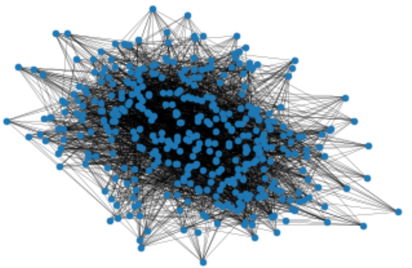
EXACT RECOVERY AND BREGMAN HARD CLUSTERING OF NODE-ATTRIBUTED STOCHASTIC BLOCK MODEL

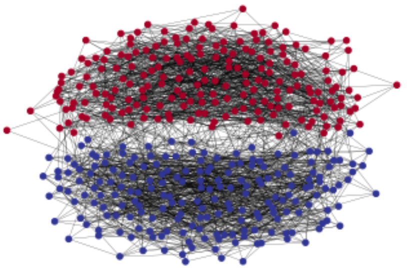
Maximilien Dreveton   Felipe S. Fernandes   Daniel R. Figueiredo

EPFL, Switzerland, & Federal University of Rio de Janeiro, Brazil

NeurIPS – 2023

Graph clustering with node attributes





Setup

- Observed data: Interactions between node pairs (network) and node attributes (features).
- Hidden data: Nodes are divided into clusters.

Main focus

- Theoretical: how much information is brought by the network and by the attributes?
- Practical: derive an algorithm that learns both from the *network* and from the *attributes*.
  - network: often sparse and possibly weighted;
  - attributes: a vector with discrete or continuous entries (or a mix of both).

NODE-ATTRIBUTED SBMS

- n nodes are divided into K latent blocks. We denote by $z \in [K]^n$ the vector of the block (cluster) memberships, and we suppose that:
  - $z_1, \cdots, z_n$ are iid such that $\mathbb{P}(z_i = a) = \pi_a$.
- Pairwise interactions $(X_{ij})_{1 \le i,j \le n}$ and node attributes $(Y_i)_{1 \le i \le n}$ are independent conditionally on the blocks:
  - $f_{ab}(X_{ij})$: probability of observing an interaction X_{ij} between a node i in block a and a node j in block b;
  - $h_a(Y_i)$: probability of observing an attribute Y_i for a node i in a block a.

Conditional distribution of the data (X, Y) given block memberships z:

$$\mathbb{P}(X, Y \mid z) = \prod_{1 \le i < j \le n} f_{z_i z_j}(X_{ij}) \prod_{i=1}^{n} h_{z_i}(Y_i).$$

How hard is it to recover z based on the observation of X and Y?

EXACT RECOVERY OF BLOCK MEMBERSHIPS

Denote by $\mathrm{D}_t(f\|g) = \frac{1}{t-1}\log\int f^t g^{1-t}$ the *Rényi divergence* of order t between two pdf f and g.
A key information-theoretic divergence is

$$I \;=\; \min_{\substack{a,b\in[K]\\ a\neq b}} \mathrm{CH}(a,b). \tag{1.1}$$

where $\mathrm{CH}(a,b) \;=\; \sup_{t\in(0,1)}(1-t)\left[\;\underbrace{\sum_{c=1}^{K}\pi_c\,\mathrm{D}_t\left(f_{bc}\|f_{ac}\right)}_{\text{information from the network}} \;+\; \underbrace{\frac{1}{n}\,\mathrm{D}_t\left(h_b\|h_a\right)}_{\text{information from the attributes}}\;\right].$

EXACT RECOVERY OF BLOCK MEMBERSHIPS

Denote by $\mathrm{D}_t(f \| g) = \frac{1}{t-1} \log \int f^t g^{1-t}$ the *Rényi divergence* of order t between two pdf f and g.
A key information-theoretic divergence is

$$I = \min_{\substack{a,b \in [K] \\ a \neq b}} \mathrm{CH}(a,b). \tag{1.1}$$

where $\mathrm{CH}(a,b) = \sup_{t \in (0,1)} (1-t) \left[\underbrace{\sum_{c=1}^{K} \pi_c \mathrm{D}_t \left(f_{bc} \| f_{ac} \right)}_{\text{information from the network}} + \underbrace{\frac{1}{n} \mathrm{D}_t \left(h_b \| h_a \right)}_{\text{information from the attributes}} \right].$

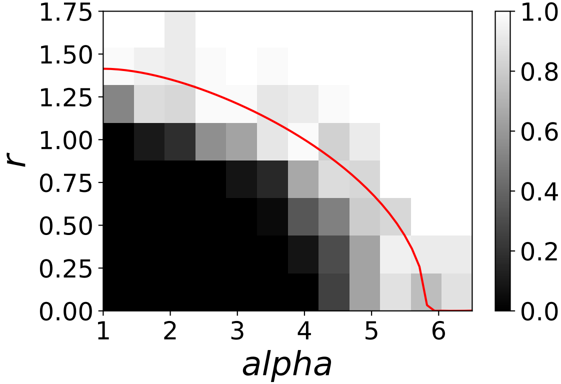
Theorem 1

Suppose $K = \Theta(1)$ and $\pi_a > 0$ for all $a \in [K]$. Denote by a^*, b^* the two hardest blocks to distinguish, that is $\mathrm{CH}(a^*, b^*) = I$. Suppose for all $t \in (0,1)$, $\lim_{n \to \infty} \frac{n}{\log n} \mathrm{CH}_t(a^*, b^*)$ exists and is strictly concave. Then the following holds:

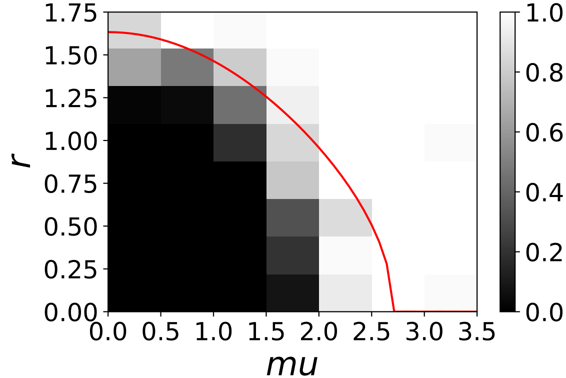
(i) *exact recovery of z is information-theoretically impossible if $\lim_{n \to \infty} \frac{n}{\log n} I < 1$;*

(ii) *exact recovery of z is information-theoretically possible if $\lim_{n \to \infty} \frac{n}{\log n} I > 1$.*

Numerical experiments



(a) Binary weights with Gaussian attributes



(b) zero-inflated Gaussian weights with Gaussian attributes.

Figure. Phase transition of exact recovery. Each pixel represents the empirical probability that Algorithm 1 succeeds at exactly recovering the clusters (over 50 runs), and the red curve shows the theoretical threshold.
(a) $n = 500$, $K = 2$, $f_{\rm in} = \mathrm{Ber}(\alpha n^{-1} \log n)$, $f_{\rm out} = \mathrm{Ber}(n^{-1} \log n)$. The attributes are 2d-spherical Gaussian with radius $(\pm r\sqrt{\log n}, 0)$ and identity covariance matrix.
(b) $n = 600$, $K = 3$, $f_{\rm in} = (1-\rho)\delta_0 + \rho\,\mathrm{Nor}(\mu, 1)$, $f_{\rm out} = (1-\rho)\delta_0 + \rho\,\mathrm{Nor}(0, 1)$ with $\rho = 5n^{-1} \log n$. The attributes are 2d-spherical Gaussian whose means are the vertices of a regular polygon on the circle of radius $r\sqrt{\log n}$.

Conclusion

In this presentation

Theoretical threshold for exact recovery of the community structure combines both the network and attribute information.

In the paper & poster

Algorithm that clusters sparse networks with weighted interactions and with node-attributes.

- We suppose the attributes are sampled from an *exponential family*;
- We suppose the network interactions are sampled from *zero-inflated exponential families*;
- We use the relationship between exponential families and *Bregman divergences* to derive an iterative algorithm based on *profile-likelihood maximisation*.