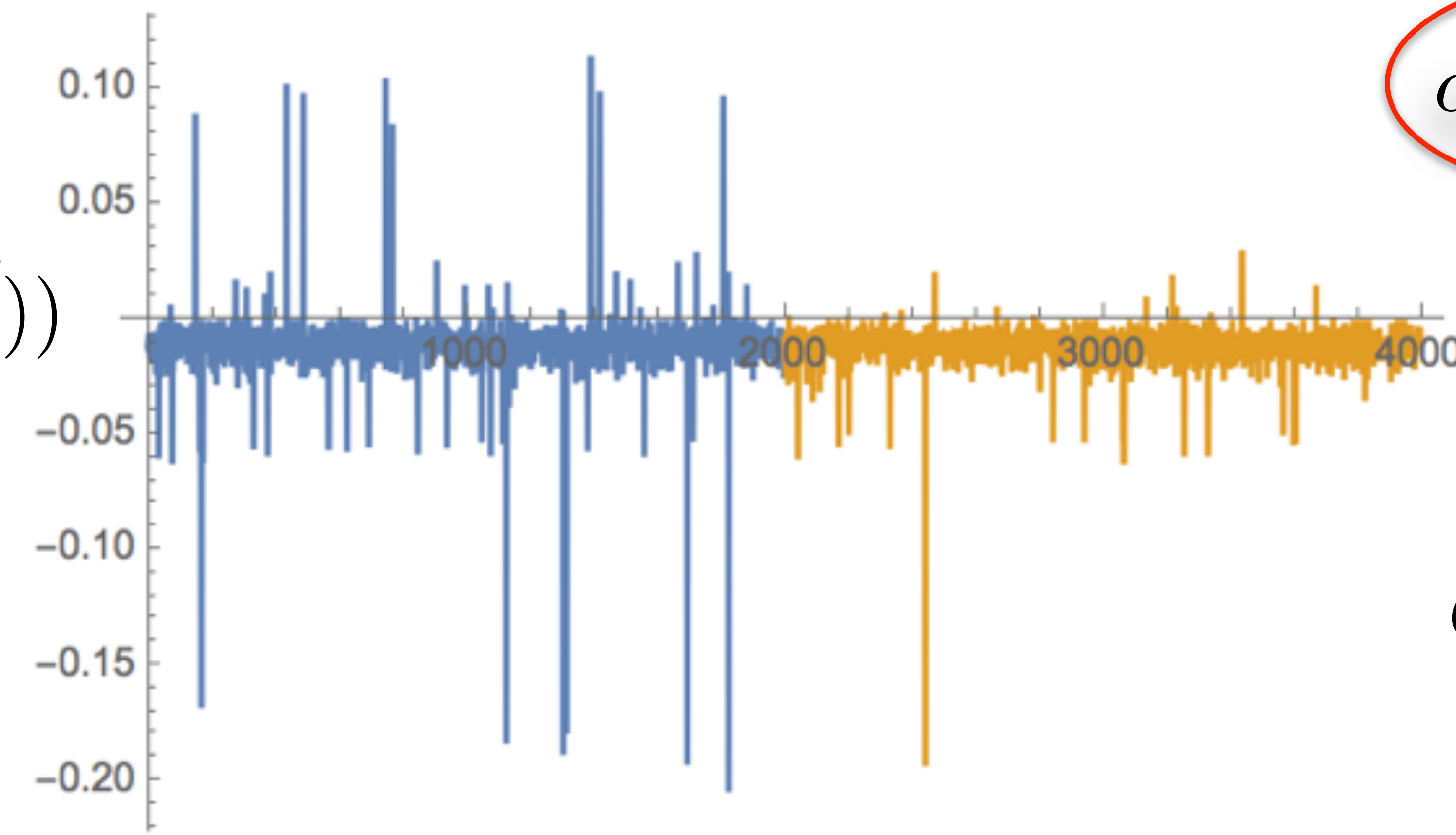


$Q = 0.439$

Improved spectral methods fail on quasi-random graphs

$$n = 4000 \quad d = 4 \quad \lambda = 1.1$$

$\boldsymbol{v}_2(H(\sqrt{d}))$

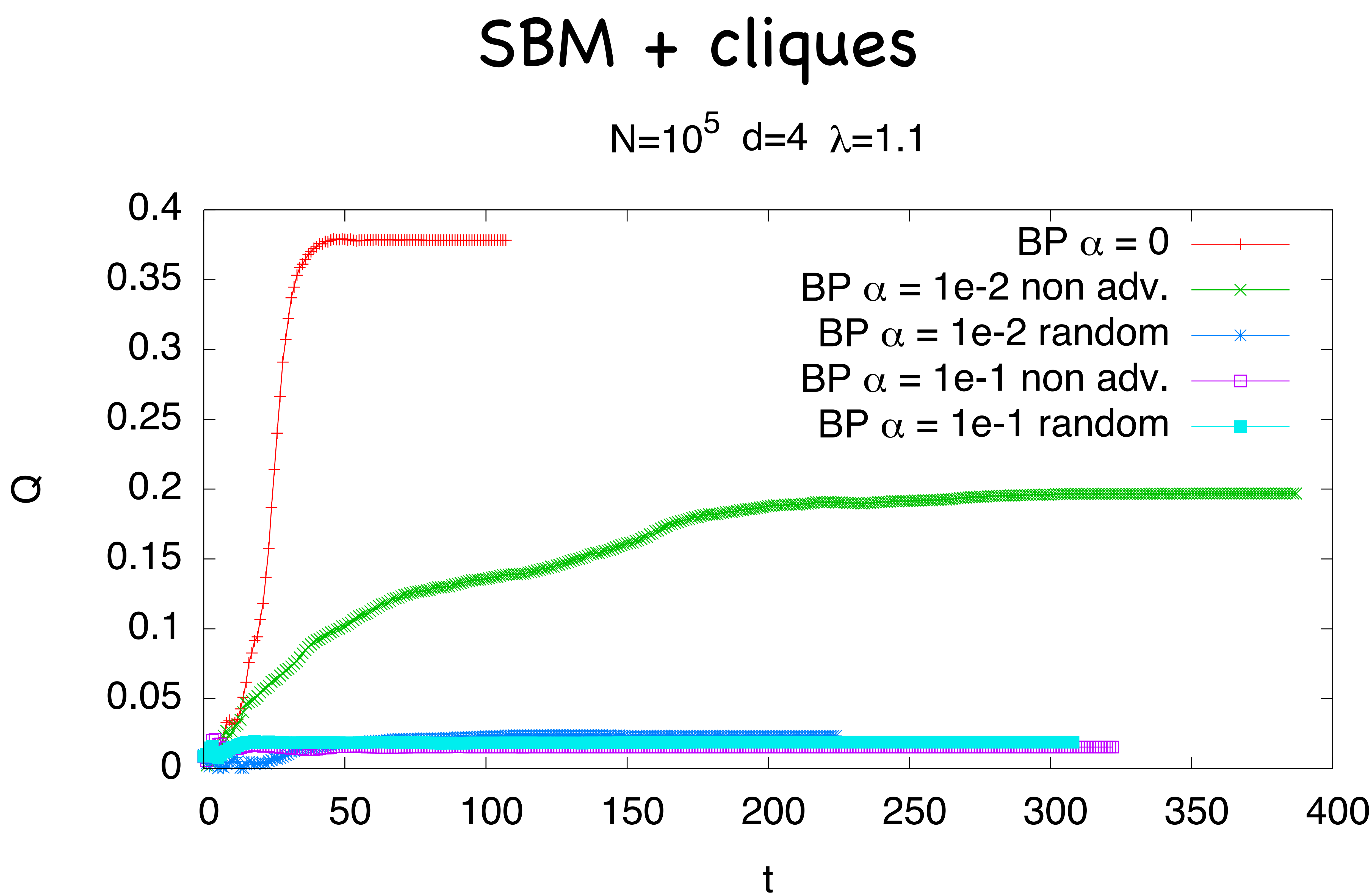


$\alpha = 0.001$

$Q \approx 0$

Quasi-random graphs (SBM + non adversarial cliques)

- Generate a graph according to the SBM
- Choose a subset S of vertices of size $|S| = \alpha n$
- For each vertex in S connect all its neighbours belonging to the same community
- Non adversarial cliques provide <u>more information</u>



SBM + cliques
N=10^5 d=4 λ=1.1
BP α = 0
BP α = 1e-2 non adv.
BP α = 1e-2 random
BP α = 1e-1 non adv.
BP α = 1e-1 random
Q
t
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
50
100
150
200
250
300
350
400

SDP: a better relaxation?

- Maximize $\sum_{i,j} A_{i,j} x_i x_j$ over $\boldsymbol{x} \in \{+1,-1\}^n$

  it is equivalent to maximize $\langle \boldsymbol{A}, \boldsymbol{X} \rangle \equiv \sum_{i,j} A_{ij} X_{ij}$

  subject to $\boldsymbol{X} \in \mathbb{R}^{n \times n}$, $\boldsymbol{X} \succeq 0$ (i.e. all eigenvalues >= 0)

  $\boxed{X_{ii} = 1}$ and $\boldsymbol{X}$ being of rank 1

- SDP relaxes the rank and maximizes $\langle \boldsymbol{A}, \boldsymbol{X} \rangle$ over the convex space of positive semidefinite matrices

- The maximizer is a matrix of rank $m \in [1, n]$ to be projected back on a rank 1 matrix...

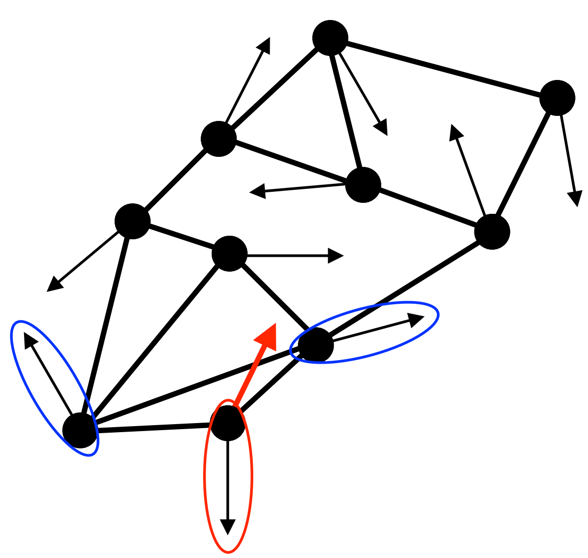
$$\boldsymbol{X}^{\mathrm{opt}} \longrightarrow \hat{\boldsymbol{x}}^{\mathrm{SDP}} (\hat{\boldsymbol{x}}^{\mathrm{SDP}})^{\mathsf{T}}$$

SDP-based algorithm

- Maximize $\langle \boldsymbol{A}, \boldsymbol{X} \rangle$ over rank-m matrices = correlation matrices between m-components variables of unit norm

$$C_{ij} = \underline{x}_i \cdot \underline{x}_j \,, \quad \text{with } \underline{x}_i \in \mathbb{R}^m \,, \; \|\underline{x}_i\|^2 = \underline{x}_i \cdot \underline{x}_i = 1$$

- Maximize $\sum_{(ij)\in E} \underline{x}_i \cdot \underline{x}_j$ subject to $\sum_i \underline{x}_i = \underline{0}$



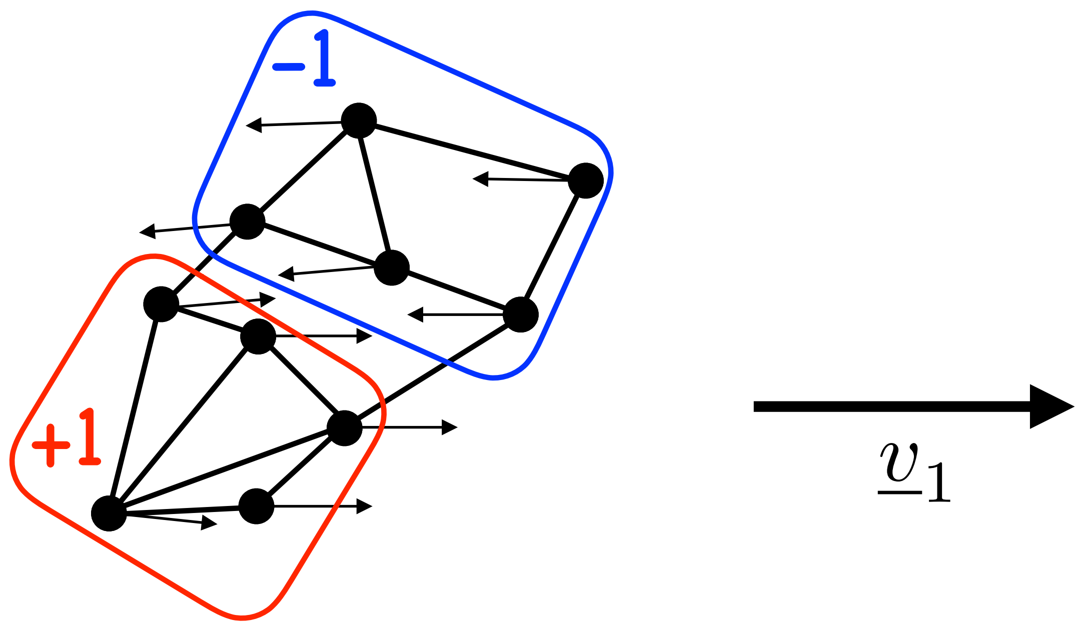
$$\underline{h}_i = \sum_{j\in\partial i} \underline{x}_j - \frac{d}{n} \sum_j \underline{x}_j$$

$$\underline{x}_i \leftarrow \frac{\underline{h}_i}{||\underline{h}_i||}$$

Greedy T=0 dynamics (very fast! no gradient used)

SDP-based algorithm

Once you reach a local maximum...



- Given the maximizer $\underline{\boldsymbol{x}}^* = \{\underline{x}_1^*, \ldots, \underline{x}_n^*\}$ compute the empirical covariance matrix (m x m)

$$\Sigma_{jk} = \frac{1}{n} \sum_{i=1}^{n} (\underline{x}_i^*)_j (\underline{x}_i^*)_k$$

- Project on its principal eigenvector $\hat{x}_i^{\mathrm{SDP}} = \mathrm{sign}(\underline{x}_i \cdot \underline{v}_1)$

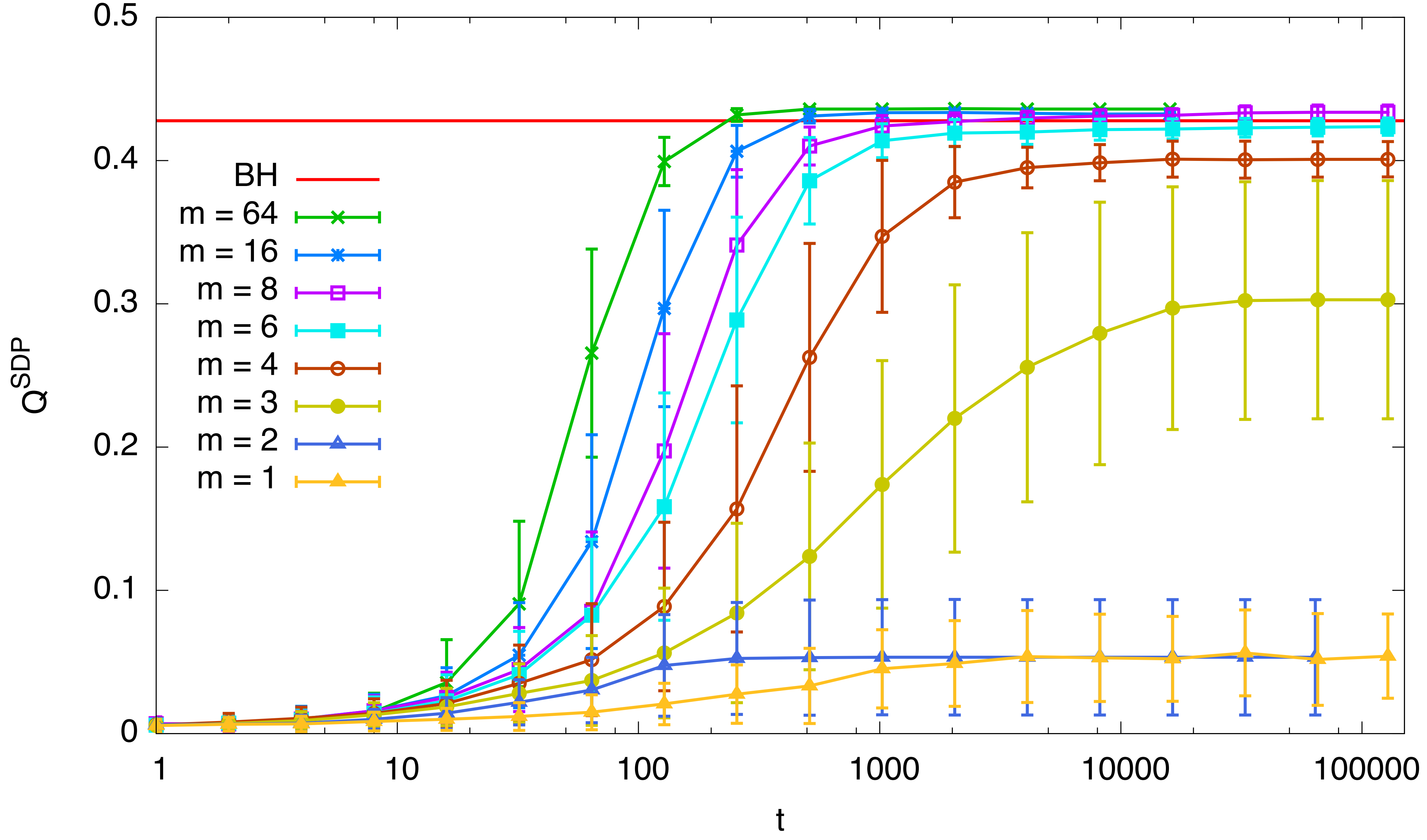
http://web.stanford.edu/~montanar/SDPgraph/

SDP-based algorithm

- Algorithm complexity $O(n\, m\, t_{\text{conv}})$ and quality of inference do depend on m
  - m=1 -> ML, very rough objective function, NP-hard
  - m=n -> SDP, convex objective function
    no local maxima for $m > \sqrt{2n}$ [Burer, Monteiro, 2003]
  - m>1, but small -> smooth enough objective function ?
    local minima are "close enough" $O(m^{-1/2})$
    to global minimum [Montanari, 2016]
- Running times grows very mildly with m and n
  e.g. if stopping rule is max variation < 10^{-3} -> $t_{\text{conv}} \propto n^{0.22}$

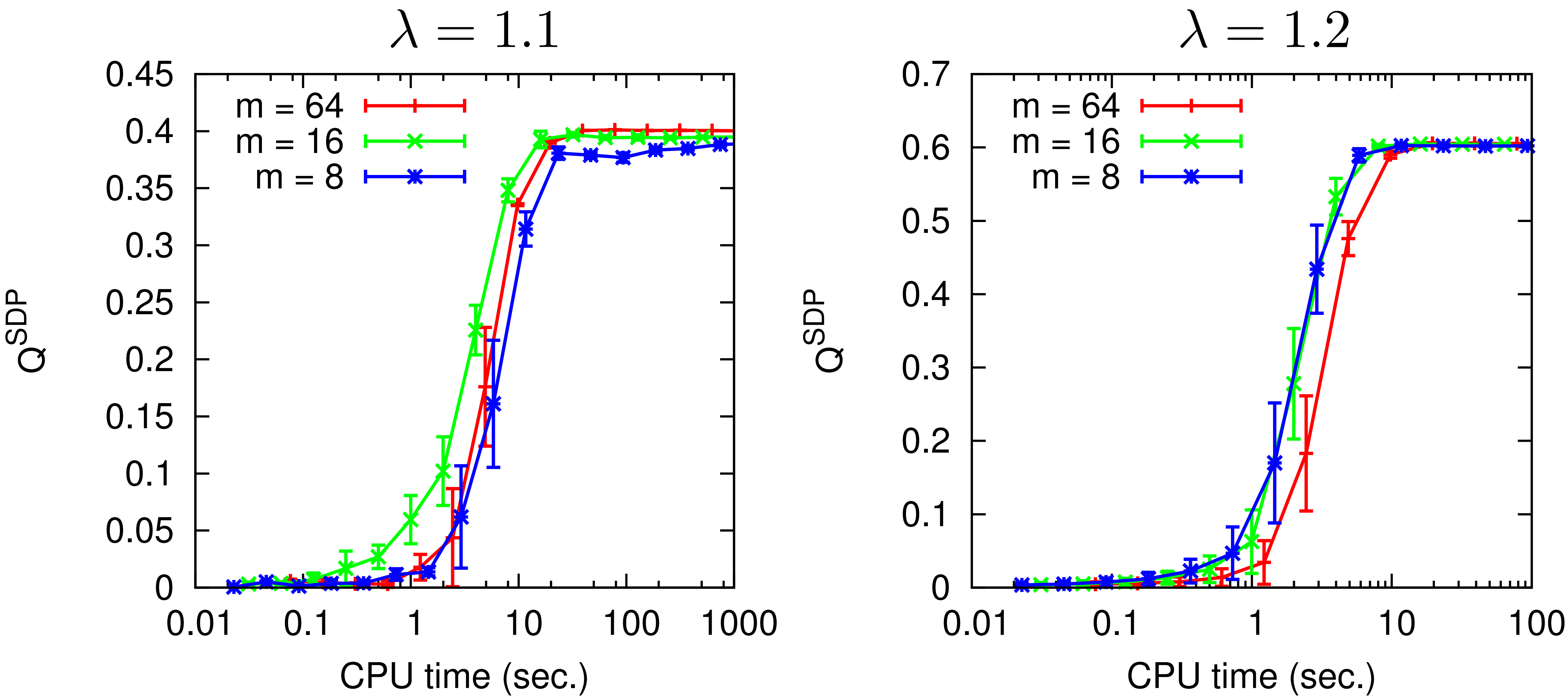
Small m values are fine!

$$n = 4 \cdot 10^4 \quad d = 3 \quad \lambda = 1.1 \quad \alpha = 0.0$$



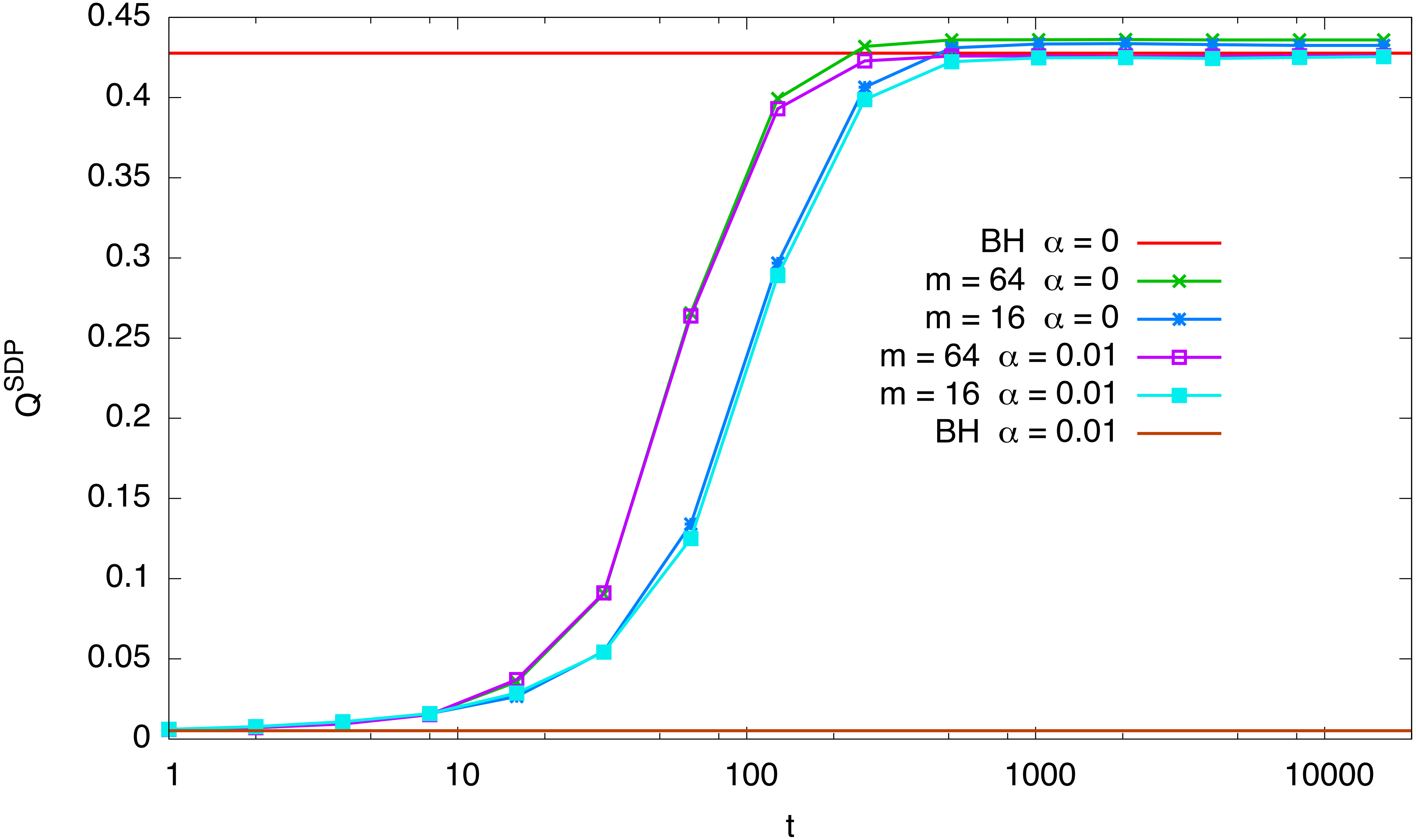
The algorithm is very fast!

$$\underline{n = 10^5} \quad d = 3$$

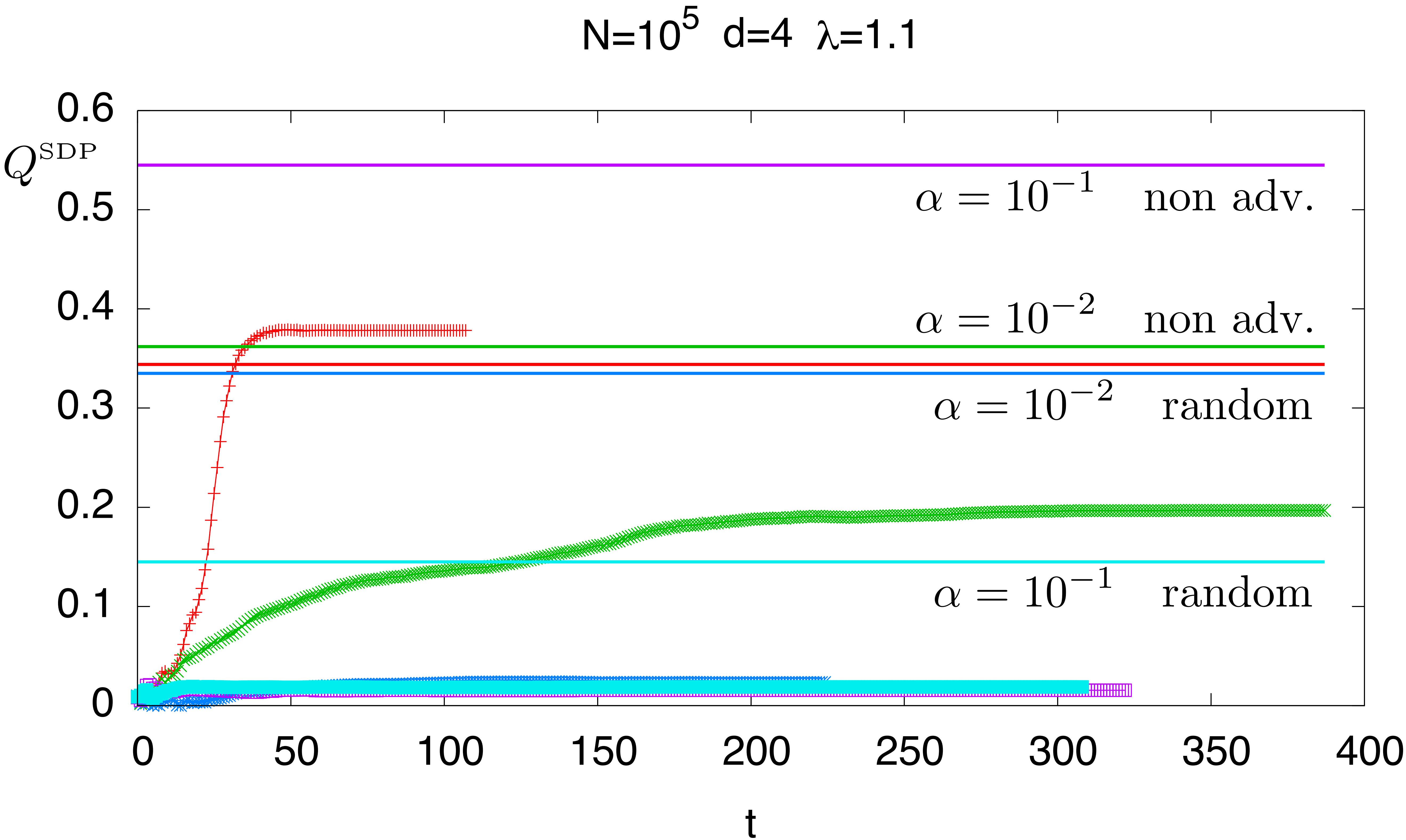


The algorithm is very robust!

$$n = 4 \cdot 10^4 \quad d = 3 \quad \lambda = 1.1$$



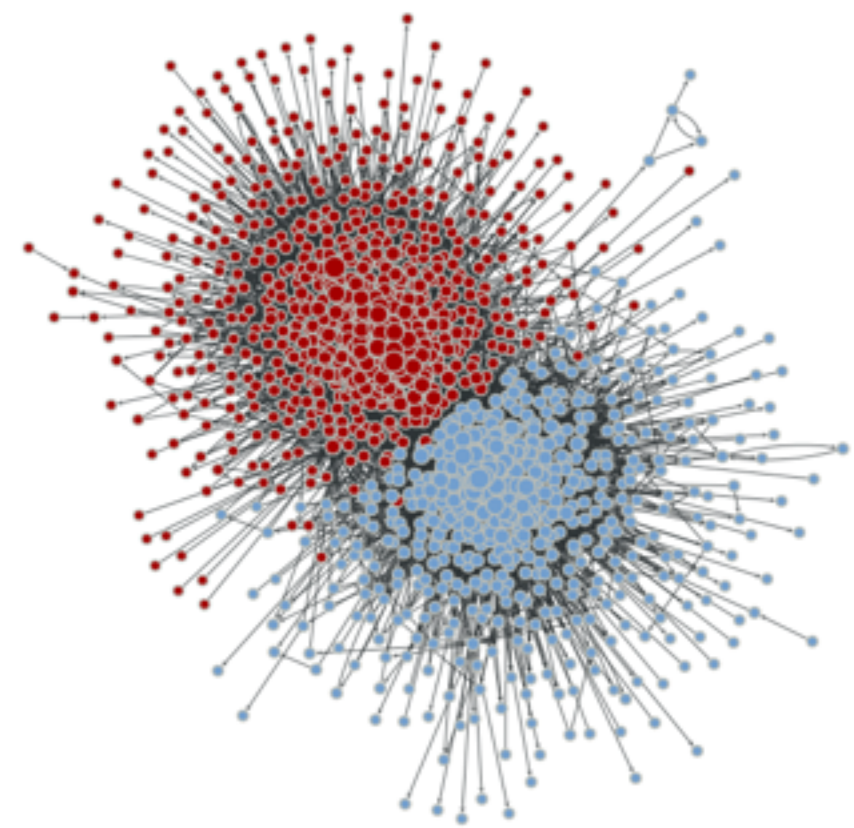
The algorithm is very robust!



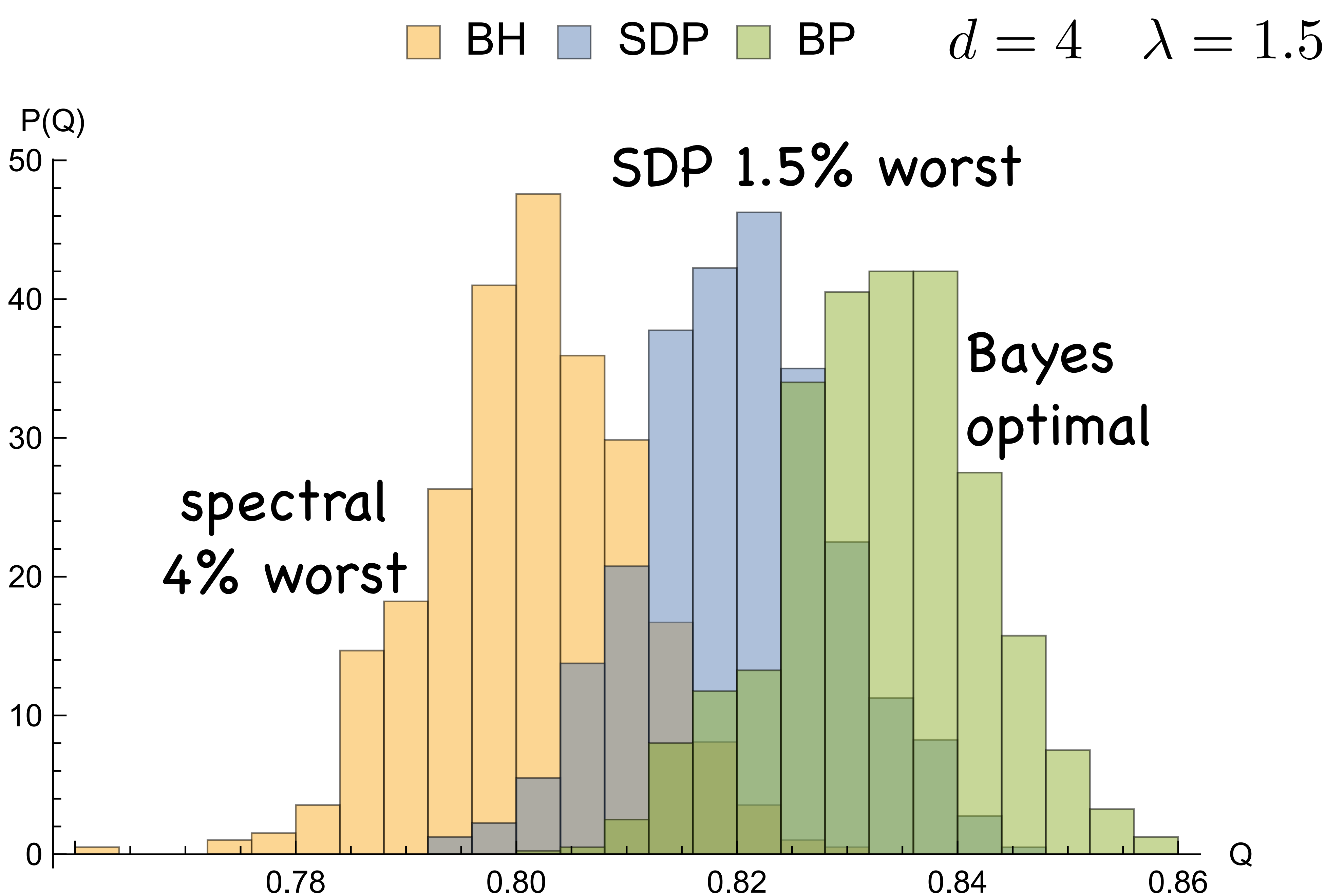
A real-world network (political blogs)

1222 nodes, 16714 edges

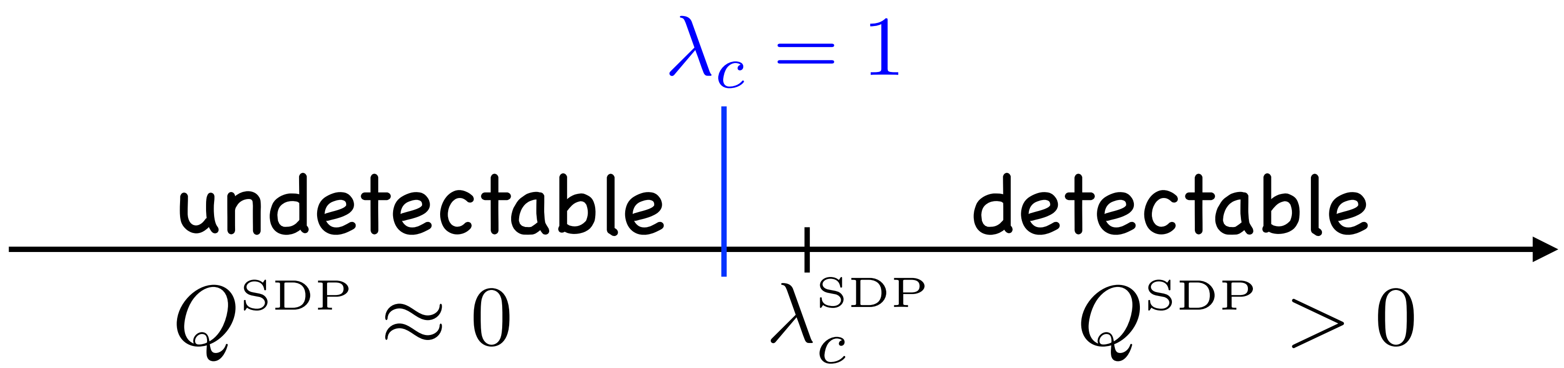
	overlap	cut size
Bethe Hessian z-Laplacian	0.865794	1271
Adjacency	0.86743	1268
X-Laplacian	0.918167	1250
Low rank SDP	0.903437	1221
“ground truth”	1.0	1575



A quantitative comparison



SDP quasi-optimality

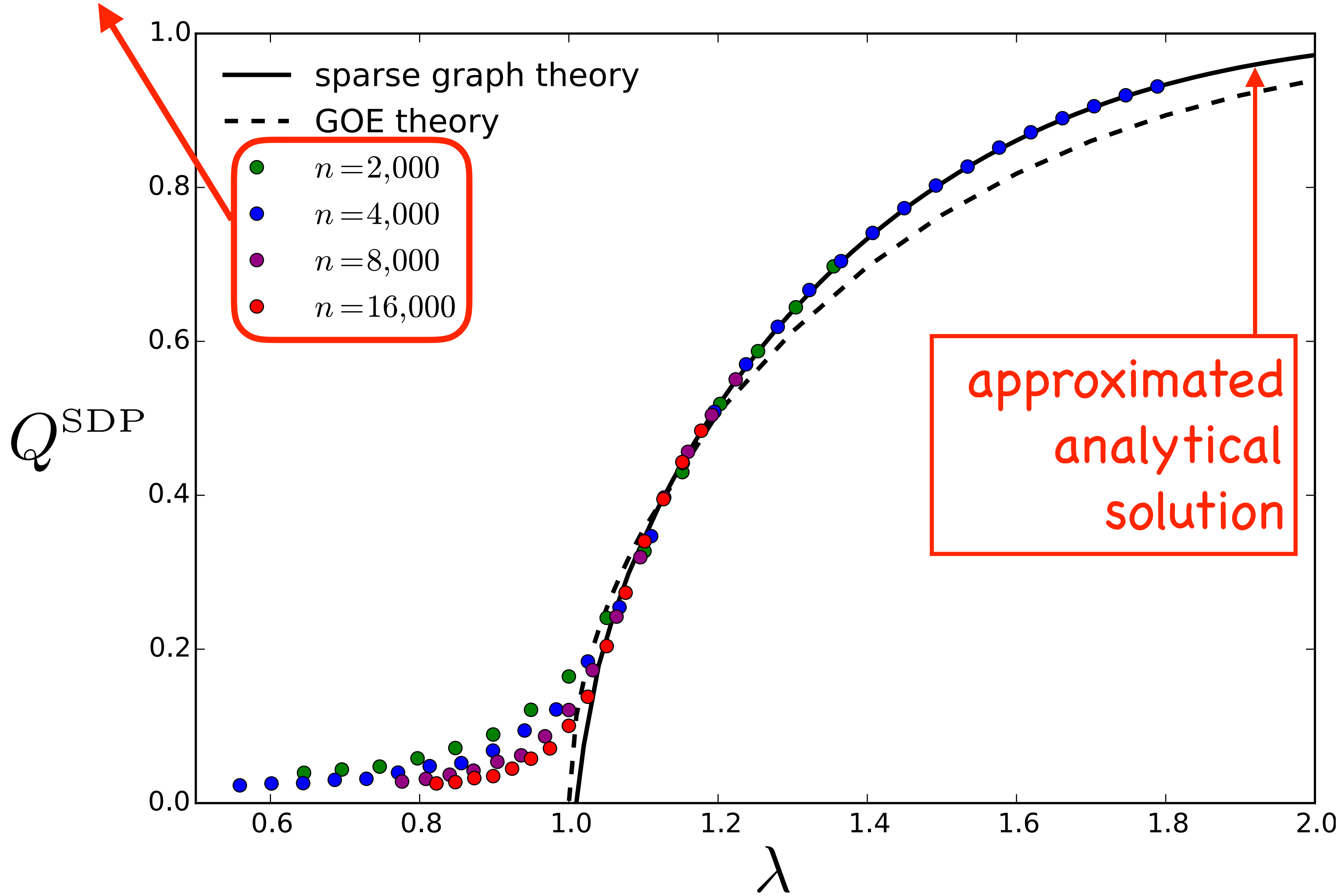


- We estimate λ_c^{SDP} by solving the statistical physics of models with m-component spin variables, in $m \to \infty$ limit
  - Running the SDP-based algorithm for very large m values (= solve the exact cavity equations)
  - Solving analytically via an approximate ansatz

SDP quasi-optimality

SDP-based algorithm
(very large m values)

$d = 5 \quad N_{\text{samples}} = 500$

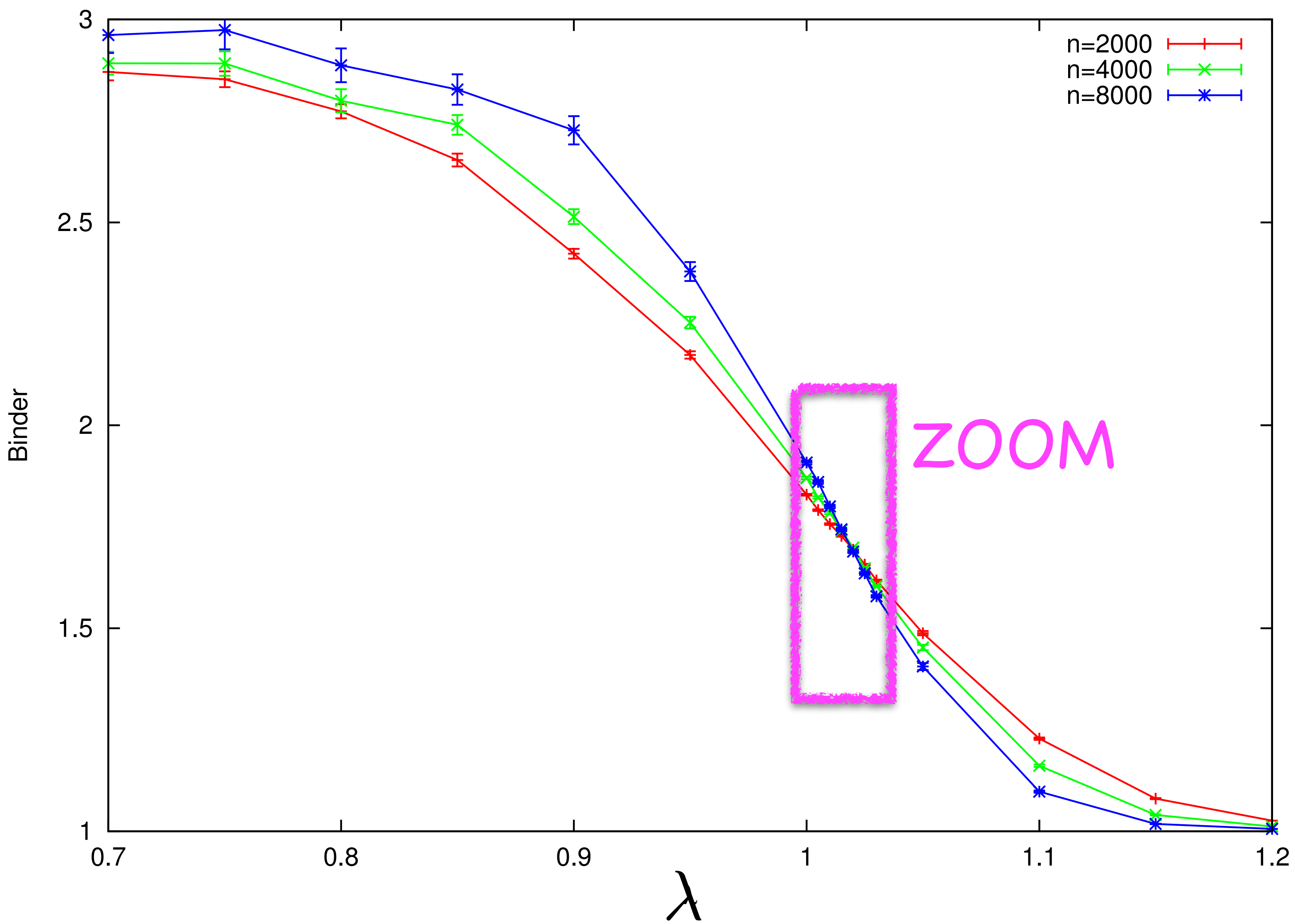


approximated
analytical
solution

$$Q^{\text{SDP}} = \frac{1}{n}\mathbb{E}\{|\langle \boldsymbol{x}^{\text{SDP}}, \boldsymbol{x}_0 \rangle|\}$$

Computing the threshold

- Crossing of the Binder cumulants to locate exactly λ_c^{SDP}



$d = 5$

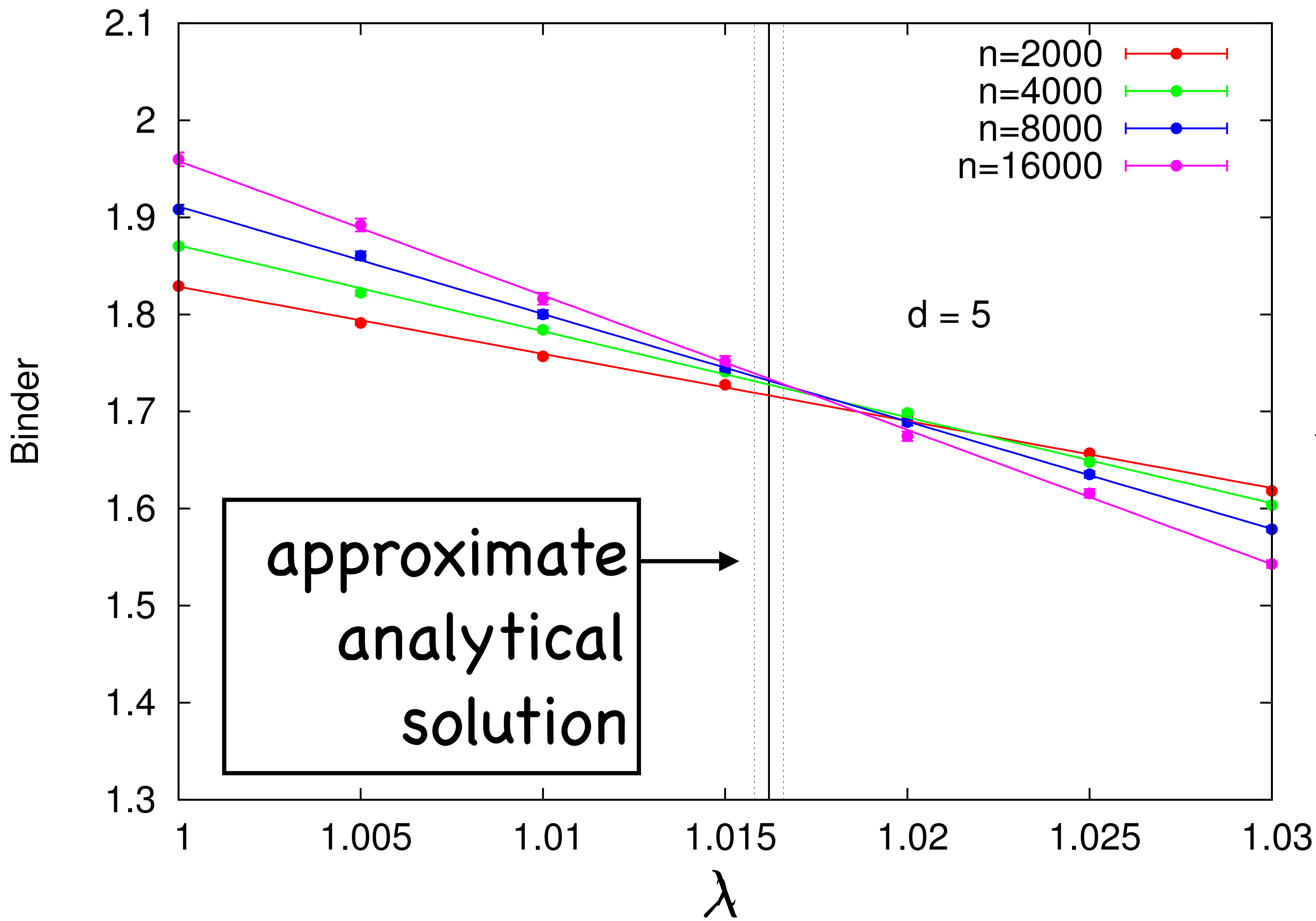
$$\boxed{N_{\mathrm{samples}} \geq 10^5}$$

$$B = \frac{n\,\mathbb{E}\{\langle \boldsymbol{x}^{\mathrm{SDP}}, \boldsymbol{x}_0\rangle^4\}}{(\mathbb{E}\{\langle \boldsymbol{x}^{\mathrm{SDP}}, \boldsymbol{x}_0\rangle^2\})^2}$$

SDP-based algorithm for very large m values

Computing the threshold

- Crossing of the Binder cumulants to locate exactly λ_c^{SDP}



$$d = 5$$

$$\boxed{N_{\text{samples}} \geq 10^5}$$

$$B = \frac{n\,\mathbb{E}\{\langle \boldsymbol{x}^{\mathrm{SDP}}, \boldsymbol{x}_0\rangle^4\}}{(\mathbb{E}\{\langle \boldsymbol{x}^{\mathrm{SDP}}, \boldsymbol{x}_0\rangle^2\})^2}$$

SDP-based algorithm for very large m values

Statistical physics analytical approach

- Unified framework: statistical physics models with m-component variables: $\underline{x}_i \in \mathbb{R}^m \, , ||\underline{x}_i|| = 1$

$$P(\underline{\boldsymbol{x}}) = \frac{1}{Z} \exp\left[\beta \sum_{(ij)\in E} \underline{x}_i \cdot \underline{x}_j\right]$$

- Bayes: $m = 1, \quad \tanh(\beta) = \lambda/\sqrt{d}$
- ML: $m = 1, \quad \beta \to \infty$
- SDP: $m \to \infty, \quad \beta \to \infty$

Statistical physics analytical approach

- Ansatz for the marginals in m-component <u>dense</u> models

$$P_i(\underline{x}_i) = \frac{1}{Z_i} \exp\left[2m\beta(\underline{\xi}_i^{\mathsf{T}} \underline{x}_i + \underline{x}_i^{\mathsf{T}} \boldsymbol{C}_i \underline{x}_i)\right]$$

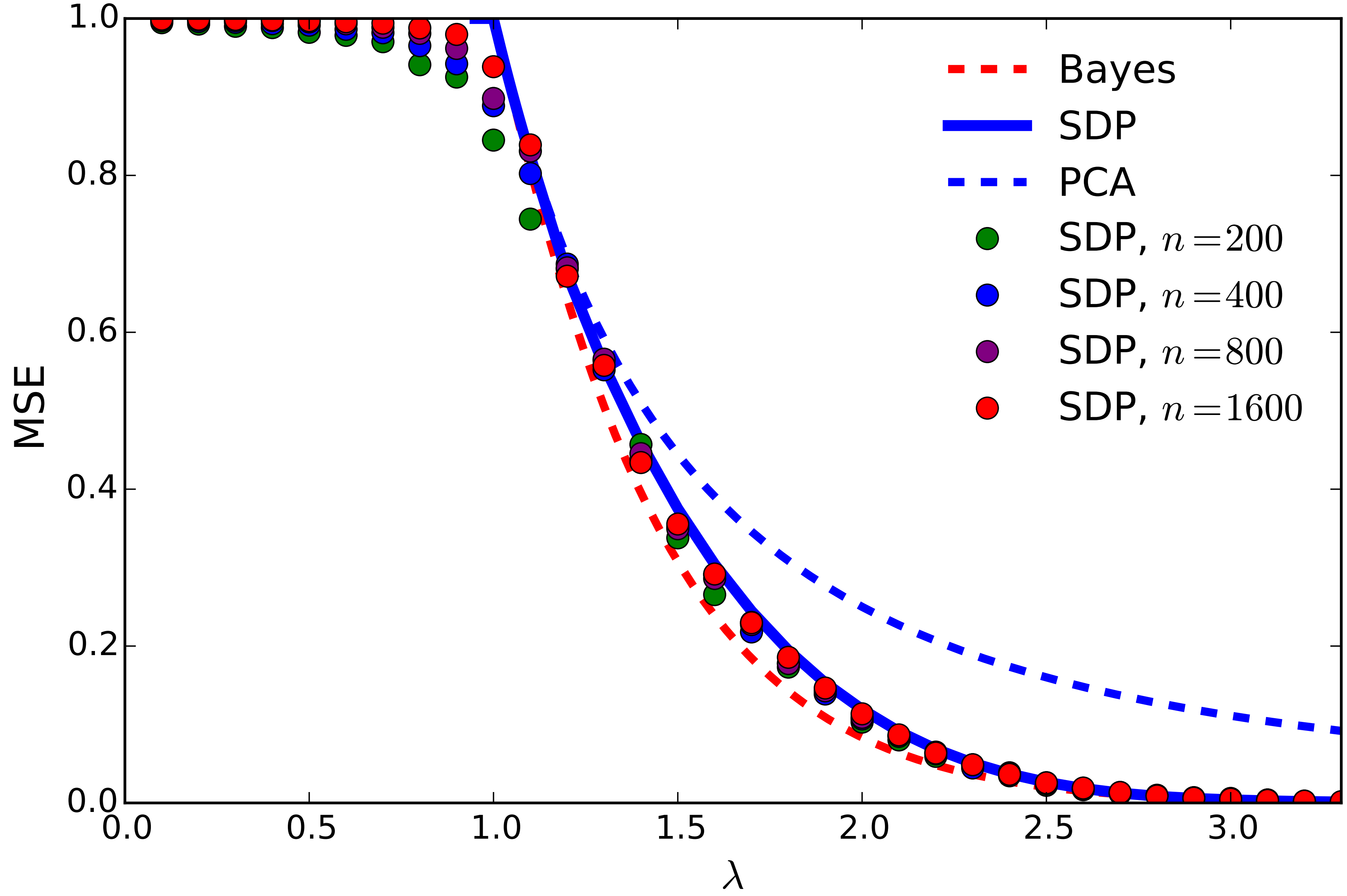
$$\underline{x}_i \in \mathbb{F}^m \,, ||\underline{x}_i|| = 1 \qquad \underline{\xi}_i \sim \mathcal{N}(\underline{\mu}, \boldsymbol{Q}) \qquad \boldsymbol{C}_i = \boldsymbol{C}$$

- Self consistency equations in the dense case

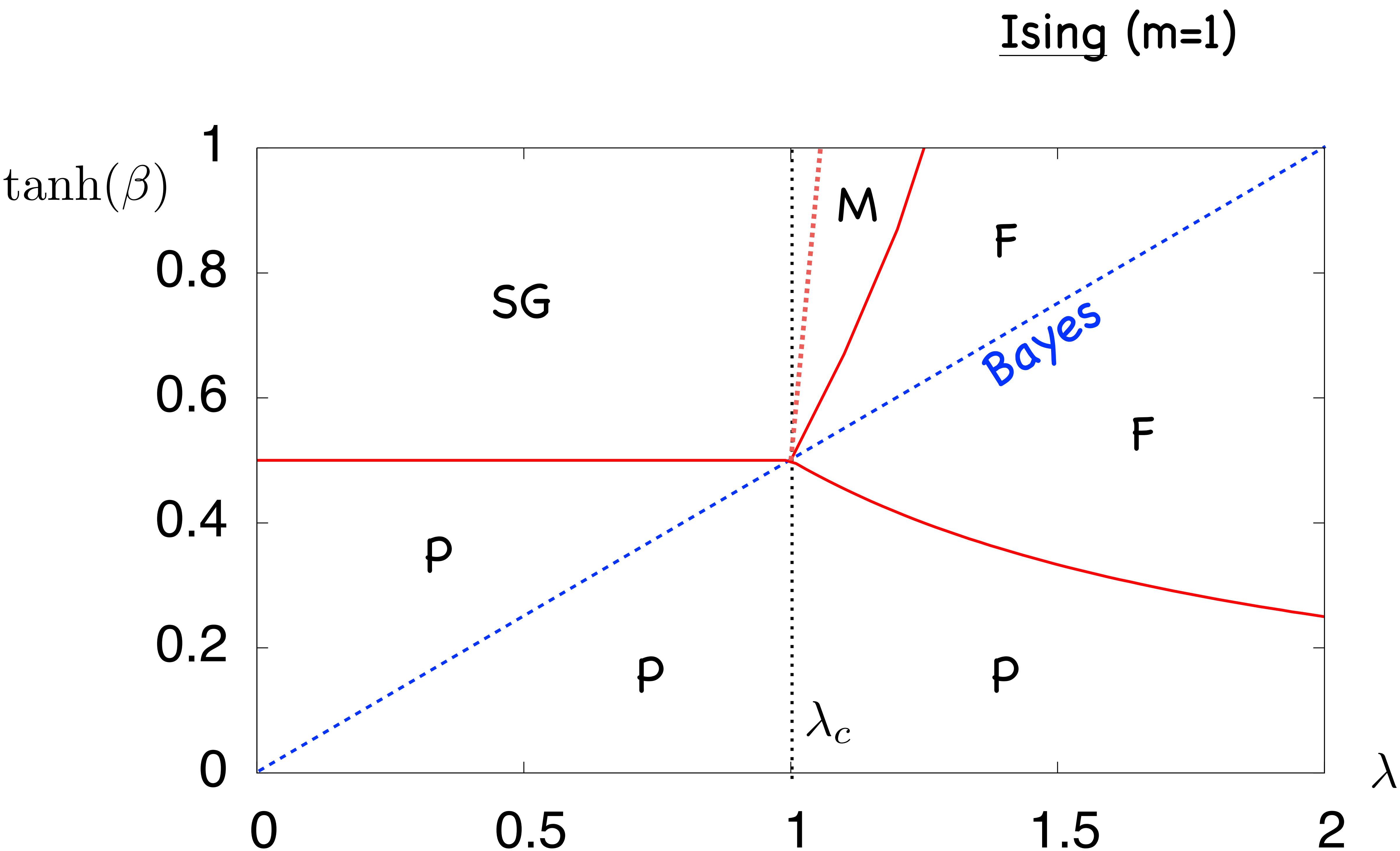
$$\begin{aligned}
\underline{\mu} &= \lambda\, \mathbb{E}[\langle \underline{x} \rangle] \\
\boldsymbol{Q} &= \mathbb{E}[\langle \underline{x} \rangle \langle \underline{x}^{\mathsf{T}} \rangle] \\
\boldsymbol{C} &= \beta\, m\, \mathbb{E}[\langle \underline{x}\, \underline{x}^{\mathsf{T}} \rangle - \langle \underline{x} \rangle \langle \underline{x}^{\mathsf{T}} \rangle]
\end{aligned}$$

Analytical solution: dense real case

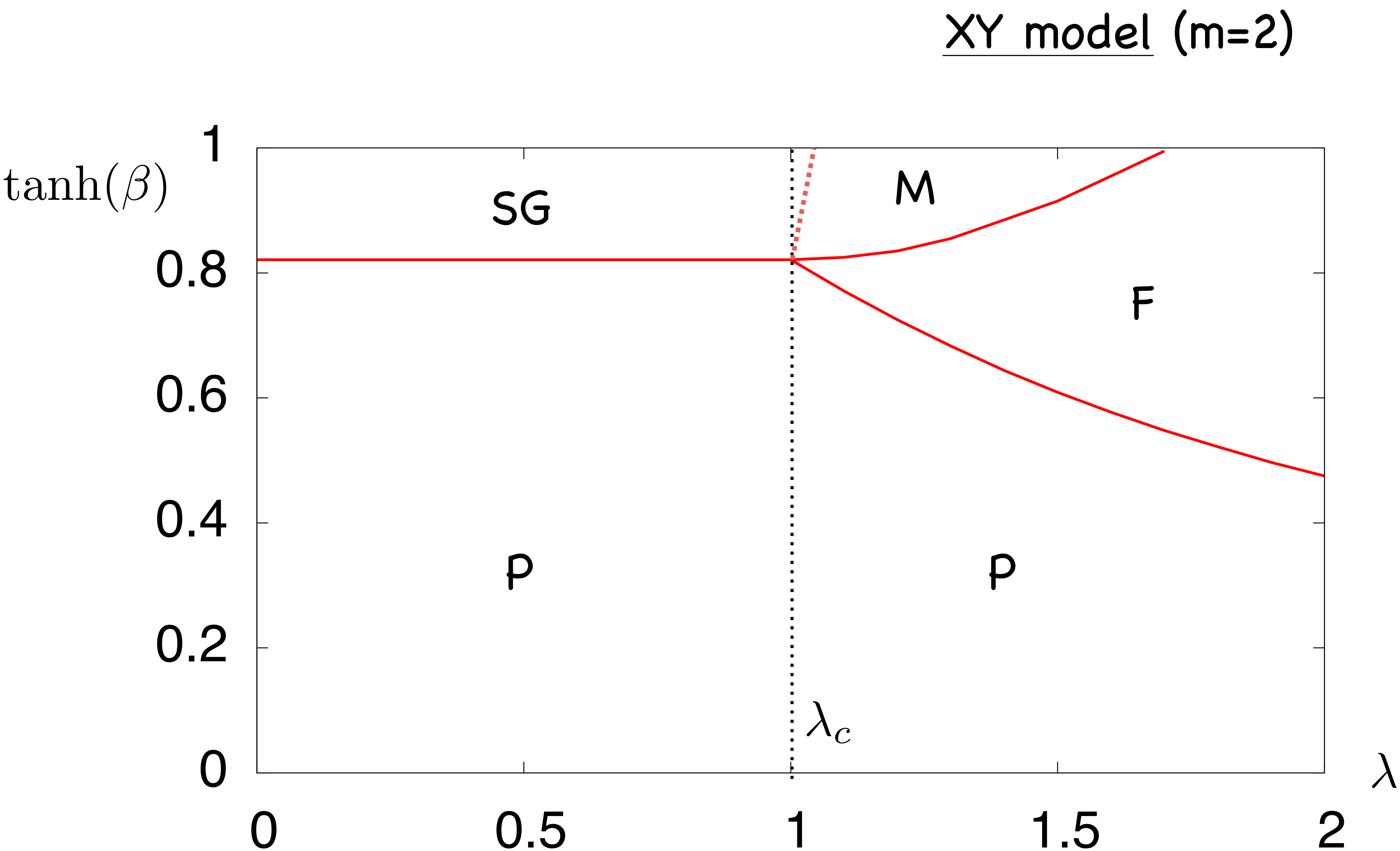
$$\mathrm{MSE}_n(\hat{\boldsymbol{x}}) \equiv \frac{1}{n}\mathbb{E}\{\min_{s\in\{+1,-1\}} \|\hat{\boldsymbol{x}}(\boldsymbol{Y}) - s\,\boldsymbol{x_0}\|_2^2\}$$



Phase diagrams in the sparse case (SBM d=4)



Phase diagrams in the sparse case (SBM d=4)



Approximate analytical solution (SBM)

- In the recovery phase we assume the O(m) symmetry to break along the first component, while preserving O(m-1)

$$\underline{x}_i = (s_i, \boldsymbol{\tau}_i),\ s_i \in \mathbb{R},\ \boldsymbol{\tau}_i \in \mathbb{R}^{m-1}$$

- We write the marginal for $\underline{x}_i$ as

$$\exp\left\{2\beta\sqrt{m\mathsf{c}_i}\,\langle \boldsymbol{z}_i, \boldsymbol{\tau}_i\rangle + 2\beta m h_i\, s_i - \beta m r_i s_i^2 + O_m(1)\right\}\delta\Big(s_i^2 + \|\boldsymbol{\tau}_i\|_2^2 - 1\Big)$$

with $\boldsymbol{z}_i \sim \mathsf{N}(0, \mathrm{I}_{m-1})$

- Approximate because the z_i are correlated
- It should be valid in the limits $d \to 1$ and $d \to \infty$

Approximate analytical solution (SBM)

$$\exp\left\{2\beta\sqrt{m\mathsf{c}_i}\,\langle \boldsymbol{z}_i, \boldsymbol{\tau}_i\rangle + 2\beta m h_i\, s_i - \beta m r_i s_i^2 + O_m(1)\right\}\delta\Big(s_i^2 + \|\boldsymbol{\tau}_i\|_2^2 - 1\Big)$$

- Cavity method -> self consistency equation for marginals

$$\mathsf{c}_0 = \sum_{i=1}^{k} \frac{\mathsf{c}_i}{\rho_i^2}\,,$$

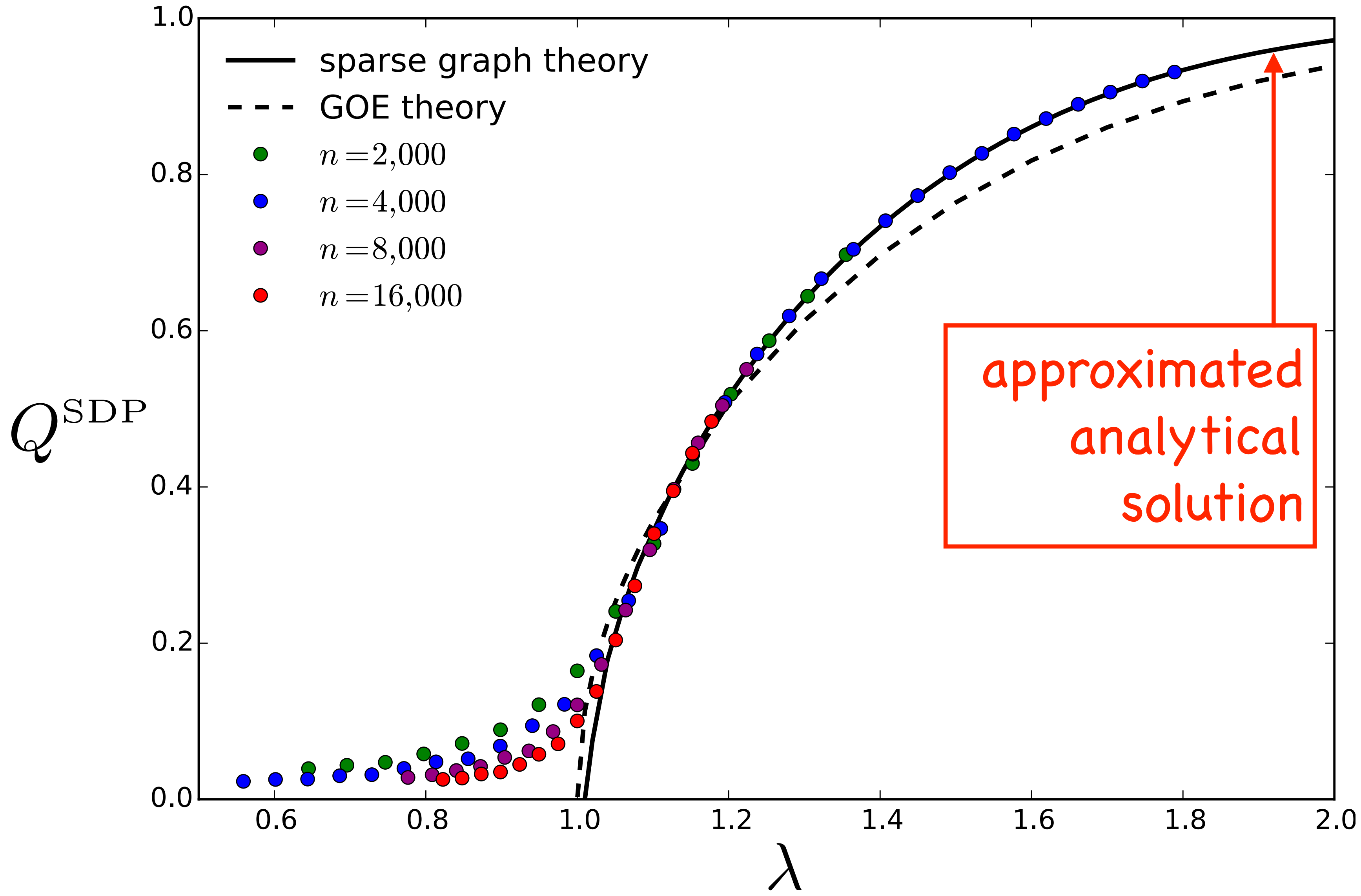
$$1 = \frac{h_i^2}{(\rho_i + r_i)^2} + \frac{1+\mathsf{c}_i}{\rho_i^2}$$

$$h_0 = \sum_{i=1}^{k} \frac{h_i}{\rho_i + r_i}\,,$$

$$r_0 = \sum_{i=1}^{k} \left\{\frac{1}{\rho_i} - \frac{1}{\rho_i + r_i} + \left(1 + \frac{(1+\mathsf{c}_i) r_i}{\rho_i^3}\right)^{-1} \frac{h_i^2}{(\rho_i + r_i)^3}\right\}$$

- Solve by population dynamics
- At the fixed point $Q^{\mathrm{SDP}} = \mathbb{E}[\mathrm{sign}(h^*)]$

Approximate analytical solution (SBM)



Approximate analytical solution (SBM)

- Linearize the cavity equations to locate the threshold
- To linear order in $h \implies r_i = 0$

$$\mathsf{c}_0 = \sum_{i=1}^{k} \frac{\mathsf{c}_i}{\rho_i^2}\,, \qquad 1 = \frac{h_i^2}{(\rho_i + r_i)^2} + \frac{1+\mathsf{c}_i}{\rho_i^2}$$

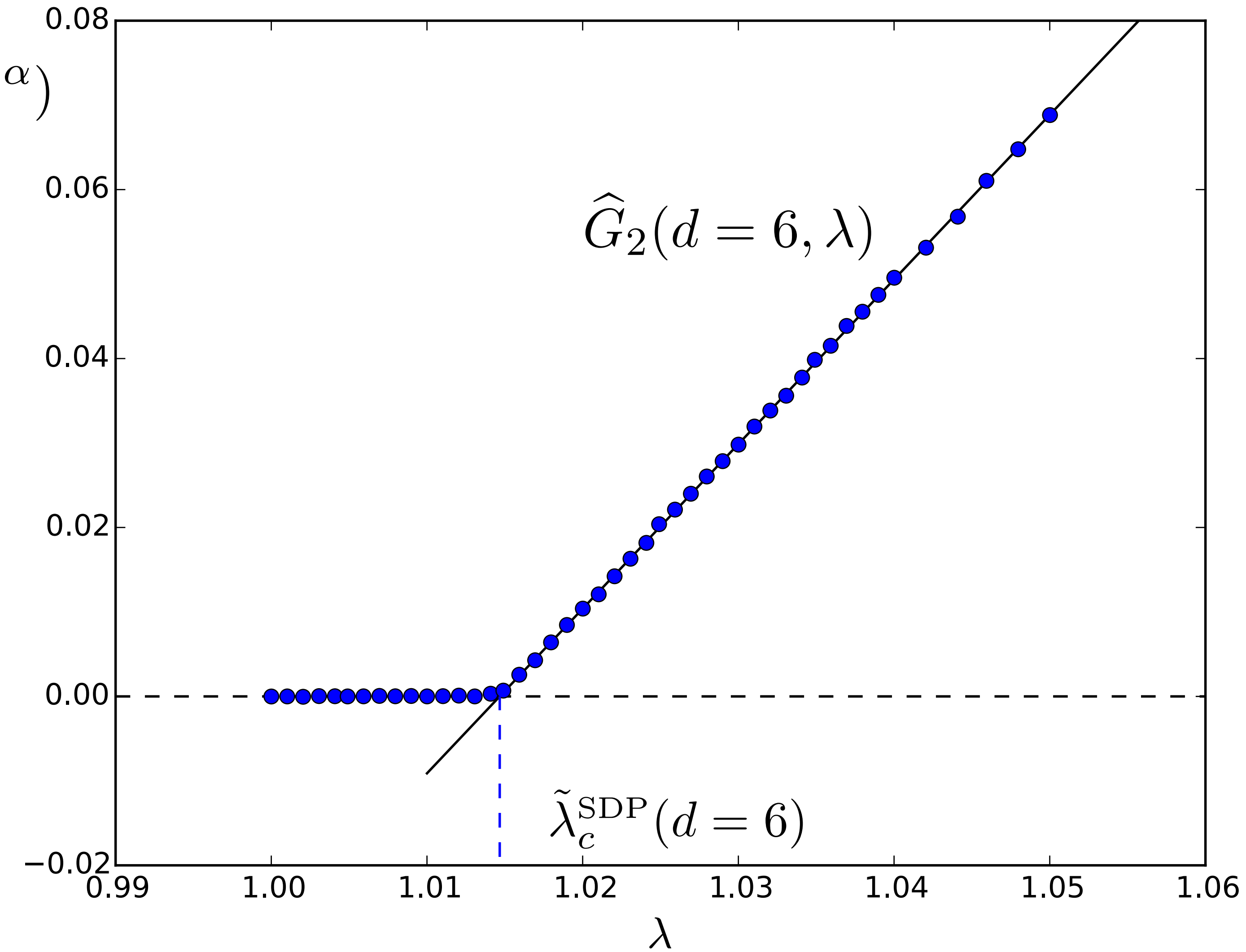
$$h_0 = \sum_{i=1}^{k} \frac{h_i}{\rho_i + r_i}\,,$$

$$r_0 = \sum_{i=1}^{k} \left\{ \frac{1}{\rho_i} - \frac{1}{\rho_i + r_i} + \left(1 + \frac{(1+\mathsf{c}_i) r_i}{\rho_i^3}\right)^{-1} \frac{h_i^2}{(\rho_i + r_i)^3} \right\}$$

$$\mathsf{c}_0 = \sum_{i=1}^{k} \frac{\mathsf{c}_i}{1+\mathsf{c}_i}\,, \qquad h_0 = \sum_{i=1}^{k} \frac{h_i}{\sqrt{1+\mathsf{c}_i}}$$

Approximate analytical solution (SBM)

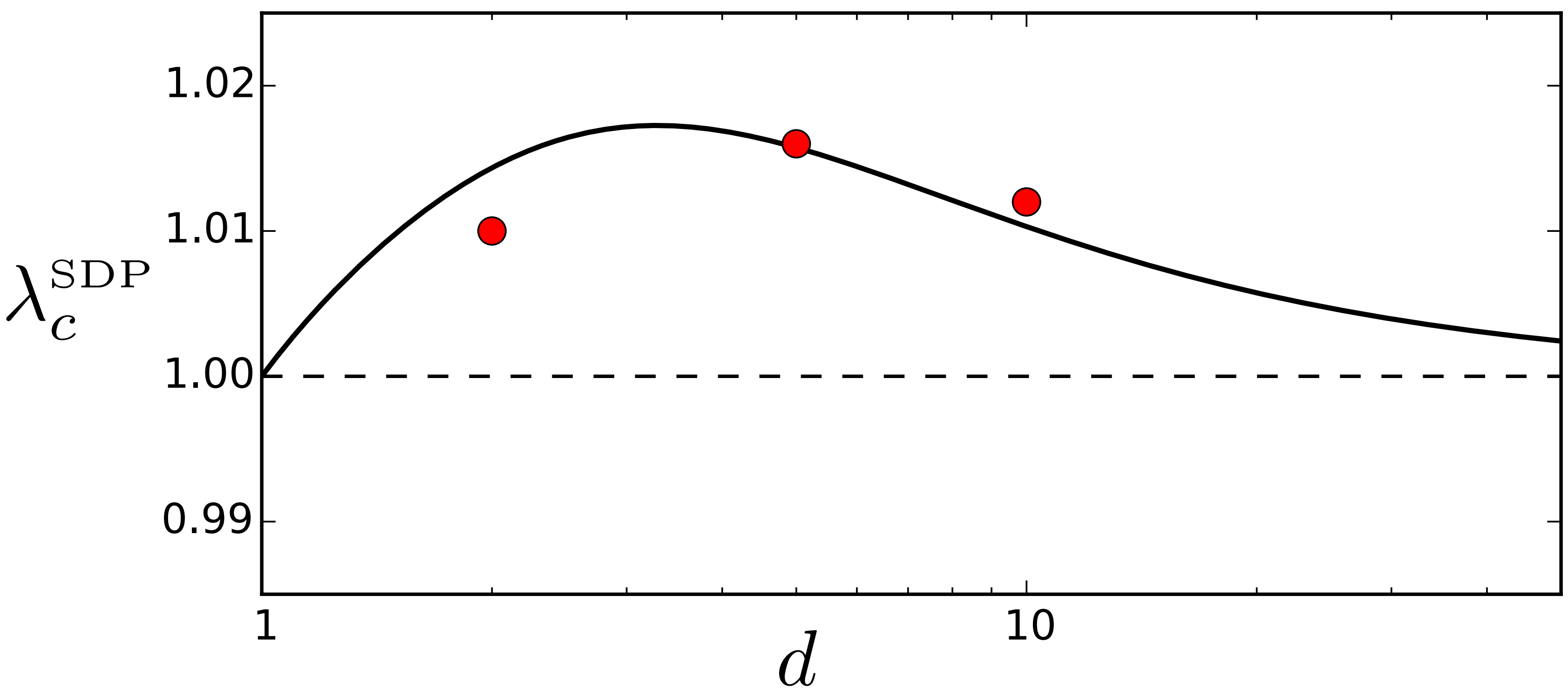
$$G_\alpha(d,\lambda) \equiv \lim\inf_{t\to\infty} \frac{1}{t\alpha} \log \mathbb{E}(|h^t|^\alpha)$$



$$\left(\mathsf{c}^{t+1}; h^{t+1}\right) \stackrel{\mathrm{d}}{=} \Big(\sum_{i=1}^{L_+ + L_-} \frac{\mathsf{c}_i^t}{1+\mathsf{c}_i^t}; \sum_{i=1}^{L_+ + L_-} \frac{s_i h_i^t}{\sqrt{1+\mathsf{c}_i^t}} \Big)$$

Analytical solution: sparse case (SBM)

- SDP at most 2% sub-optimal!



- Red points: numerical solution of the replica/cavity equations (crossing of Binder cumulants)
- Black line: approximated analytical solution

Some conclusions...

- SDP relaxations are very effective:
  - robust and quasi-optimal
  - may outperform spectral relaxations
- Better than SDP are SDP-inspired algorithms (small m)
  `http://web.stanford.edu/~montanar/SDPgraph/`
- It is worth studying the statistical physics of models with m-component variables:
  - unifying framework to study and solve several estimators in statistical inference
  - different physics, better algorithms