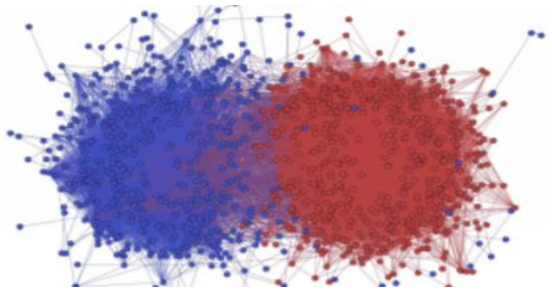
Cluster-Size Constrained Network Partitioning (M. Mironov, K. Avrachenkov)

We consider network partitioning with the desired cluster sizes, use the technique similar to Glauber Dynamics and apply it for the SBM model.



Stochastic Block Model is defined as follows

- 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

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

We also define the mean-field SBM 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

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

We consider labels for each node from {-1, 1} and thus have set of configurations

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

We measure clustering result with *global energy* defined by

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

The contribution of a single node to energy we denote as *local energy* and define by

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

The algorithm we offer does the discrete optimization where the simple step is the swap of labels of two arbitrary nodes in case it reduces the global energy.

The algorithm has purely local nature, which means it can be distributed over any number of machines with shared memory and we emphasize that.

So, the algorithm is described as follows

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

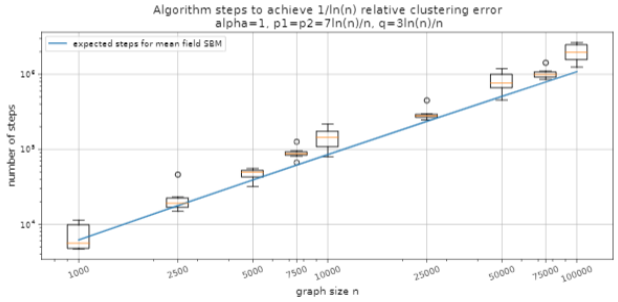
Stop criteria might be just the number of steps.
For the mean-field SBM we have theoretical upper bound on running time:

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

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

For the SBM graph simulations show almost the same time complexity. Alpha=1 means equal blocks



In terms of accuracy algorithm is comparable with Spectral clustering, which is proven to be optimal for SBM model (but works in $O(n^3)$):

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

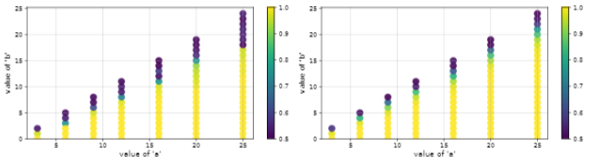


Figure: Our algorithm

Figure: Spectral clustering

To conclude:

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.