Improved clustering algorithms for the Bipartite Stochastic Block Model

Talk by S. Sigalla (CREST, ENSAE),
joint work with M. Ndaoud (USC Math Department) and A.B. Tsybakov (CREST, ENSAE).

Meeting in Mathematical Statistics – CIRM
December 18th 2020
ArXiv preprint, arXiv:1911.07987

Introduction

The Stochastic Block Model

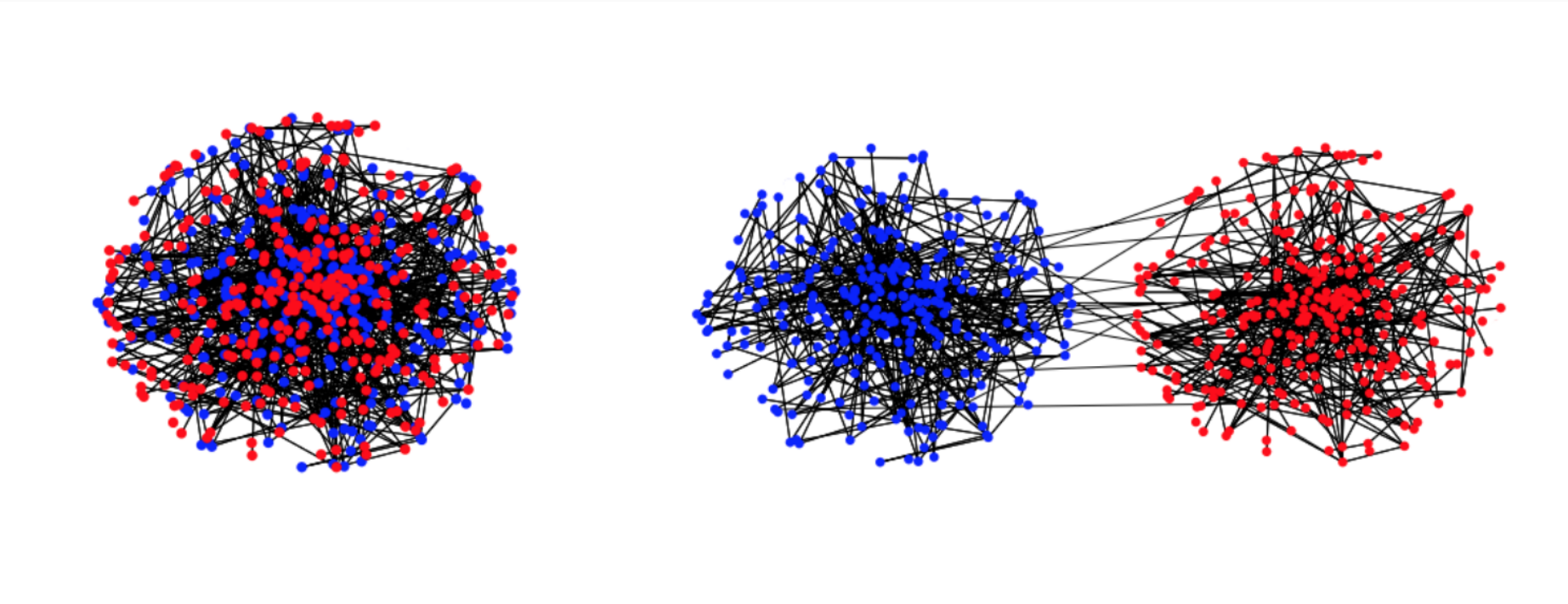
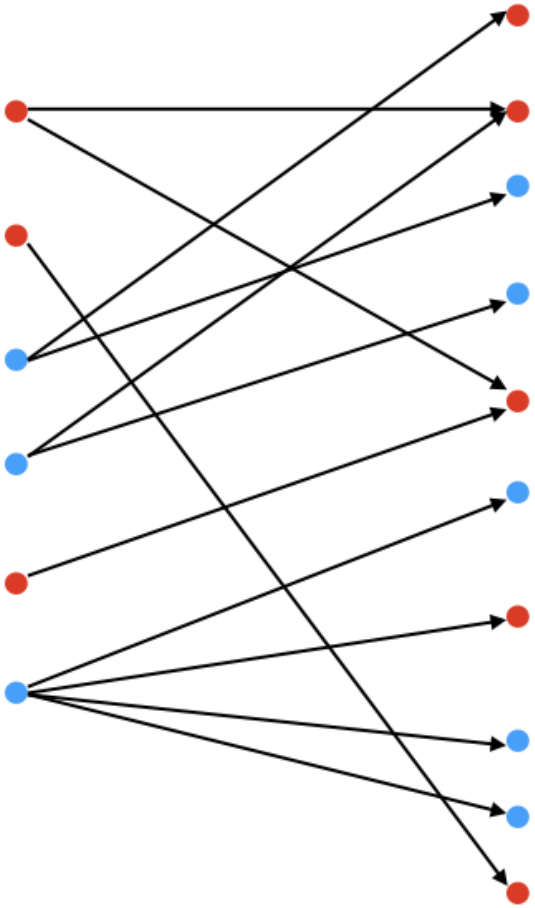
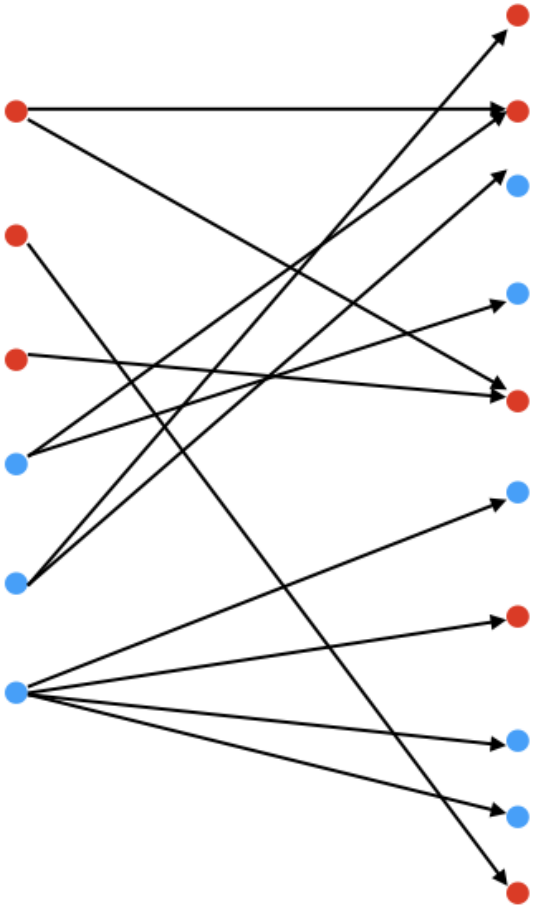


Figure 1: A graph generated from the stochastic block model with 600 nodes and 2 communities, scrambled on the left and clustered on the right. Nodes in this graph connect with probability $p = 6/600$ within communities and $q = 0.1/600$ across communities. Courtesy to [Abbe *et al.*, 2015].

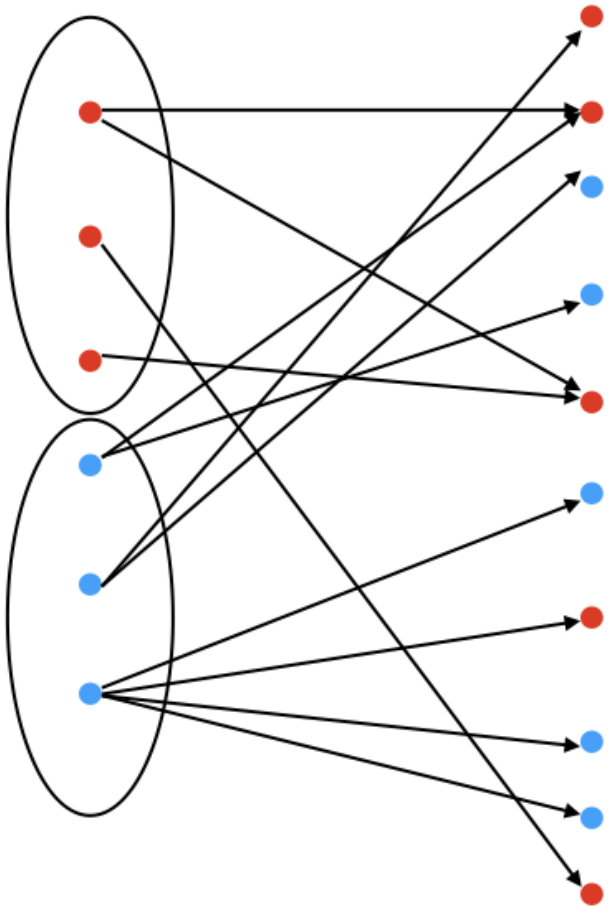
The Bipartite Stochastic Block Model (BSBM)



The Bipartite SBM (cont.)

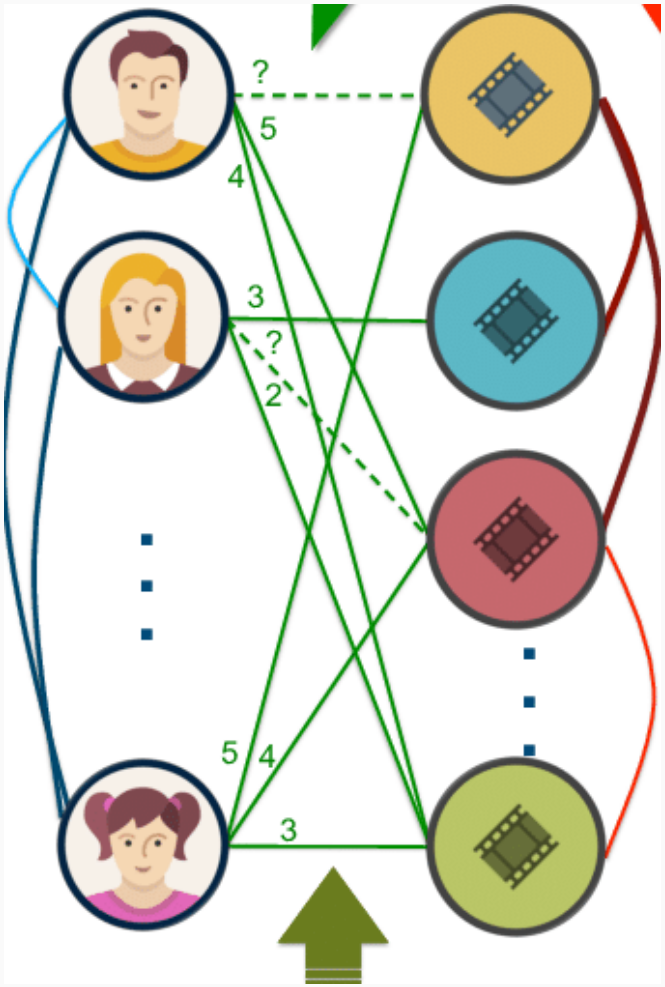


The Bipartite SBM (cont.)



Why this model?

- In-network interactions are not **accessible** or not **informative**.
- **Huge** amount of out of network interactions.
- Example: recommendation systems.



Statement of the problem

Statement of the problem

- Consider **two** sets of vertices V_1 and V_2 of respective sizes $n_1 = n_{1+} + n_{1-}$ and $n_2 = n_{2+} + n_{2-}$. We denote by $\sigma(u) \in \{-1, 1\}$ the label corresponding to vertex u.

- Let A denote the **biadjacency matrix**. We say that matrix A is drawn according to a *BSBM* model if:
  1. $A_{ij} \sim Ber(\delta p)$ if $\sigma(i) = \sigma(j)$,
  2. $A_{ij} \sim Ber((2-\delta)p)$ if $\sigma(i) \neq \sigma(j)$,
  3. $(A_{ij})_{i,j}$ are independent,

  where $0 < \delta < 2, 0 < p < 1/2$.

- Define $\gamma_1 := |n_{1+} - n_{1-}|/n_1$ (resp. $\gamma_2 := |n_{2+} - n_{2-}|/n_2$) the **imbalance** of the set V_1 (respectively, V_2).
- Interesting case: $\mathbf{n_2 >> n_1}$.

Statement of the problem (cont.)

Denote by $\eta_1 \in \{\pm 1\}^{n_1}$ the vector of vertex labels in V_1.

An estimator $\hat{\eta} = \hat{\eta}(A)$ is a binary valued estimator:

$$\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_{n_1}), \quad \hat{\eta}_j \in \{-1, 1\}.$$

Hamming loss of an estimator $\hat{\eta} = \hat{\eta}(A)$ is

$$|\hat{\eta} - \eta_1| \triangleq \sum_{j=1}^{n_1} |\hat{\eta}_j - \eta_{1j}| = 2 \sum_{j=1}^{n_1} \mathbf{1}(\hat{\eta}_j \neq \eta_{1j}).$$

A more appropriate loss of an estimator $\hat{\eta}$ is

$$r(\hat{\eta}, \eta_1) := \min_{\nu \in \{-1,+1\}} |\hat{\eta} - \nu \eta_1| .$$

Types of recovery

Definition (weak recovery)
The estimator $\hat{\eta}$ achieves **weak recovery** *if there exists $\alpha \in (0,1)$ such that*

$$\lim_{n_1 \to \infty} \sup_{BSBM} \mathbb{P}\left(\frac{r(\eta_1, \hat{\eta})}{n_1} \geq \alpha \right) = 0,$$

where $\sup_{BSBM}$ denotes the maximum over all distributions of A drawn from $BSBM(\delta, n_{1+}, n_{1-}, n_{2+}, n_{2-}, p)$.

Weak recovery can be interpreted as the fact that $\hat{\eta}$ classifies the vertices **better than chance**.

Types of recovery (cont.)

Definition (almost full recovery)
The estimator $\hat{\eta}$ achieves **almost full recovery** *if for all $\alpha \in (0,1)$ we have*

$$\lim_{n_1 \to \infty} \sup_{BSBM} \mathbb{P}\left(\frac{r(\eta_1, \hat{\eta})}{n_1} \geq \alpha \right) = 0.$$

Almost full recovery means that $\hat{\eta}$ correctly classifies the vertices **on average**.

Types of recovery (cont.)

Definition (exact recovery)
The estimator $\hat{\eta}$ achieves **exact recovery** *if*

$$\lim_{n_1 \to \infty} \inf_{BSBM} \mathbb{P}\big(r(\eta_1, \hat{\eta}) = 0\big) = 1.$$

Exact recovery means that $\hat{\eta}$ correctly classifies **all the vertices**.

Statistical question

Sufficient conditions on p to achieve exact recovery?

$$\lim_{n_1 \to \infty} \inf_{\mathrm{BSBM}} \mathbf{P}\left(r(\hat{\eta}, \eta_1) = 0\right) = 1.$$

A spiked model

- The biadjacency matrix A can be written as

$$A = \mathbf{E}(A) + W$$

where A is observed and W is a centered random matrix.

- One can check that

$$\mathbf{E}(A) = p\mathbf{1}_{n_1}\mathbf{1}_{n_2}^\top + (\delta - 1)p\eta_1\eta_2^\top.$$

- the non-informative matrix $p\mathbf{1}_{n_1}\mathbf{1}_{n_2}^\top$ can be **eliminated by estimating** p by

$$\hat{p} = \frac{1}{n_1 n_2}\mathbf{1}_{n_1}^\top A\mathbf{1}_{n_2} \tag{1}$$

and then considering

$$\hat{A} = A - \hat{p}\mathbf{1}_{n_1}\mathbf{1}_{n_2}^\top = (\delta - 1)p\eta_1\eta_2^\top + \underbrace{W + (p - \hat{p})\mathbf{1}_{n_1}\mathbf{1}_{n_2}^\top}_{\text{noise}}$$

Previous de-biasing spectral methods

- Spectral methods are based on classical SVD on **the Gram matrix** $AA^\top$.
- Why $AA^\top$ rather than A? Because it reduces the dimension.
- Classical SVD: **bias** of order

$$\Sigma = \mathbf{E}\left(WW^\top\right)$$

  that **grows with** n_2.
- Strict improvement: **adaptive de-biasing procedure** [Royer, 2017] by considering

$$\frac{1}{n_1}AA^\top - \hat{\Sigma}$$

  where $\hat{\Sigma}$ is an estimator of the covariance matrix Σ.
- Σ and $\hat{\Sigma}$ are both **diagonal** matrices.

Previous de-biasing spectral methods (cont.)

- Idea: **diagonal deletion SVD** [Florescu & Perkins, 2016]: SVD applied to matrix

$$AA^\top - \mathrm{diag}(AA^\top).$$

- Sufficient condition to obtain exact recovery for diagonal deletion SVD when $n_2 > n_1$:

$$p = \Omega\left(\frac{\log n_1}{\sqrt{n_1 n_2}}\right).$$

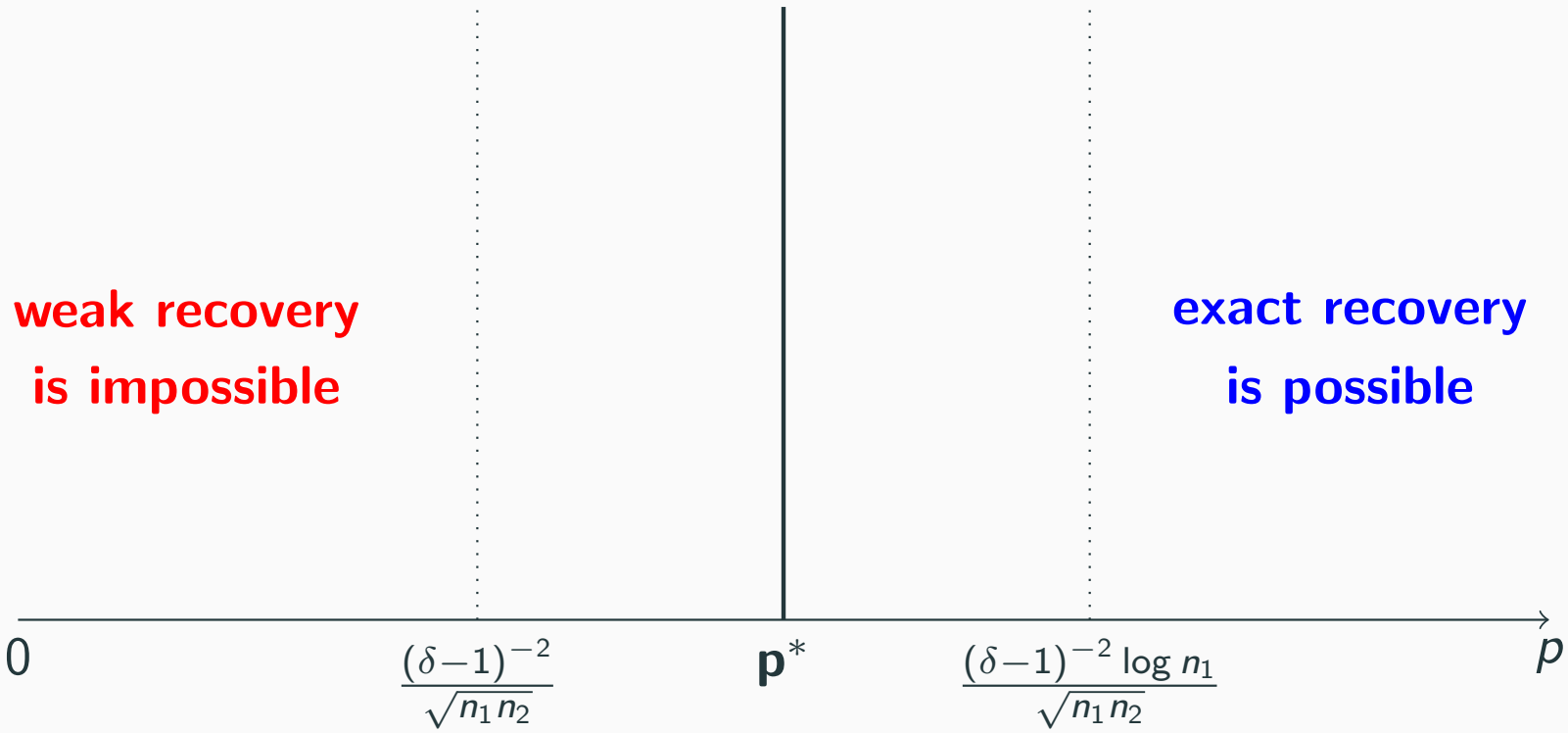
Related literature

Reference	Results	Conditions	Algorithm
[Feldman *et al.*, 2015]	Exact recovery	$\begin{cases} n_2 \geq n_1, \text{ known } p, \\ p \geq C(\delta - 1)^{-2} \frac{\log n_1}{\sqrt{n_1 n_2}} \end{cases}$	Subsampled iterations
[Florescu & Perkins, 2016]	Almost full recovery	$\begin{cases} n_2 \geq n_1 \log^4 n_1, \gamma_1 = \gamma_2 = 0 \\ p \geq C_\delta \frac{\log n_1}{\sqrt{n_1 n_2}} \end{cases}$	Diagonal deletion SVD
[Florescu & Perkins, 2016]	Weak recovery	$\begin{cases} n_2 \geq n_1, \gamma_1 = \gamma_2 = 0 \\ p > \frac{(\delta-1)^{-2}}{\sqrt{n_1 n_2}} \end{cases}$	SBM reduction

Table 1: Summary of the results of [Feldman *et al.*, 2015] and [Florescu & Perkins, 2016].

State of the art results

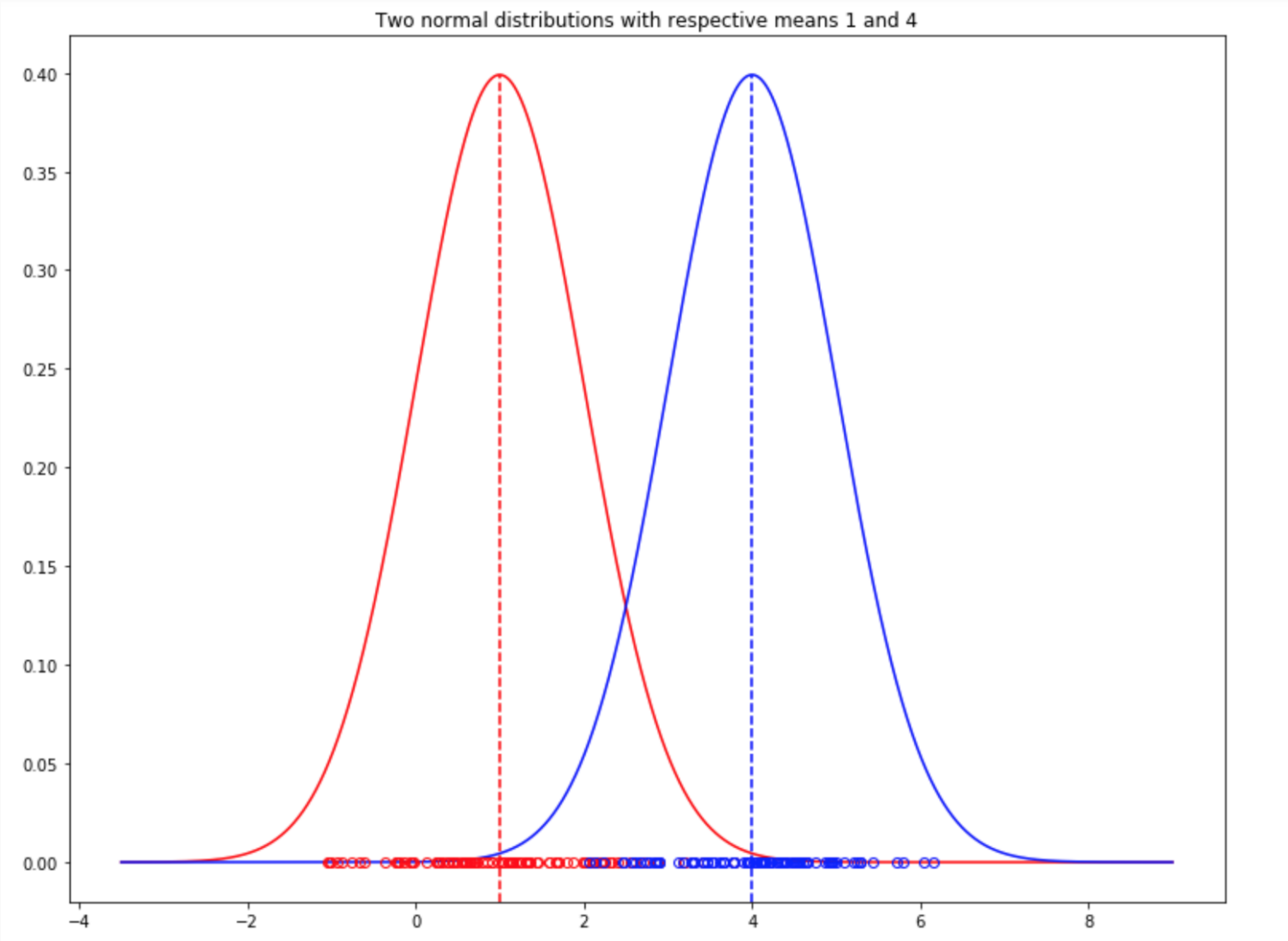
Assume that $n_2 \geq n_1 \log n_1$.



What about the gap ?

Heuristics on the optimal condition

The two components Gaussian Mixture Model: analogy with BSBM



The two components GMM: analogy with BSBM (cont.)

- [Ndaoud, 2018, Giraud & Verzelen, 2019, Lu & Zhou, 2016]
- **GMM**:
  $A^{GMM} = \eta_1 \theta^\top + \sigma W^{GMM}$, $\eta_1 \in \{-1, 1\}^{n_1}, \theta \in \mathbb{R}^{n_2}$,
  where $W_{ij}^{GMM} \sim \mathcal{N}(0, 1)$ i.i.d.
- **BSBM**:
  $A^{BSBM} = (\delta - 1)p\eta_1\eta_2^\top + W^{BSBM}$, $\eta_1 \in \{-1, 1\}^{n_1}, \eta_2 \in \{-1, 1\}^{n_2}$,
  where $W_{ij}^{BSBM} \sim Ber(\delta p)$ or $W_{ij}^{BSBM} \sim Ber((2 - \delta)p)$.
- **Moments matching** through:

$$\|\boldsymbol{\theta}\|^2 = (\delta - 1)^2 p^2 n_2$$

and

$$\sigma^2 \approx p.$$

A sharp phase transition

- In [Ndaoud, 2018], the **phase transition** happens around

$$\|\boldsymbol{\theta}\|^{*2} = \sigma^2(\log n_1)\left(1+\sqrt{1+2\frac{n_2}{n_1\log n_1}}\right).$$

- If $n_2 >> n_1 \log n_1$, this phase transition corresponds, in the **BSBM** model, to $(\delta-1)^2 p^2 n_2 = p\sqrt{(n_2 \log n_1)/n_1}$, i.e.

$$p^* = (\delta-1)^{-2}\sqrt{\frac{\log n_1}{n_1 n_2}}.$$

- Questions: is it possible to achieve **exact recovery** under the condition $p = \Omega\left((\delta-1)^2\sqrt{\frac{\log n_1}{n_1 n_2}}\right)$ in the BSBM model ? What about **computational issues** ?

Contributions

A new de-biasing technique

Define the linear operator $\mathbf{H} : \mathbf{R}^{n\times n} \to \mathbf{R}^{n\times n}$, such that

$$\forall M \in \mathbf{R}^{n\times n}, \quad \mathbf{H}(M) = M - \mathrm{diag}(M).$$

Recall that $\hat{A} = A - \hat{p}\mathbf{1}_{n_1}\mathbf{1}_{n_2}^\top$, then

$$\mathbf{H}(\hat{A}\hat{A}^\top) = \underbrace{(\delta-1)^2 p^2 n_2 \mathbf{H}(\eta_1\eta_1^\top)}_{\textit{signal}} + \underbrace{\cdots + \mathbf{H}(WW^\top) + (p-\hat{p})\mathbf{H}(Z)}_{\textit{noise}}.$$

Observe that

$$\|\mathbf{H}(WW^\top)\|_{op} \leq 2\left\|WW^\top - \mathbf{E}\left(WW^\top\right)\right\|_{op},$$

and

$$\|\mathbf{H}(\eta_1\eta_1^\top)\|_{op} = \left(1 - \frac{1}{n_1}\right)\|\eta_1\eta_1^\top\|_{op}.$$

$\Longrightarrow$ **If you don't like bias get rid of it!**

A better spectral estimator

- Consider $\mathbf{H}(\hat{A}\hat{A}^\top)$ rather than $\mathbf{H}(AA^\top)$.
- Define the following **spectral estimator**:

$$\eta_1^0 = \text{sign}(\hat{v}), \tag{2}$$

where $\hat{v}$ is the eigenvector corresponding to the **top** eigenvalue of $\mathbf{H}(\hat{A}\hat{A}^\top)$ and $\text{sign}(\hat{v})$ is the vector of signs of each entry of vector $\hat{v}$.

Almost full recovery through spectral clustering

Theorem 1 [Ndaoud, Sigalla and Tsybakov, 2019]

Let η_1^0 be the estimator given by (2). Let (C_{n_1}) be a sequence of positive numbers that tends to ∞ as $n_1 \to \infty$. If

$$\begin{cases} n_2 > n_1 \log n_1, \\ \gamma_1 \gamma_2 \leq 1/C_{n_1}, \\ p \geq C_{n_1}(\delta - 1)^{-2} \sqrt{\frac{\log n_1}{n_1 n_2}}. \end{cases}$$

Then η_1^0 achieves **almost full recovery** of η_1.

- **Almost full recovery**: $\frac{1}{n_1} r(\eta_1^0, \eta_1) \to 0$ as $n_1 \to \infty$.
- $\gamma_1 = |n_{1+} - n_{1-}|/n_1$ and $\gamma_2 = |n_{2+} - n_{2-}|/n_2$: **imbalance** parameters.
- Improves upon the result of [Florescu & Perkins, 2016].

Towards improved clustering conditions

A new iterative procedure for bipartite clustering

We introduce the **hollowed Lloyd's algorithm**:

- Inspiration: analogy with the **Gaussian Mixture Model (GMM)**, cf. [Ndaoud, 2018].
- We define a sequence of iterations $(\hat{\eta}^k)_{k\geq 0}$ such that

$$\forall k \geq 0, \quad \hat{\eta}^{k+1} = \operatorname{sign}\left(\mathbf{H}\left(\hat{A}\hat{A}^\top\right)\hat{\eta}^k\right). \tag{3}$$

  with the spectral estimator $\hat{\eta}^0 = \eta_1^0$ as initializer.
- Our final estimator: $\hat{\eta}^m$ with $m \geq 3\log n_1$.
- Difference of (3) from original **Lloyd's iterations**: we replace $\hat{A}\hat{A}^\top$ by $\mathbf{H}\left(\hat{\mathbf{A}}\hat{\mathbf{A}}^\top\right)$.

Improved clustering conditions

Theorem 2 [Ndaoud, Sigalla and Tsybakov, 2019]

Let $(\hat{\eta}^k)_{k\geq 0}$ be the recursion (3) initialized with the spectral estimator (2) for $\hat{p}$ given by (1). There exists a constant $C > 0$ such that if:

$$\begin{cases} n_2 > n_1 \log n_1, \\ \gamma_1\gamma_2 \leq 1/480, \\ p \geq C(\delta-1)^{-2}\sqrt{\frac{\log n_1}{n_1 n_2}}, \end{cases}$$

then, the estimator $\hat{\eta}^m$ with $m \geq 3\log n_1$ achieves **exact recovery** of η_1.

- $\gamma_1 = |n_{1+} - n_{1-}|/n_1$ and $\gamma_2 = |n_{2+} - n_{2-}|/n_2$: **imbalance** parameters.
- **Improves** upon the result of [Feldman *et al.*, 2015].
- **Conjecture**: in the setting $n_2 \geq n_1 \log n_1$, the condition $p \geq C(\delta-1)^{-2}\sqrt{\frac{\log n_1}{n_1 n_2}}$ **cannot be improved**.

Impossibility for a supervised oracle

We define the following oracle estimator:

$$\tilde{\eta}_1 = \text{sign}(\mathbf{H}(\tilde{A}\tilde{A}^\top)\eta_1)$$

with $\tilde{A} = A - p\mathbf{1}_{n_1}\mathbf{1}_{n_2}^\top$.

Proposition 1 [Ndaoud, Sigalla and Tsybakov, 2019]

Assume that $n_2 \geq n_1 \log n_1$ and $\gamma_1 = \gamma_2 = 0$. There exists $c_\delta > 0$ depending only on δ such that if $p^2 = c_\delta \frac{\log n_1}{n_1 n_2}$ then for the oracle $\tilde{\eta}_1$ we have

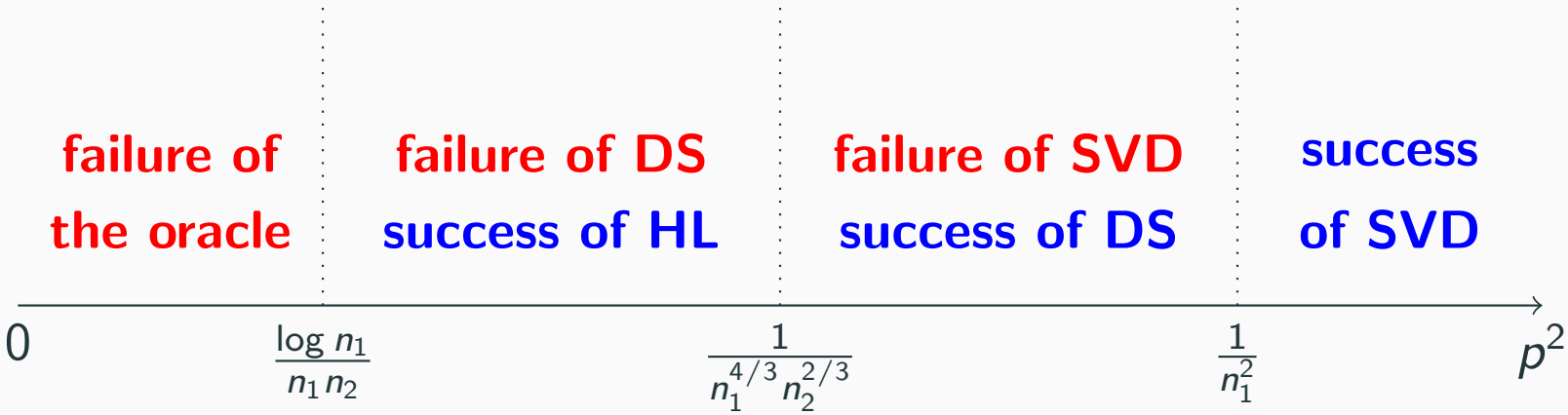
$$\lim_{n_1 \to \infty} \sum_{i=1}^{n_1} \mathbb{P}(\tilde{\eta}_{1i} \neq \eta_{1i}) = \infty.$$

Hence, the condition $p = \Omega\left(\sqrt{\log n_1/(n_1 n_2)}\right)$ is necessary for the supervised oracle to achieve exact recovery when $n_2 \geq n_1 \log n_1$.

Numerical experiments

Comparison of different estimators

- Comparison of the three methods: SVD, debiased spectral (DS) and hollowed Lloyd's (HL), in the general case of imbalance, can be summarized as follows :



Numerical experiments

- For the sake of readability of plots, we define the parameters a and b such that

$$p = \sqrt{a}/n_1 \quad \text{and} \quad b = n_1(\log n_1)/n_2.$$

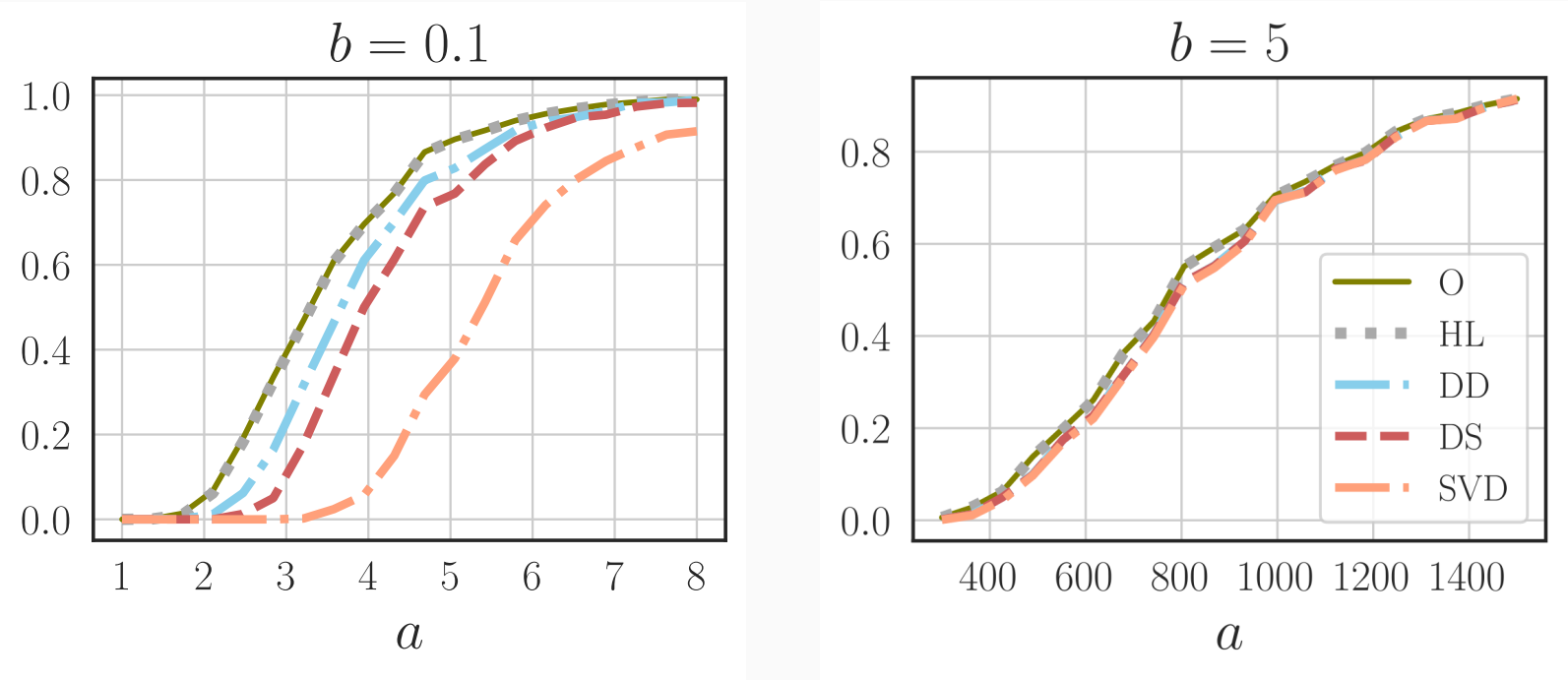


Figure 2: Empirical probability of success over 1000 runs of the experiment for: $b = 0.1$ (left) and $b = 5$ (right).

Control of the spectral norm of the noise

Control of the spectral norm of the hollowed Gram matrix

- Control of the spectral norm of $H(WW^\top) = \sum_{j=1}^{n_2} H(W_j W_j^\top)$?

Theorem 3 Matrix Bernstein inequality - adapted from [Tropp, 2012] - theorem 6.2

Let $(Y_j)_{j=1}^n$ be a sequence of independent symmetric random matrices of size $d \times d$, and $a, R > 0$. Assume that for all j in $\{1, \dots, n\}$ we have

$$\mathbb{E}(Y_j) = 0 \text{ and } \|\mathbb{E}(Y_j^q)\|_{op} \leq \frac{q!}{2} R^{q-2} a^2 \text{ for } q = 2, 3, \dots \ .$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left(\left\| \sum_{j=1}^{n} Y_j \right\|_{op} \geq t \right) \leq d \exp\left(-\frac{t^2}{2\sigma^2 + 2Rt} \right) \text{ with } \sigma^2 = na^2.$$

- Then: apply Theorem 3 with $Y_j = H(W_j W_j^\top)$, $d = n_1$, $n = n_2$, $R = 3(1 + 2n_1 p)$ and $a^2 = 4p^2 n_1$.

Insight of the proof

- **Difficulty**: to prove for $q = 2, 3, \ldots$ that

$$\left\|\mathbb{E}(H(W_j W_j^\top)^q)\right\|_\infty \leq 2q!(3(1+2n_1 p))^{q-2} p^2 n_1.$$

- **Case $q = 2$**: simple.
- **Case $q \geq 3$**: requires sophisticated combinatorial arguments.

Conclusion

Summary

1. Improved **sufficient** conditions for spectral procedures achieving almost full recovery.
2. An **efficient adaptive** procedure that achieves exact recovery under milder conditions. Outperforms previous algorithms (cf. simulations).
3. Hint that our conditions are **necessary** through the study of an oracle estimator.
4. **More general approach** with p unknown and imbalance parameters γ_1, γ_2.

Discussion

Discussion:

1. Necessary condition.
2. Recent works [Abbe *et al.*, 2020, Löffler *et al.*, 2019] show that **spectral initialization** already achieves exact recovery for **SBM** and **GMM**.
   Is it still the case for **BSBM** in the **high-dimensional** regime?

Thank you for your attention.

Bibliography i

Abbe, Emmanuel, Bandeira, Afonso S, & Hall, Georgina. 2015.
Exact recovery in the stochastic block model.
IEEE Transactions on Information Theory, **62**(1), 471–487.

Abbe, Emmanuel, Fan, Jianqing, Wang, Kaizheng, Zhong, Yiqiao, *et al.* 2020.
Entrywise eigenvector analysis of random matrices with low expected rank.
Annals of Statistics, **48**(3), 1452–1474.

Bibliography ii


Feldman, Vitaly, Perkins, Will, & Vempala, Santosh. 2015.
Subsampled power iteration: a unified algorithm for block models and planted csp's.
Pages 2836–2844 of: Advances in Neural Information Processing Systems.


Florescu, Laura, & Perkins, Will. 2016.
Spectral thresholds in the bipartite stochastic block model.
Pages 943–959 of: Conference on Learning Theory.


Giraud, Christophe, & Verzelen, Nicolas. 2019.
Partial recovery bounds for clustering with the relaxed K-means.
Mathematical Statistics and Learning, **1**(3), 317–374.

Bibliography iii


Löffler, Matthias, Zhang, Anderson Y, & Zhou, Harrison H. 2019.
Optimality of spectral clustering for gaussian mixture model.
arXiv preprint arXiv:1911.00538.


Lu, Yu, & Zhou, Harrison H. 2016.
Statistical and computational guarantees of lloyd's algorithm and its variants.
arXiv preprint arXiv:1612.02099.


Ndaoud, Mohamed. 2018.
Sharp optimal recovery in the Two Component Gaussian Mixture Model.
In: arXiv preprint, arXiv:1812.08078.

Bibliography iv

Ndaoud, Mohamed, Sigalla, Suzanne, & Tsybakov, Alexandre B. 2019.
Improved clustering algorithms for the Bipartite Stochastic Block Model.

Royer, Martin. 2017.
Adaptive clustering through semidefinite programming.
Pages 1795–1803 of: Advances in Neural Information Processing Systems.

Tropp, Joel A. 2012.
User-friendly tail bounds for sums of random matrices.
Foundations of computational mathematics, **12**(4), 389–434.