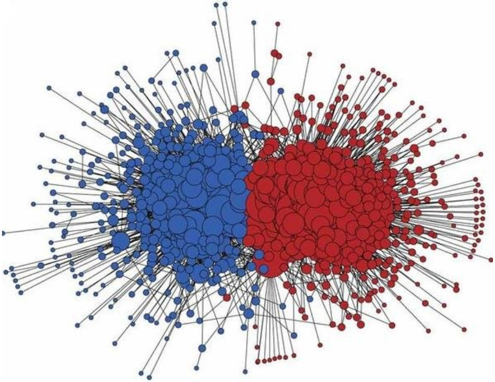
An adaptive spectral algorithm for the recovery of overlapping communities in networks

Emilie Kaufmann,

joint work with Thomas Bonald and Marc Lelarge

LINCS Seminar, June 10th, 2015

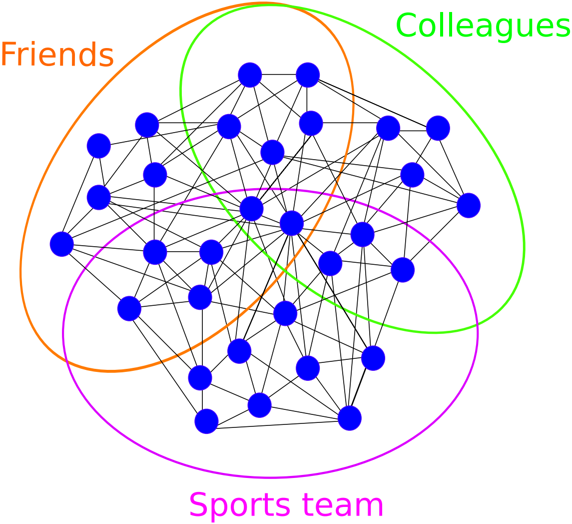
Example : partitionning a network



Political blogs network

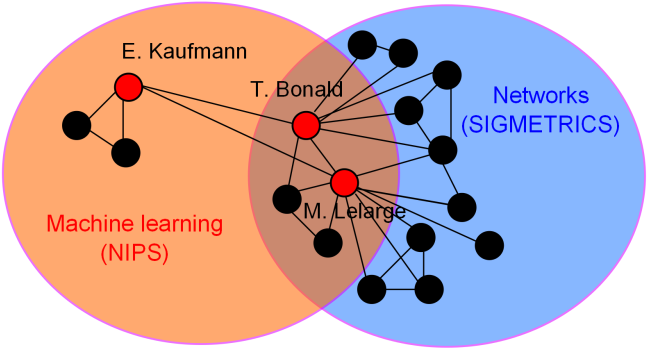
Overlapping communities : examples

- **Ego-network**



Overlapping communities : examples

- **Co-authorship network**



Random graph models

Idea : Assume that the observed graph is drawn from a random graph model that depends on (hidden) communities

- inspires model-based methods for community detection (community detection = estimation problem)
- can be used for evaluation purpose :
  - ➜ try algorithms on simulated data
  - ➜ consistency results : proof that the hidden communities are recovered (if the network is sufficiently large/dense)

Outline

Outline

The Stochastic Block-Model (SBM)

Definition

An undirected, unweighted graph with n nodes is drawn under the random graph model with expected adjacency matrix A if

$$\forall i \leq j, \quad \hat{A}_{i,j} \sim \mathcal{B}(A_{i,j})$$

where $\hat{A}_{i,j}$ is the observed adjacency matrix.

The stochastic block-model with parameter K,Z,B :

- n nodes, K communities
- a mapping $k : \{1, \ldots, n\} \longrightarrow \{1, \ldots, K\}$
- a connectivity matrix $B \in \mathbb{R}^{K \times K}$

The expected adjacency matrix is

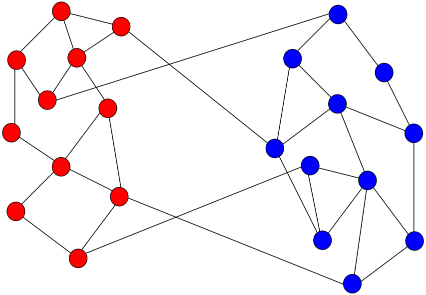
$$A_{i,j} = B_{k(i),k(j)} = (ZBZ^T)_{i,j}$$

for a membership matrix $Z \in \mathbb{R}^{n \times K}$: $Z_{i,l} = \delta_{k(i),l}$.

The Stochastic Block-Model (SBM)

Example : $K = 2$, for $p > q$,

$$B = \begin{pmatrix} p & q \\ q & p \end{pmatrix}$$



SBM : a motivation for a spectral algorithm

$$A_{i,j} = B_{k(i),k(j)}$$

Observation 1 : A is constant on communities :

$$A_{i,\cdot} = A_{j,\cdot} \;\Leftrightarrow\; k(i) = k(j)$$

(due to noise, won't be the case for $\hat{A}$)

Obsevation 2 : this property is preserved for the matrix

$$U = [u_1 | \dots | u_K] \in \mathbb{R}^{n \times K}$$

that contains eigenvectors of A associated to non-zero eigenvalues :

$$U_{i,\cdot} = U_{j,\cdot} \;\Leftrightarrow\; k(i) = k(j)$$

(not too far from the truth for an empirical version $\hat{U}$?)

Spectral clustering with the adjacency matrix

$(\hat{A}_{i,j})$ adjacency matrix of the observed graph

Step 1 : *spectral embedding*
Compute $\hat{U} = [\hat{u}_1|\dots|\hat{u}_K] \in \mathbb{R}^{n\times K}$, matrix of K eigenvectors of $\hat{A}$ associated to largest eigenvalues

$$\text{node } i \ \rightarrow \ \text{vector } \hat{U}_{i,\cdot} \in \mathbb{R}^K$$

Step 2 : *clustering phase*
Perform clustering in $\mathbb{R}^K$ on the vectors representing the nodes (the rows of $\hat{U}$), e.g. K-means clustering

Remarks :

- other possible spectral embeddings (e.g. Laplacian)
- other possible justifications for spectral algorithms

[Von Luxburg 08, Newman 13]

Outline

The model

Definition

The Stochastic Block-Model with Overlap (SBMO) has expected adjacency matrix

$$A = ZBZ^T$$

that depends on K, a connectivity matrix $B \in \mathbb{R}^{K\times K}$, and a membership matrix $Z \in \{0,1\}^{n\times K}$.

$Z_i := Z_{i,\cdot} \in \{0,1\}^{1\times K}$: indicates the communitieS to which node i belongs

Our goal : Given $\hat{A}$ drawn under SBMO, build an estimate $\hat{K}$ of K and $\hat{Z}$ of Z (up to a permutation of its columns).

Performance metrics

Two criterion to minimize :

- **number of misclassified nodes** :

$$\mathrm{MisC}(\hat{Z}, Z) = \min_{\sigma \in \mathfrak{S}_K} \left| \{ i \in \{1, \dots, n\} : \exists k \in \{1, \dots, K\}, \hat{Z}_{i,\sigma(k)} \neq Z_{i,k} \} \right|$$

- **estimation error** :

$$\mathrm{Error}(\hat{Z}, Z) = \frac{1}{nK} \inf_{\sigma \in \mathfrak{S}_K} ||\hat{Z} P_\sigma - Z||_F^2$$

(if $\hat{K} \neq K$, $\mathrm{MisC}(\hat{Z}, Z) = n$ and $\mathrm{Error}(\hat{Z}, Z) = 1$).

Identifiability

To perform estimation, the model needs to be identifiable :

$$Z'B'Z'^T = ZBZ^T \quad \Rightarrow \quad \mathrm{MisC}(Z', Z) = 0.$$

- Not always the case ! $ZBZ^T = Z'B'Z'^T = Z''B''Z''^T$, with

$$B = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \quad Z = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} \end{pmatrix}$$

$$B' = \begin{pmatrix} a+b & b & a \\ b & b+c & c \\ a & c & a+c \end{pmatrix} \quad Z' = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \end{pmatrix}$$

$$B'' = \begin{pmatrix} a+b-c & b-c & a-c & 0 \\ b-c & b & 0 & 0 \\ a-c & 0 & a & 0 \\ 0 & 0 & 0 & c \end{pmatrix} \quad Z'' = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{1} \end{pmatrix}$$

Identifiability

To perform estimation, the model needs to be identifiable :

$$Z'B'Z'^T = ZBZ^T \quad \Rightarrow \quad \mathrm{MisC}(Z', Z) = 0.$$

Theorem

The SBMO is identifiable under the following assumptions :

(SBMO1) B is invertible ;

(SBMO2) each community contains at least one pure node :

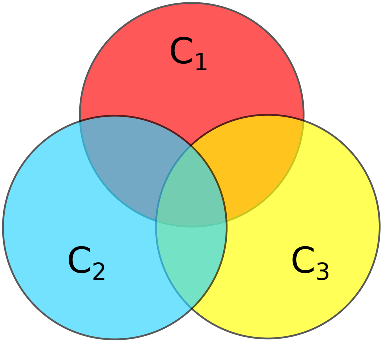
$$\forall k \in \{1, \ldots, K\}, \exists i \in \{1, \ldots, n\} : Z_{i,k} = \sum_{\ell=1}^{K} Z_{i,\ell} = 1.$$

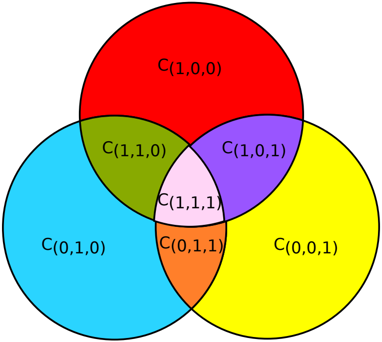
Outline

SBMO or SBM ?

SBMO(K,B,Z) can be viewed as a particular case of SBM with

- communities indexed by $\mathcal{S} = \{z \in \{0,1\}^{1\times K} : \exists i : Z_i = z\}$
- $B'_{z,z'} = zBz'^T$





Start by reconstructing the underlying SBM ? Not a good idea.

Spectrum of the adjacency matrix under the SBMO

$A = ZBZ^T$ the expected adjacency matrix of an identifiable SBMO :

- $Z \in \mathcal{Z} := \{Z \in \{0,1\}^{n\times K}, \forall k \in \{1,\dots,K\} \exists i : Z_i = \mathbb{1}_{\{k\}}\}$.
- A is of rank K

$U = [u_1|\dots|u_K]$ a matrix whose columns are K normalized eigenvectors associated to the non-zero eigenvalues of A.

Proposition

1. there exists $X \in \mathbb{R}^{K\times K}$: $U = ZX$
2. for all $Z' \in \mathcal{Z}$ and $X' \in \mathbb{R}^{K\times K}$, if $U = Z'X'$, there exists $\sigma \in \mathfrak{S}_k : Z = Z'P_\sigma$

($u_1,\dots,u_K$ form a basis of $\mathrm{Im}(A)$ and $\mathrm{Im}(A) \subset \mathrm{Im}(Z)$)

Combinatorial spectral clustering

This motivates the following estimation procedure :

$$(\mathcal{P}): \quad (\hat{Z}, \hat{X}) \in \underset{Z' \in \mathcal{Z}, X' \in \mathbb{R}^{K \times K}}{\operatorname{argmin}} ||Z'X' - \hat{U}||_F^2,$$

where $\hat{U}$ is a matrix that contains eigenvector associated to the K largest eigenvalues of $\hat{A}$ (in absolute value).

$$||M||_F^2 = \sum_{i,j} M_{i,j}^2 = \sum_i ||M_{i,\cdot}||^2 = \sum_j ||M_{\cdot,j}||^2$$

In practice : Combinatorial spectral clustering computes an (approximate) solution of

$$(\mathcal{P})': \quad (\hat{Z}, \hat{X}) \in \underset{\substack{Z' \in \{0,1\}^{n \times K} : \forall i, Z'_i \neq 0 \\ X' \in \mathbb{R}^{K \times K}}}{\operatorname{argmin}} ||Z'X' - \hat{U}||_F^2.$$

If K is unknown, let $\hat{K}$ be the number of eigenvalues λ of $\hat{A}$ satisfying $|\lambda| \geq \sqrt{2(1+\eta)\, \hat{d}_{\max}(n) \log(4n^{1+r})}$.

Outline

Goal

- Under which conditions is

$$(\mathcal{P}) : \quad (\hat{Z}, \hat{X}) \in \underset{Z' \in \mathcal{Z}, X' \in \mathbb{R}^{K \times K}}{\operatorname{argmin}} ||Z'X' - \hat{U}||_F^2,$$

a good estimation procedure ?

- We present the analysis of a slight variant :

$$(\mathcal{P}_\epsilon) : \quad (\hat{Z}, \hat{X}) \in \underset{Z' \in \mathcal{Z}_\epsilon, X' \in \mathbb{R}^{K \times K}}{\operatorname{argmin}} ||Z'X' - \hat{U}||_F^2,$$

$$\mathcal{Z}_\epsilon = \left\{ Z' \in \{0,1\}^{n \times K}, \, \forall k \in \{1, \dots, K\}, \frac{|\{i : Z'_i = \mathbb{1}_{\{k\}}\}|}{n} > \epsilon \right\}.$$

for ϵ smaller than the smallest proportion of pure nodes.

Scaling

To analyze the solution of $(\mathcal{P}_\epsilon)$ when the network grows,

$$A = \frac{\alpha_n}{n} Z B Z^T,$$

with α_n a degree parameter, B independent of n, $Z \in \{0,1\}^{n \times K}$.

$$d_i(n) = \sum_{j=1}^{n} A_{i,j} = \alpha_n \left(\frac{1}{n} Z_i B Z^T \mathbf{1} \right)$$

Assumption : overlap matrix

There exists some matrix $O \in \mathbb{R}^{K \times K}$, called the overlap matrix :

$$\frac{1}{n} Z^T Z \to O.$$

$O_{k,l}$: (limit) proportion of nodes belonging to communities k and l

A precise characterization of the spectrum

The spectrum of A can be related to the spectrum of $K \times K$ matrices that are independent on n :

Proposition

Let $\mu \neq 0$. The following statements are equivalent :

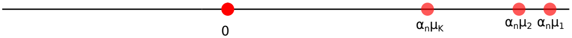
1. x is an eigenvector of $M_0 := O^{1/2} B O^{1/2}$ associated to μ
2. $u = Z O^{-1/2} x$ is an eigenvector of A associated to $\alpha_n \mu$

In particular, the non-zero eigenvalues of A are of order $O(\alpha_n)$.

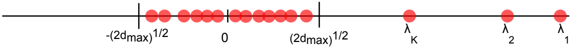
Step 1 : Why is $\hat{U}$ close to U ?

Heuristic :

- Spectrum of A



- Spectrum of $\hat{A} = A +$ perturbation



Extra ingredient : the Davis-Kahan theorem (linear algebra) to prove that the associated eigenvectors are close

An adaptive eigenvectors perturbation result

Let $\hat{K} = \left| \left\{ \lambda \in \mathrm{Sp}(\hat{A}) : |\lambda| \geq \sqrt{2\,(1+\eta)\,\hat{d}_{\max}(n) \log(4n/\delta)} \right\} \right|$

and $\hat{U} \in \mathbb{R}^{n \times \hat{K}}$ a matrix that contains normalized eigenvectors of $\hat{A}$ associated with the largest $\hat{K}$ eigenvalues. If

$$\begin{aligned} d_{\max}(n) &\geq C_1(\eta) \log\left(n/\delta\right), \\ \lambda_{\min}(A)^2/d_{\max}(n) &> C_2(\eta) \log\left(n/\delta\right), \end{aligned}$$

for some constants $C_1(\eta), C_2(\eta)$, then with probability larger than $1-\delta$, $\hat{K} = \mathrm{Rank}(A)$ and there exists $\hat{P} \in \mathcal{O}_K(\mathbb{R})$ such that

$$\left|\left|\hat{U} - U\hat{P}\right|\right|_F^2 \leq 32 \left(1 + \frac{\eta}{\eta+2}\right) \left(\frac{d_{\max}(n)}{\lambda_{\min}(A)^2}\right) \log\left(\frac{4n}{\delta}\right).$$

In the SBMO, $\left\{ \begin{array}{rcl} d_{\max}(n) & = & O(\alpha_n) \\ \lambda_{\min}(A) & = & \mu_0 \alpha_n \end{array} \right.$: we need $\frac{\alpha_n}{\log(n)} \to \infty$.

Step 2 : Sensitivity to noise

There exists $V \in \mathcal{O}_K(\mathbb{R})$ (eigenvectors of M_0) such that

$$U = ZX \quad \text{with} \quad X = \frac{1}{\sqrt{n}} O^{-1/2} V.$$

Let

$$d_0 := \min_{\substack{z \in \{-1,0,1,2\}^{1\times K} \\ z \neq 0}} \left|\left| z O^{-1/2} \right|\right| > 0.$$

Lemma

Let $Z' \in \mathbb{R}^{n\times K}$, $X' \in \mathbb{R}^{K\times K}$ and $\mathcal{N} \subset \{1, \dots, n\}$. Assume that

1. $\forall i \in \mathcal{N}$, $||Z'_i X' - U_i|| \leq \frac{d_0}{4K\sqrt{n}}$
2. there exists $(i_1, \dots, i_K) \in \mathcal{N}^K$ and $(j_1, \dots, j_K) \in \mathcal{N}^K$: $\forall k \in [1, K]$, $Z_{i_k} = Z'_{j_k} = \mathbb{1}_{\{k\}}$

Then there exists a permutation matrix P_σ such that

$$\forall i \in \mathcal{N}, Z_i = (Z' P_\sigma)_i.$$

The result

Let $\eta \in]0, 1/2[$ and $r > 0$. Let

$$\hat{K} = \left| \left\{ \lambda \in \mathrm{Sp}(\hat{A}) : |\lambda| \geq \sqrt{2\,(1+\eta)\,\hat{d}_{\max}(n) \log(4n^{1+r})} \right\} \right|$$

and $\hat{U} \in \mathbb{R}^{n \times \hat{K}}$ a matrix that contains normalized eigenvectors of $\hat{A}$ associated with the largest $\hat{K}$ eigenvalues.

$$(\mathcal{P}_\epsilon) : \quad (\hat{Z}, \hat{X}) \in \underset{Z' \in \mathcal{Z}_\epsilon, X' \in \mathbb{R}^{\hat{K} \times \hat{K}}}{\mathrm{argmin}} ||Z'X' - \hat{U}||_F^2.$$

Assume that $\frac{\alpha_n}{\log n} \to \infty$. There exists a constant $C_1 > 0$ such that, for n large enough,

$$\mathbb{P}\left(\frac{\mathrm{MisC}(\hat{Z}, Z)}{n} \leq \frac{C_1 K^2}{d_0^2 \mu_0^2} \frac{\log(4n^{1+r})}{\alpha_n} \right) \geq 1 - \frac{1}{n^r}.$$

Outline

Combinatorial Spectral Clustering (CSC)

- **Step 1** : spectral embedding based on the adjacency matrix : compute $\hat{U}$, the matrix of K leading eigenvectors of $\hat{A}$
- **Step 2** : compute an approximation of the solution of $(\mathcal{P}')$

$$(\mathcal{P})' : \quad (\hat{Z}, \hat{X}) \in \underset{\substack{Z' \in \{0,1\}^{n \times K} : \forall i, Z'_i \neq 0 \\ X' \in \mathbb{R}^{K \times K}}}{\operatorname{argmin}} ||Z'X' - \hat{U}||_F^2 .$$

using alternate minimization.

$$||Z'X' - \hat{U}||_F^2 = \sum_{i=1}^{n} ||Z'_i X' - \hat{U}_i||^2$$

Combinatorial Spectral Clustering (CSC)

Algorithm 1 Adaptive Combinatorial Spectral Clustering for Overlapping Community Detection

Require: Parameters $\epsilon, r, \eta > 0$. Upper bound on the maximum overlap $O_{\max}$.

Require: $\hat{A}$, the adjacency matrix of the observed graph.

♯ Selection of the eigenvectors

Form $\hat{U}$ a matrix whose columns are $\hat{K}$ eigenvectors of $\hat{A}$ associated to eigenvalues λ satisfying

$$|\lambda| > \sqrt{2(1+\eta)\hat{d}_{\max}(n)\log(4n^{1+r})}$$

♯ Initialization

$\hat{Z} = 0 \in \mathbb{R}^{n\times\hat{K}}$

$\hat{X} \in \mathbb{R}^{\hat{K}\times\hat{K}}$ initialized with k-means++ applied to $\hat{U}$, the first centroid being chosen at random among nodes with degree smaller than the median degree

Loss $= +\infty$

♯ Alternating minimization

while (*Loss* $- ||\hat{Z}\hat{X} - \hat{U}||_F^2 > \epsilon$) **do**

    Loss $= ||\hat{Z}\hat{X} - \hat{U}||_F^2$

    Update membership vectors: $\forall i,\ \hat{Z}_{i,\cdot} = \underset{z\in\{0,1\}^{1\times\hat{K}}:1\leq||z||_1\leq O_{\max}}{\arg\min} ||\hat{U}_{i,\cdot} - z\hat{X}||$.

    Update centroids: $\hat{X} = (\hat{Z}^T\hat{Z})^{-1}\hat{Z}^T\hat{U}$.

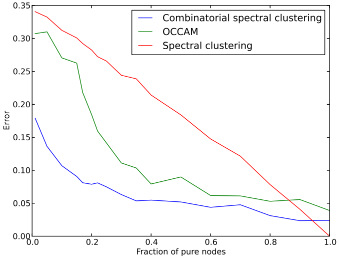
end while

Experiments on simulated data

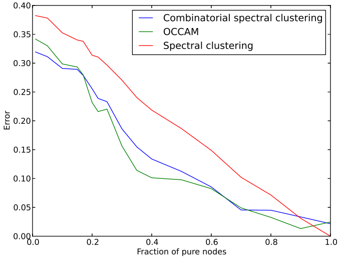
CSC versus two spectral algorithms :

- Normalized Spectral Clustering (SC)
- the OCCAM spectral algorithm (OCCAM) [Zhang et al. 14]

$n = 500$, $K = 5$, $\alpha_n = (\log n)^{1.5}$, $B = \mathrm{Diag}(5, 4, 3, 3, 3)$,
Z : fraction p of pure nodes, $O_{\max} \leq 3$.



under SBMO



under OCCAM

Experiments on real-world networks

Ego-networks from the ego-networks dataset
(SNAP, [Mc Auley, Leskovec 12])

	n	K	c	O_{max}	FP	FN	Error
SC	190 (173)	3.17 (1.07)	1.09 (0.06)	2.17 (0.37)	0.200 (0.110)	0.139 (0.107)	0.120 (0.083)
OCC.	190 (173)	3.17 (1.07)	1.09 (0.06)	2.17 (0.37)	0.176 (0.176)	0.113 (0.084)	0.127 (0.102)
CSC	190 (173)	3.17 (1.07)	1.09 (0.06)	2.17 (0.37)	0.125 (0.067)	0.101 (0.062)	0.102 (0.049)

TABLE: Spectral algorithms recovering overlapping friend circles in ego-networks from Facebook (average over 6 networks).

Experiments on real-world networks

Co-authorship networks built from DBLP

$$\mathcal{C}_1 = \{\text{NIPS}\},\ \mathcal{C}_2 = \{\text{ICML}\},\ \mathcal{C}_3 = \{\text{COLT}, \text{ALT}\}$$

$$n = 9272,\ \ K = 3,\ \ d_{\text{mean}} = 4.5$$

	c	$\hat{c}$	FP	FN	Error
SC	1.22	1.	0.38	0.39	0.39
OCCAM	1.22	1.02	0.43	0.41	0.42
CSC	1.22	1.04	0.26	0.28	0.27

$$\mathcal{C}_1 = \{\text{ICML}\},\ \mathcal{C}_2 = \{\text{COLT}, \text{ALT}\}.$$

$$n = 4374,\ \ K = 2,\ \ d_{\text{mean}} = 3.8$$

	c	$\hat{c}$	FP	FN	Error
SC	1.09	1.	0.39	0.55	0.46
OCCAM	1.09	1.01	0.29	0.44	0.36
CSC	1.09	1.03	0.21	0.31	0.25

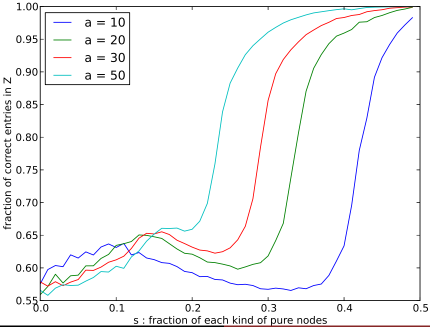
Experiments in the sparse case

A simple SBMO : $A = \frac{\alpha_n}{n} Z B Z^T$

$$B = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} \quad Z = \begin{pmatrix} \mathbf{1}_{sn} & 0 \\ \mathbf{1}_{(1-2s)n} & \mathbf{1}_{(1-2s)n} \\ 0 & \mathbf{1}_{sn} \end{pmatrix},$$

$s \in]0, 1/2[$: fraction of pure nodes in each community.

We set $\alpha_n = 1$ (very sparse network) :



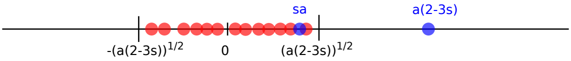
Spectrum of A ($\alpha_n = 1$) :

$$a(2-3s) \quad > \quad sa$$

$$X = \begin{pmatrix} \mathbf{1}_{sn} \\ \mathbf{2}_{(1-2s)n} \\ \mathbf{1}_{sn} \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} -\mathbf{1}_{sn} \\ \mathbf{0}_{(1-2s)n} \\ \mathbf{1}_{sn} \end{pmatrix}.$$

The eigenvectors λ of $\hat{A}$ associated to the noise should satisfy

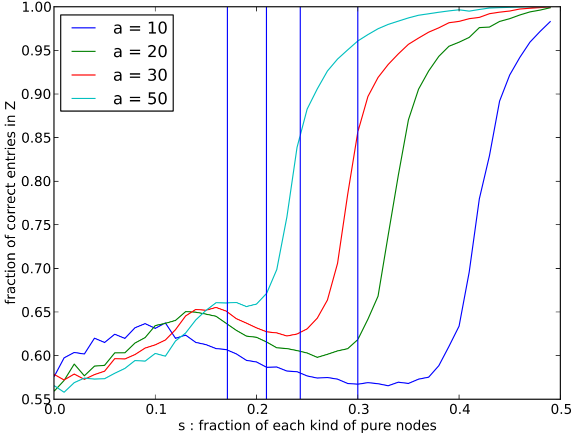
$$|\lambda| < \sqrt{a(2-3s)}$$



Conjecture :
If $s^2 a < 2 - 3s$, it is impossible to classify the pure nodes better than by random guessing.

Experiments in the sparse case

- Adding the threshold



Conclusion

Combinatorial Spectral Clustering = a spectral algorithm that uses the geometry of the eigenvectors of the adjacency matrix under the SBMO to directly identify overlapping communities

Future work :

- further explore the phase transition in the sparse case
- find heuristics for solving $(\mathcal{P}')$ more efficiently
- are other spectral embeddings possible ?
- can the pure nodes assumption be relaxed ?

References

- E. Kaufmann, T. Bonald, M. Lelarge, *An adaptive spectral algorithm for the recovery of overlapping communities in networks* (soon on arXiv !)
- J. Mc Auley and J. Leskovec, *Learning to discover social circles in ego networks*, 2012
- M. Newman, *Spectral methods for network community detection and graph partitioning*, 2013
- U. Von Luxburg, *A tutorial on Spectral Clustering*, 2007
- Y. Zhang, E. Levina, J. Zhu, *Detecting Overlapping Communities in Networks with Spectral Methods*, 2014