Cluster size constrained network partitioning

Maksim Mironov[1] Konstantin Avrachenkov[2]

[1]Moscow Institute of Physics and Technology, Moscow, Russia

[2]Inria Sophia Antipolis, Nice Area, France

Stochastic Block Model

- We have a random graph $G_{sbm} = (V, E)$ with two blocks V_1, V_2, where $V = V_1 \sqcup V_2$
- For each pair of nodes $\{v, u\}$ an edge is drawn independently according to

$$\mathrm{P}\left(\{v, u\} \in E\right) = \begin{cases} p_1, & v, u \in V_1, \\ p_2, & v, u \in V_2, \\ q, & \text{overwise}, \end{cases}$$

- For simplicity of presentation we assume the case of equal blocks, $|V_1| = |V_2|$.
- Values p_1, p_2, q are considered as functions of n.

Mean-field Stochastic Block Model

- We have a **full** graph $G_{mf} = (V, E)$ with two blocks V_1, V_2, where $V = V_1 \sqcup V_2$
- For each pair of nodes $\{v, u\}$ an edge has its weight according to

$$Weight\left(\{v, u\}\right) = \begin{cases} p_1, & v, u \in V_1, \\ p_2, & v, u \in V_2, \\ q, & \text{overwise}, \end{cases}$$

- There is no need to partition the G_{mf} graph – everything is already clear

Formal description of the problem

- We have a graph $G(V, E)$ and set of configurations

$$\{-1, 1\}^V \supset \Sigma = \left\{ \sigma \mid \sum_{v \in V} \sigma(v) = 0 \right\}$$

  here $\sigma(v)$ denotes label of v.
- We introduce the **global energy** of the configuration as

$$\varepsilon(\sigma) = - \sum_{\{u,v\} \in E} \sigma(u)\sigma(v). \tag{1}$$

- We naturally introduce a **local energy** defined by

$$\varepsilon(\sigma, v) = -\sigma(v) \sum_{w \sim v} \sigma(w). \tag{2}$$

Algorithm

1. Choose $n/2$ nodes to have label -1 at random and $n/2$ others are set with 1
2. Choose randomly a pair of nodes with different labels
3. Calculate the sum of their local energies ε_1 as if they are labeled as it is, and ε_2 in case they swap their labels
4. Choose either original labels or swapped ones based on flip of a biased coin with probability

$$p = \frac{\exp(-\beta\varepsilon_1)}{\exp(-\beta\varepsilon_1) + \exp(-\beta\varepsilon_2)}$$

5. If stop criteria is not met go to step 2
6. Iterate over all nodes and update the label of each of them based on weighted majority of its neighbours

Theoretical running time of Algorithm for G_{mf}

Theorem

Let $p_1 + p_2 > 2q$ and $p_1 > q$. Then, the expected number of steps T to obtain the exact global optimum in the mean-field SBM is upper bounded by

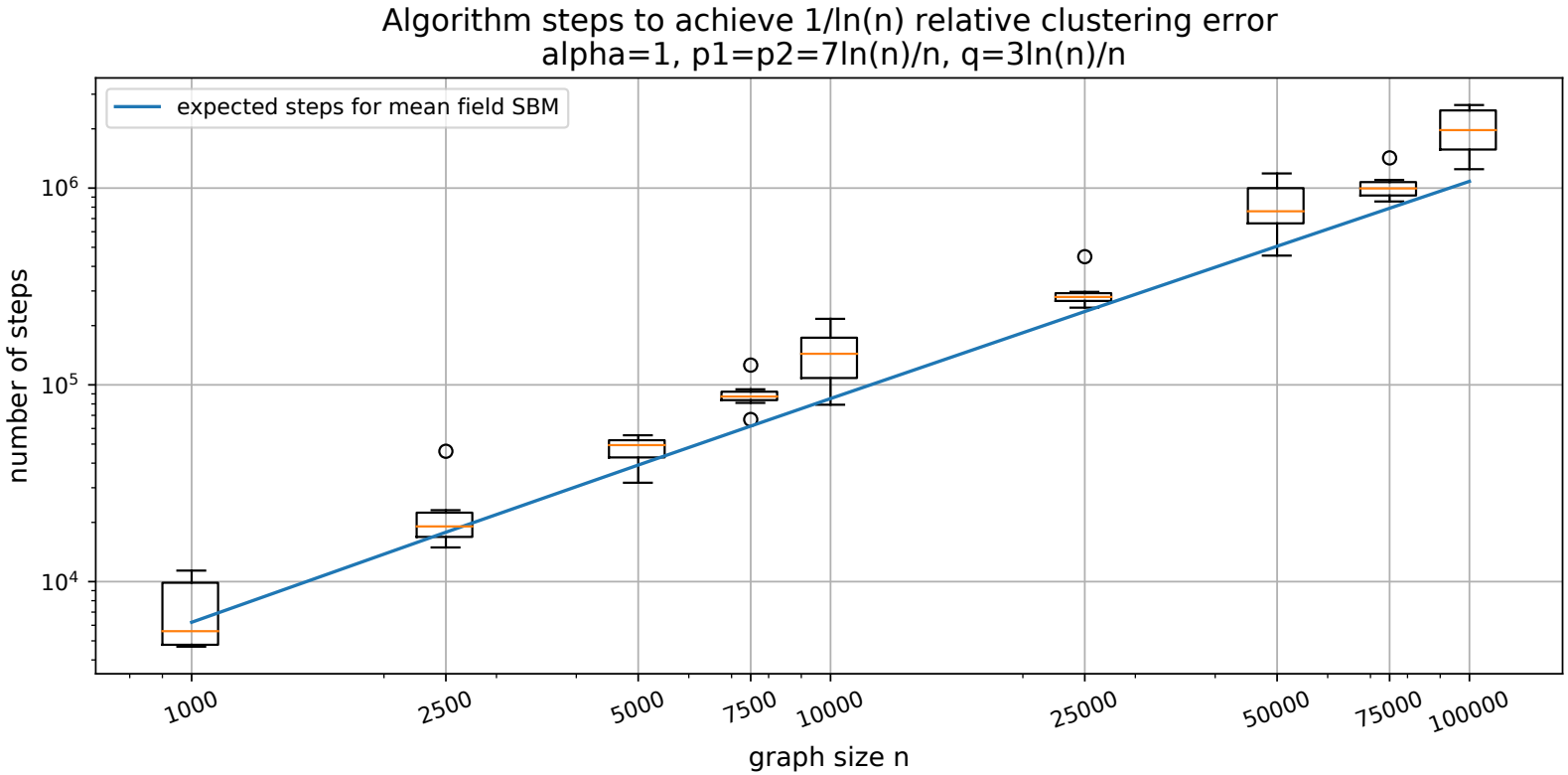
$$\mathrm{E}T = O(n^2)$$

Theorem

Let $p_1 + p_2 > 2q$ and relative clustering error $\delta = o(1)$. Then, the expected number of steps T to obtain the almost exact global optimum in the mean-field SBM is upper bounded by

$$\mathrm{E}T = O\left(\frac{n}{\delta}\right)$$

Simulations for SBM



Simulations for SBM

$$n = 10000, \alpha = 1, p = \frac{a \ln(n)}{n}, q = \frac{b \ln(n)}{n}$$

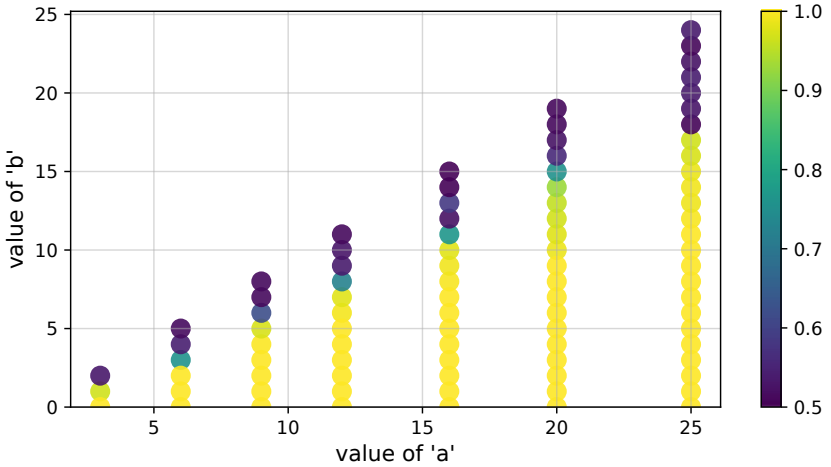


Figure: Our algorithm

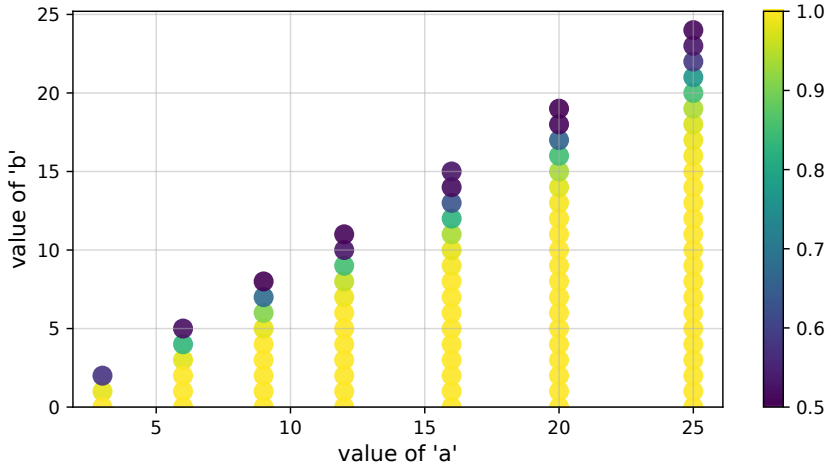


Figure: Spectral clustering

Advantages and drawbacks

Advantages:

- the running time complexity of the algorithm is roughly $\bar{d} \cdot n/\delta$ for the SBM graphs;
- the algorithm can be effectively distributed over any number of machines with shared memory and with no need in synchronization;
- the approach can be customized with different objective functions.

Drawbacks:

- the output of the algorithm is not reproducible, it is a result of a random process;
- for the extremely difficult problems it works worse than the spectral clustering in the case of balanced clusters.

Thanks for the attention!