Stochastic Block Model

Q. Duchemin

Université Paris Est Marne La Vallée

April 2020

Definitions of the SBM and the SSBM

Let $n, k \in \mathbb{N}^*$, $\mathbf{p} = (p_1, \dots, p_k)$ a probability vector and $W \in S_k([0,1])$.

Definition : SBM

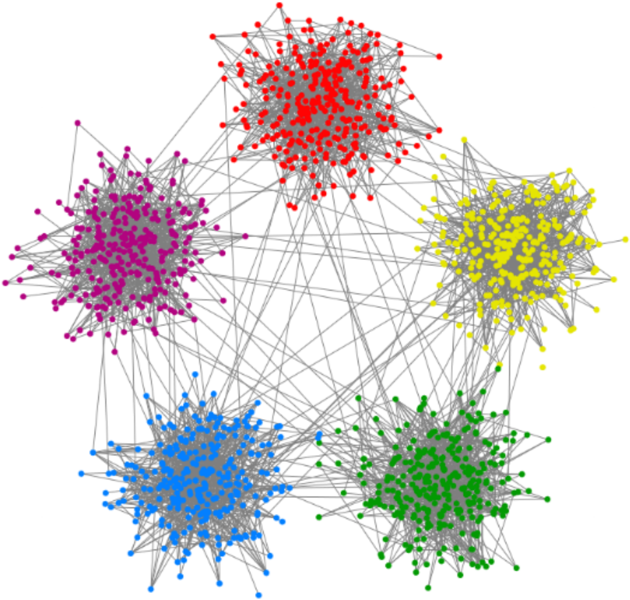
(X, G) is drawn under $SBM(n, p, W)$ if :

- $X_u \sim \mathbf{p}$, $\forall u \in [n]$.
- $G_{u,v} \sim \mathcal{B}(W(X_u, X_v))$, , $v \in [n]$, $u \neq v$.

Definition of the **Symmetric** SBM : SSBM

(X, G) is drawn under $\mathbf{S}SBM(n, k, q_{in}, q_{out})$ if :

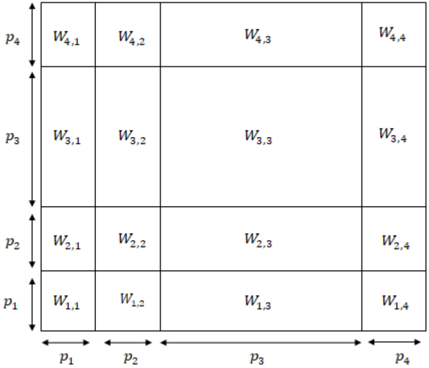
- $W(i,i) = q_{in}$ and $W(i,j) = q_{out}$, $\forall i, j \in [k]$ with $i \neq j$.
- $p = (\frac{1}{k}, \dots, \frac{1}{k})$.



SBM as a graphon model

Node v is in community i if $\sum_{l=1}^{i-1} p_l \leq U_v \leq \sum_{l=1}^{i} p_l$, $U_v \sim \mathcal{U}([0,1])$.

$u \longleftrightarrow v$ w.p. $g(U_u, U_v)$ where $g : [0,1]^2 \to [0,1]$ is defined below.



Recovery requirements

Definition

An algorithm recovers clusters in the SBM with accuracy α if it outputs clusters which agree with the true clusters on a fraction α of the vertices with probability that tends to 1 as n tends to infinity.

Agreement

Let $x, y \in [k]^n$.

$$A(x,y) = \max_{\pi \in \mathfrak{S}_k} \frac{1}{n} \sum_{u=1}^{n} \mathbb{1}_{x_u = \pi(y_u)}.$$

Exact recovery is solved if one can achieve an accuracy of $\alpha = 1$.

Exact recovery	$\mathbb{P}(A(X,\hat{X}) = 1) = 1 - o(1).$
Almost exact recovery	$\mathbb{P}(A(X,\hat{X}) = 1 - o(1)) = 1 - o(1).$

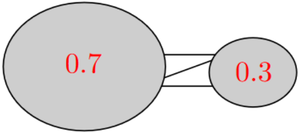
Weak recovery

What is the smallest non-trivial value of α ?

In the SSBM

$$\alpha = \frac{1}{k} + \epsilon, \quad \epsilon > 0.$$

In the asymmetrical SBM



For a random guess, the agreement is : $0.7^2 + 0.3^2 = 0.58$.
All in community 1 gives an agreement of 0.7.

Weak recovery

Weak recovery.

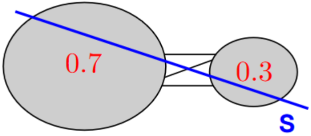
WR requires us to separate at least two communities.

Weak recovery is solvable in $SBM(n, \mathbf{p}, W)$ if :

$\exists \epsilon > 0, i, j \in [k]$, and an algo returning a partition (S, S^c) of $[n]$ s.t.

$$\mathbb{P}\left(\frac{|\Omega_i \cap S|}{|\Omega_i|} - \frac{|\Omega_j \cap S|}{|\Omega_j|} \geq \epsilon\right) = 1 - o(1),$$

where $|\Omega_i| = \{u \in [n] \ : \ X_u = i\}$.



Relevant regimes

Consider the SSBM with $b = 0$ (i.e., no edges between clusters).

Exact recovery when $b = 0 \Leftrightarrow$ Connectivity

Theorem (Erdos-Rényi '60)

Connectivity in $G(n, a\log(n)/n)$ iff $a > 1$.

Weak recovery when $b = 0 \Leftrightarrow$ Presence of a giant component

Theorem (Erdos-Rényi '60)

Presence of a Giant component in $G(n, a/n)$ iff $a > 1$.

Relevant regimes

Erdos Renyi	
Connectivity w.h.p in $G(n, c\log(n)/n)$	$\Leftrightarrow c > 1$
Giant Component in $G(n, c/n)$	

$\mathbf{SSBM(n, k, q_{in}, q_{out})}$	
Connectivity w.h.p in $SSBM(n, k, a\log(n)/n, b\log(n)/n)$	$\Leftrightarrow \frac{a+(k-1)b}{k} > 1$
Giant Component in $SSBM(n, k, a/n, b/n)$	

$\mathbf{SBM(n, p, W)}$	
Connectivity w.h.p in $SBM(n, p, Q\log(n)/n)$	$\Leftrightarrow \min_{i\in[k]} \|\,((diag(p)Q)_i\,\|_1 > 1$

To keep in mind for the talk

Connectivity matrix W

We will consider connectivity matrix of the form

$$W := \alpha_n Q,$$

where $Q \in [0,1]^{k \times k}$ is a symmetric $k \times k$ matrix and α_n is a sparsity parameter. Typically

$$\alpha_n = \frac{1}{n} \text{ or } \alpha_n = \frac{\log(n)}{n}.$$

Implications for recovery requirements

exact recovery $\implies$ almost exact recovery $\implies$ weak recovery $\implies$ distinguishability.

1. Exact Recovery

2. Weak recovery

3. Partial recovery

Section 2

Exact Recovery

Exact Recovery

Theorem [MNS14]

Exact recovery in $SSBM\left(n, 2, a\frac{\log(n)}{n}, b\frac{\log(n)}{n}\right)$ is

- solvable and efficiently if $|\sqrt{a} - \sqrt{b}| > \sqrt{2}$.
- unsolvable if $|\sqrt{a} - \sqrt{b}| < \sqrt{2}$.

Remarks.

- $|\sqrt{a} - \sqrt{b}| > \sqrt{2} \Leftrightarrow \frac{a+b}{2} > 1 + \sqrt{ab}$: connectivity is necessary but not sufficient.
- The case of equality is also solved in [MNS14].
- In [MNS14], the $SSBM(n, 2, a_n, b_n)$ is tackled when $a_n \sim b_n$. Exact recovery is solvable if and only if $((\sqrt{a_n} - \sqrt{b_n})^2 - 1)\log(n) + \log\log(n/2) = \omega(1)$.

Towards the general SBM : Community profile

The expected number of neighbors that a node in community i has in community j is

$$np_j W_{i,j}.$$

If $W = \log(n)Q/n$, then $n diag(p)W = \log(n) diag(p)Q$.

Definition

Let us consider a graph sampled from $SBM(n, p, \log(n)Q/n)$.
The degree profile of a node $u \in [n]$ belonging to community $i \in [k]$ is defined by

$$d(u) = \log(n) diag(p)Q_i \in \mathbb{R}^k.$$

The general SBM : Exact Recovery

Exact recovery in the phase transition regime $W = Q\log(n)/n$. [AS15]

Exact recovery in $SBM(n, p, \log(n)Q/n)$ is solvable and efficiently so if

$$I_+(p, Q) := \min_{1\leq i<j\leq k} D_+\left((diag(p)Q)_i || (diag(p)Q)_j\right) > 1,$$

and is not solvable if $I_+(p, Q) < 1$, where D_+ is defined by

$$D_+(\mu, \nu) := \max_{t\in[0,1]} \underbrace{\sum_x \nu(x) f_t\left(\mu(x)/\nu(x)\right)}_{:=D_t(\mu,\nu)}, \quad f_t(y) := 1 - t + ty - y^t.$$

Remarks.

- $\forall t \in [0, 1], \quad D_t$ is a f-divergence.
- In $SSBM(n, k, a\log(n)/n, b\log(n)/n)$, the CH-divergence is maximized at $t = 1/2$ and is reduced to the Hellinger divergence between any two columns of Q. The theorem's inequality becomes $\frac{1}{k}(\sqrt{a} - \sqrt{b})^2 > 1$.

Information-Computational gap and open problems.

No Information Computational gap in the standard SBM.

Open problems

If $k = o(\log(n))$, and the communities remain reasonably balanced, most of the developed techniques extend for exact recovery. However new phenomena seem to take place beyond this regime, with again gaps between information and computational thresholds. In [CX14], some of this is captured by looking at coarse regimes of the parameters.

Exact Recovery : Achievability

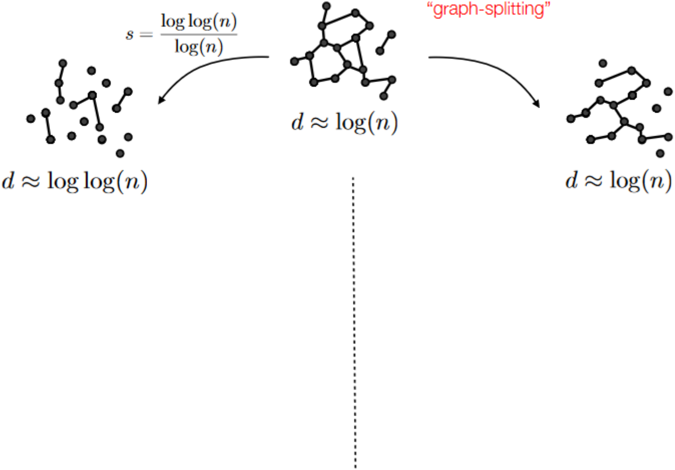
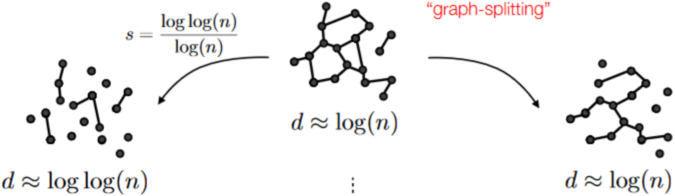


Figure created by A.Bandeira and C.Sandon.



$s = \frac{\log\log(n)}{\log(n)}$

"graph-splitting"

$d \approx \log\log(n)$

$d \approx \log(n)$

$d \approx \log(n)$

Solve here "almost exact" recovery
- all vertices but a vanishing fraction -
using a "smart enough" algorithm

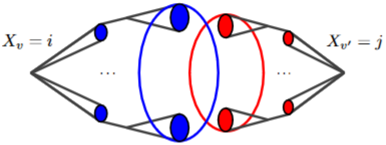
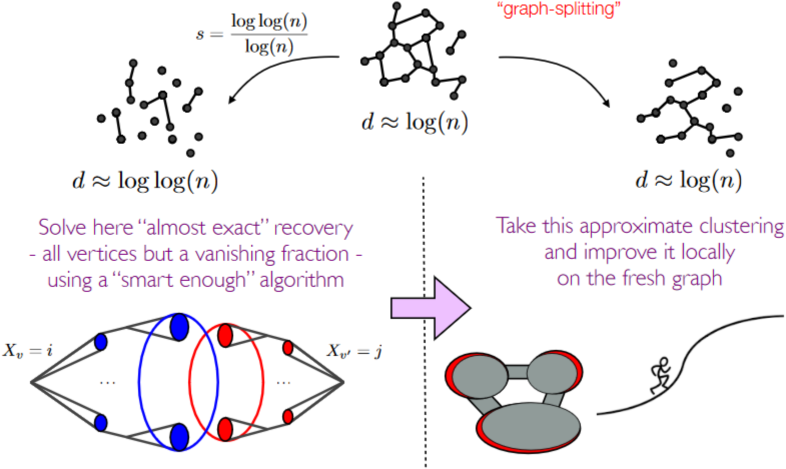


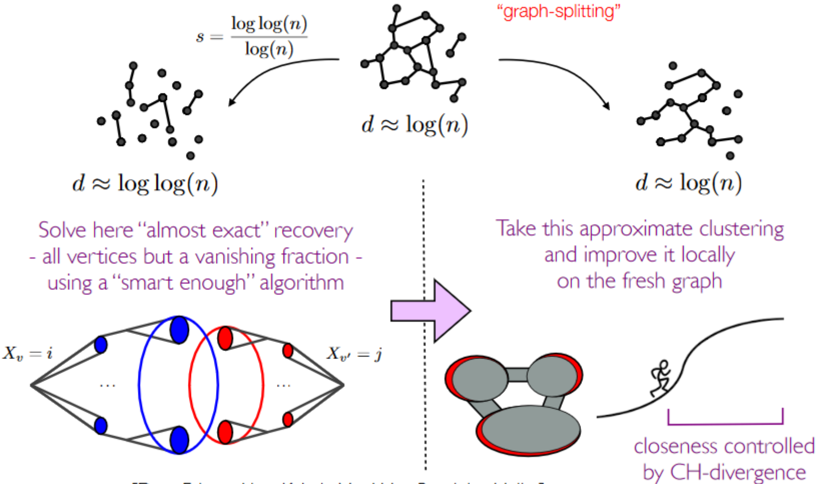
Figure created by A.Bandeira and C.Sandon.



Solve here "almost exact" recovery
- all vertices but a vanishing fraction -
using a "smart enough" algorithm

Take this approximate clustering
and improve it locally
on the fresh graph

Figure created by A.Bandeira and C.Sandon.



"graph-splitting"

$s = \frac{\log\log(n)}{\log(n)}$

$d \approx \log(n)$

$d \approx \log\log(n)$

$d \approx \log(n)$

Solve here "almost exact" recovery - all vertices but a vanishing fraction - using a "smart enough" algorithm

Take this approximate clustering and improve it locally on the fresh graph

closeness controlled by CH-divergence

[Dyer-Frieze, Alon-Kahale, Vu, Abbe-Bandeira-Hall,...]

Figure created by A.Bandeira and C.Sandon.

Le Cam's inequality

$$\left\|\mathrm{Bin}\left(np_i, \frac{\log(n)}{n} Q_{i,j}\right) - \mathcal{P}(p_i Q_{i,j} \log(n))\right\|_{TV} \leq 2 p_i Q_{i,j}^2 \frac{\log(n)^2}{n}.$$

Idea : **Test between** k- **multivariate Poisson distributions** of different means $\log(n)\theta_1, \ldots, \log(n)\theta_k \in \mathbb{Z}_+^k$, where $\theta_j = diag(p) Q_j$. The probability of error of this hypothesis test is controlled by $(*)$.

Theorem

$$(*) := \sum_{x \in \mathbb{Z}_+^k} \min\left(\mathcal{P}_{\log(n)\theta_1}(x), \mathcal{P}_{\log(n)\theta_2}(x)\right) = n^{-D_+(\theta_1, \theta_2) + o(1)},$$

where for any $\lambda \in \mathbb{Z}_+^k$ and any degree profile $d \in \mathbb{Z}_+^k$,

$$\mathcal{P}_\lambda(d) = \prod_{i \in [k]} \frac{\lambda_i^{d_i}}{d_i!} e^{-\lambda_i}.$$

Summary : Exact Recovery Achievability

Method

1. Graph splitting : $G \to (G_1$ and $G_2)$.
2. Solve almost exact recovery with degrees that grow sub-logarithmically.
3. A robust version of the genie-aided hypothesis test can be run to reclassify each node successfully when $I_+(p, Q) > 1$.

From almost exact recovery to exact recovery

If almost exact recovery is solvable in $SBM(n, p, \omega(1)Q/n)$, then exact recovery is solvable in $SBM(n, p, \log n/nQ)$ if $I_+(p, Q) > 1$.

Almost exact recovery efficiently solvable

Almost exact recovery is efficiently solvable in $SBM(n, p, \omega(1)Q/n)$.

Section 3

Weak recovery

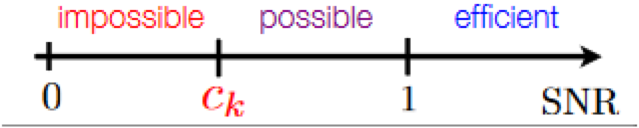
Decelle et al. conjecture

Conjecture (based on deep but non-rigorous statistical physics arguments)

Let (X, G) be drawn from $SSBM(n, k, a/n, b/n)$.

Let $SNR = \dfrac{(a-b)^2}{k(a+(k-1)b)}$.

- For any $k \geq 2$, it is possible to solve weak recovery efficiently if and only if $\mathbf{SNR} > \mathbf{1}$ (the Kesten-Stigum (KS) threshold).
- If $k \geq 3$, it is possible to solve weak recovery information-theoretically (i.e.,not necessarily in polynomial time in n) for some SNR strictly below 1.



New algorithmic challenge : Standard clustering algorithms (Laplacians, SDPs) fail to achieve the KS threshold !

Weak Recovery : An overview.

- **Weak recovery is closed for $k = 2$ in SSBM.**
  1. WR solvable efficiently in $SSBM(n, 2, a/n, b/n)$ when $a, b = O(1)$ if and only if $(a - b)^2 > 2(a + b)$.
  2. WR solvable efficiently in $SSBM(n, 2, a_n/n, b_n/n)$ when $a_n, b_n = \omega(1)$ and $(a_n - b_n)^2/(2(a_n + b_n)) \to \lambda$ if and only if $\lambda > 1$.
  3. For $SNR \leq 1$, distinguishability is not solvable.
- The achievability parts of previous conjecture for $k \geq 3$ are resolved in [AS15], with an extended result applying to the general SBM.
- In [Abb18], a non-efficient algorithm crosses the KS threshold for $k \geq 4$.

Achieving the threshold for WR

What algorithm is used ?
A linearized version of the Belief Propagation algorithm (LBPA).
In the SSBM, one can show that the LBPA is equivalent to a **power iteration method** on a linear operator involving the *non-backtracking walk matrix* of the graph.

Definition (Non-backtracking walk matrix)

Given a graph (V, E), the graph's non-backtracking walk matrix B is a matrix of dimension $|\vec{E}_2| \times |\vec{E}_2|$, where $\vec{E}_2$ is the set of directed edges on E (with $|\vec{E}_2| = 2|E|$), such that for two directed edges $e = (i, j)$, $f = (k, l)$,

$$B_{e,f} = \mathbb{1}\{j = k, i \neq l\}.$$

Achieving the threshold for WR

Definition ($r-$non-backtracking walk matrix)

Let (V, E) be a graph and $\vec{E}_r$ be the set of directed paths of length $r-1$ obtained on E.

The graph's $r-$non-backtracking walk matrix $B^{(r)}$ is a matrix of dimension $|\vec{E}_r| \times |\vec{E}_r|$, such that for $e = (e_1, \dots, e_{r-1}) \in \vec{E}_r$ and $f = (f_1, \dots, f_{r-1}) \in \vec{E}_r$,

$$B^{(r)}_{e,f} = \prod_{i=1}^{r-2} \mathbb{1}\{(e_{i+1})_2 = (f_i)_1\} \times \mathbb{1}\{(e_1)_1 \neq (f_{r-1})_2\}.$$

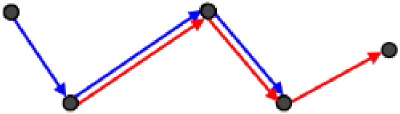


Figure – Two paths of length 3 that contribute to an entry of 1 in $B^{(4)}$.

Achieving the threshold for WR

Why the adjacency matrix is a bad idea ?
Powers count walks from a vertex to another, and these get amplified around high-degree vertices.

$$\Longrightarrow$$

Large eigenvalues with eigenvectors **localized around high-degree vertices**.

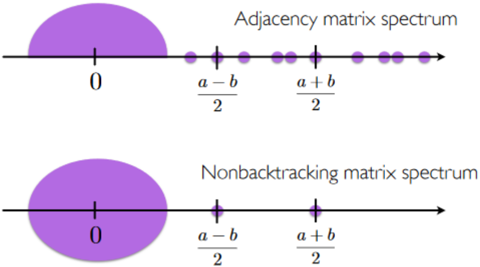


Figure – Illustration of the spectrum of the adjacency and non-backtracking matrices for the SBM with two symmetric communities above the KS threshold.

Algorithm 1 Algorithm to solve weak recovery.

Data : Adjacency matrix of a graph $G = (V, E)$.

```
Build the $r-$non-backtracking matrix of the graph $B^{(r)}$.
Extract the eigenvector $y$ corresponding to the second largest eigenvalue of $B^{(r)}$.
For each $v$, set $y'_v = \sum_{v':(v',v)\in E(G)} y_{v,v'}$ and return $(\{v : y'_v > \tau/\sqrt{n}\}, \{v : y'_v \leq \tau/\sqrt{n}\})$
```

Extension to the general SBM

Definition

Let $\lambda_1, \lambda_2, \dots, \lambda_k$ be the distinct eigenvalues of $diag(p)Q$ in order of nonincreasing magnitude. We define the signal-to-noise ratio of $SBM(n, p, Q/n)$ by

$$SNR := \frac{\lambda_2^2}{\lambda_1}.$$

Theorem

If $SNR > 1$, then the power iteration algorithm on the $r-$non-backtracking matrix solves efficiently weak recovery in $SBM(n, p, Q/n)$.

Section 4

Partial recovery

Finding the right balance between *SNR* and classification error.

1. A tight result close to KS threshold
   In [HS15], an exponential bound for partial recovery in general sparse SBM : $SBM(n, k, Q/n)$. Their results are optimal in the vicinity of the weak recovery threshold.

2. A better result with more signal
   In [GV19] a relaxed SDP is solved to recover communities. They show that their algorithm outputs a partition of the nodes which leads to a misclassification error that decays exponentially fast with respect to their SNR. They show that their result improve the partial recovery bound from [HS15] in a setting with slightly more signal.

Let us detail the method of Giraud and Verzelen in [GV19].

Peng and Wei in [PW07] showed that any partition G of $[n]$ can be uniquely represented by $B^* \in \mathbb{R}^{n\times n}$ defined by

$$\forall u, v \in [n], \quad B^*_{u,v} = \begin{cases} \dfrac{1}{m_i} & \text{if } u \text{ and } v \text{ belong to the same community } i \in [k] \\ 0 & \text{otherwise.} \end{cases}$$

The set of such matrices B^* is

$$\mathcal{S} = \{B \in \mathbb{R}^{n\times n} : \text{ symmetric}, \; B^2 = B, \; \mathrm{Tr}(B) = k, \; B\mathbf{1} = \mathbf{1}, \; B \geq 0\}.$$

$$B^* \to \arg\min \sum_{i=1}^{k} \sum_{u \in G_i} \left\| A_{:,u} - \frac{1}{|G_i|} \sum_{v \in G_i} A_{:,v} \right\|^2$$

$$\Leftrightarrow$$

$$B^* = \arg\max_{B \in \mathcal{S}} \langle AA^\top, B \rangle.$$

$$\mathcal{S} = \{B \in \mathbb{R}^{n \times n} : \text{ symmetric}, \; B^2 = B, \; \operatorname{Tr}(B) = k, \; B\mathbf{1} = \mathbf{1}, \; B \geq 0\}$$

$$B^* \to \arg\min \sum_{i=1}^{K} \sum_{u \in G_i} \left\| A_{:,u} - \frac{1}{|G_i|} \sum_{v \in G_i} A_{:,v} \right\|^2$$

$$\Leftrightarrow$$

$$B^* = \underset{B \in \mathcal{S}}{\arg\max} \, \langle AA^\top, B \rangle. \tag{1}$$

In [GV19], Verzelen and Giraud propose the following relaxation of problem (1)

$$\hat{B} \in \underset{B \in \mathcal{C}_\beta}{\arg\max} \, \langle AA^\top, B \rangle,$$

where for $k/n \leq \beta \leq 1$,

$$\mathcal{C}_\beta := \{B \in \mathbb{R}^{n \times n} : \text{ symmetric}, \; \operatorname{Tr}(B) = k, \; B\mathbf{1} = \mathbf{1}, 0 \leq B \leq \beta\}.$$

A final rounding step on the rows of $\hat{B}$ is necessary to conclude.

Theorem [GV19]

Let $W = \alpha_n Q$. We define the signal-to-noise ratio $s^2 = \Delta^2/(\alpha_n \|Q\|_\infty)$, where $\Delta^2 = \min\limits_{k \neq j} \Delta^2_{k,j}$ with

- $\forall l \in [k]$, m_l is the size of community l.
- $\Delta^2_{k,j} = \sum_l m_l (W_{k,l} - W_{j,l})^2 = \alpha_n^2 \sum_l m_l (Q_{k,l} - Q_{j,l})^2$.

Then, there exist positive constants c, c', c'', such that if

$$s^2 \geq c'' n / (\min_l \ m_l),$$

it holds with probability at least $1 - c/n^2$,

$$A(\hat{X}, X) \geq 1 - e^{-c' s^2}.$$

Conclusion

Open problems

Weak recovery : IT gap.
Partial recovery : Fundamental tradeoff between the SNR and the agreement.

Beyond the SBM

Dynamic SBMs

- Models where the size of the graph is fixed and communities can change with time (C.Mathias, L.Longepierre).
- Model with a Markovian assignment of the communities.

When SBMs meet Bandits
[C.Giraud and M.Lerasle] : Algorithm that aims at finding the largest number of pairs with the same labels.