# Graph clustering

Quentin Moyedpour [1] Florian Oubaha [2]

[1]qmoayedpour@ensae.fr [2]florian.oubaha@polytechnique.edu

## Stochastic Block Model

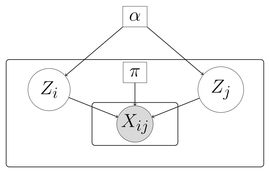



Figure 1. Latent Model

- $X$ is the adjacency matrix.
- $Z_{i,q}$: the probability of a vertex $i$ belonging to cluster $q$.
- $\tau_{i,q} = P(Z_{i,q} = 1|X)$.
- $\alpha_q = P(Z_{i,q} = 1)$: the prior probability for vertex $i$ to belong to $q$.
- $\pi \in [0,1]^{Q\times Q}$, satisfying:

$$X_{i,j}|\{i \in q, j \in l\} \sim \mathcal{B}(\pi_{q,l})$$

## EM-algorithm

The log-likelihood of the SBM is given by:

$$\log \mathcal{L}(\mathcal{X}) = \sum_{i=1}^{n}\sum_{q=1}^{Q} Z_{iq}\log(\alpha_q) + \frac{1}{2}\sum_{i\neq j}\sum_{q,l} Z_{iq}Z_{jl}\pi_{ql}$$

Using a variational approach, we want to maximize this likelihood minus the Kullback-Leibler divergence $\mathcal{J}(R_{\mathcal{X}})$.

$$\begin{aligned} J(R_{\mathcal{X}}) &= \log L(\mathcal{X}) - \mathrm{KL}[R_{\mathcal{X}}(\cdot), P_r(\cdot|\mathcal{X})] \\ &= \sum_i\sum_q \tau_{iq}\log\alpha_q + \frac{1}{2}\sum_{i\neq j}\sum_q \tau_{iq}\tau_{jq}\log\left(\pi_q^{X_{i,j}}(1-\pi_q)^{1-X_{i,j}}\right) - \sum_i\sum_q \tau_{iq}\log\tau_{iq}. \end{aligned}$$

One can show that given $\boldsymbol{\tau}$, the parameters $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\pi}}$ that maximize $\mathcal{J}(R_{\mathcal{X}})$ are:

$$\left.\begin{aligned} \hat{\alpha}_q &= \frac{1}{n}\sum_{i=1}^{n}\tau_{iq}, \\ \hat{\pi}_{ql} &= \frac{\sum_{i\neq j}\tau_{iq}\tau_{jl}X_{ij}}{\sum_{i\neq j}\tau_{iq}\tau_{jl}} \end{aligned}\right\} \text{M-Step}$$

Similarly, given $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$, the parameters $\hat{\boldsymbol{\tau}}$ that maximize $\mathcal{J}(R_{\mathcal{X}})$ are:

$$\left.\hat{\tau}_{iq} \propto \alpha_q \prod_{j\neq i}\prod_l b(X_{ij};\pi_{ql})^{\hat{\tau}_{jl}}\right\} \text{E-Step}$$

Which is estimated by Fixed Point Algorithm

## Choice of the number of clusters

We choose $Q$ that minimizing the following Information Criterion Loss:

$$\mathrm{ICL}(m_Q) = \max_{\theta}\mathcal{L}(\mathcal{X} \mid \theta, m_Q) - \frac{Q-1}{2}\log(n) - \frac{1}{2}\times\frac{Q(Q+1)}{2}\log\frac{n(n-1)}{2}$$

by testing for all $Q$ between two values $Q_{min}$ and $Q_{max}$.

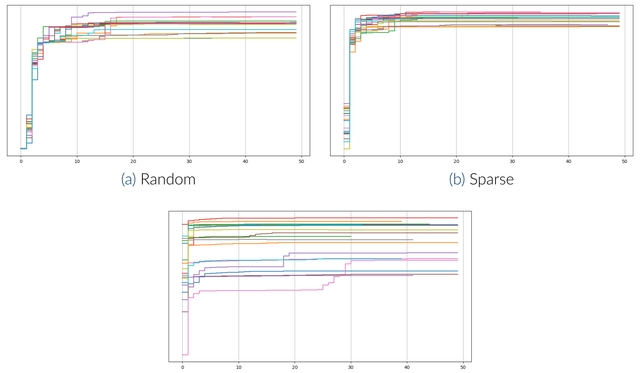
## Initialization of the EM-algorithm

Since the EM algorithm has no guarantee of converging to a global minimum, different initialization strategies generate diverse starting points, leading to variations in the exploration of local optima based on the chosen generation method.

1. **Uniform**. $\tau_{i,q}$ follow an uniform law on $[0,1]$ and then we do the normalization :

$$\tau_{i,q} \leftarrow \frac{\tau_{i,q}}{\sum_{q=1}^{Q}\tau_{i,q}}$$

2. **Sparse**. For each $i$, before the normalisation we choose a given number $S < Q$ and then we choose $S$ random indice $q$ such that $\tau_{i,q} = 0$.
3. **K-means**. We use a K-means algorithm to initialize the weights.



(a) Random (b) Sparse

(c) K-means

Figure 2. Log-likelihood paths on different initializations

## Results on synthetic graphs

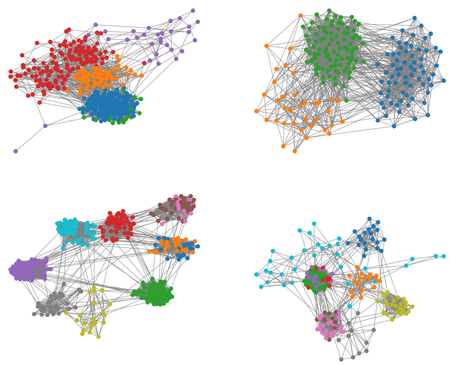


Figure 3. Predictions on synthetic graphs.

## Results on a real-world graph

The graph *sp-school-day-2* measures face-to-face contact patterns in a primary school. This graph has 238 nodes, 5539 edges and 11 communities.

We obtains these classifications depending on the initialization method:

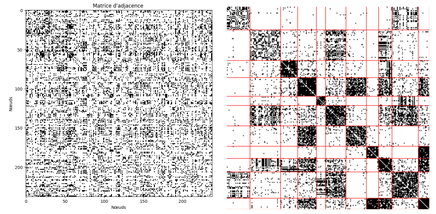



Figure 4. Adjacency matrix (left) and with uniform init (right)

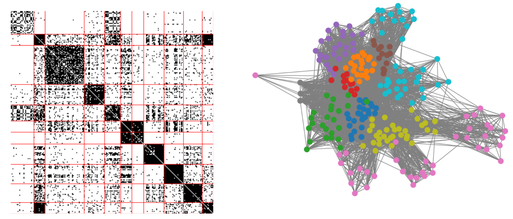


Figure 5. Adjacency matrix (left) and graph (right) with K-means init

## Quantification of the results on a real-world graph

| Init | Metrics | Sp-school | Polbooks |
|---|---|---|---|
| Random | Likelihood | -9153.10 | -1426.42 |
| | NMI | 0.695 | 0.090 |
| Sparse | Likelihood | **-9152.65** | -1310.66 |
| | NMI | 0.719 | **0.570** |
| K-means | Likelihood | -9207.5 | **-1306.67** |
| | NMI | **0.760** | 0.516 |

Table 1. Maximum likelihood and NMI with 20 initializations on real-world data.

## Takeaway

The EM algorithm we used performs well on SBM graphs, which is expected since it was specifically designed for such graphs. It also yields good results on real-world graphs.

However, the random uniform initialization appears to be limited because the optimization paths tend to be quite similar. In practice, these initializations do not allow the algorithm to explore sufficiently diverse clustering to reach higher-quality local maxima.

Therefore, when applying the EM algorithm to real graphs, it is preferable to test different initializations that follow various generative models, like Sparse or K-means. This approach allows for exploring more diverse starting points, ultimately leading to better results.