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

Emilie Kaufmann, Thomas Bonald and Marc Lelarge

In a nutshell

We present *combinatorial spectral clustering* (CSC), a simple spectral algorithm designed to identify overlapping communities, motivated by a random graph model called *stochastic blockmodel with overlap* (SBMO).

A random graph model : the SBMO

A network with n nodes is drawn from the SBMO if its observed adjacency matrix $\hat{A}$ satisfies $\mathbb{E}[\hat{A}] = A$, with

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

where

- K is the number of communities
- $B \in \mathbb{R}^{K\times K}$ is the community connectivity matrix, independent of n
- $Z \in \{0,1\}^{n\times K}$ is the community membership matrix, satisfying

$$\forall z \in \mathcal{S}, \ \frac{|\{i : Z_i = z\}|}{n} \to \beta_z.$$

- α_n is a degree parameter

Goal: Propose a good estimate $\hat{Z}$ of Z, up to a permutation of the rows

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

Identifiability: (needed to perform estimation !) If B, B' are invertible and Z, Z' have at least one pure node per community, ie belong to

$$\mathcal{Z} = \{Z \in \{0,1\}^{n\times K}, \forall k \in \{1,\dots,K\}, \exists i \in \{1,\dots,n\}, Z_{i,k} = \sum_\ell Z_{i,\ell} = 1\},$$

then

$$\frac{\alpha_n}{n} Z B Z^T = \frac{\alpha'_n}{n} Z' B' (Z')^T \ \Rightarrow \ \text{Err}(Z', Z) = 0.$$

Motivation: spectral analysis

- Spectral analysis of the expected adjacency matrix

Let $U = [u_1|\dots|u_K] \in \mathbb{R}^{n\times K}$ be a matrix whose columns are K normalized eigenvectors associated to the K non-zero eigenvalues of A.

Proposition 1 *1. There exists $X \in \mathbb{R}^{K\times K}$ such that $U = ZX$.*

2. If $U = Z'X'$ for some $Z' \in \mathcal{Z}$, $X' \in \mathbb{R}^{K\times K}$, then there exists $\sigma \in \mathfrak{S}_K$ such that $Z = Z'P_\sigma$.

- Spectral analysis of the observed adjacency matrix

Let $\hat{U}$ be at matrix whose columns are K normalized eigenvectors associated to the K largest eigenvalues of $\hat{A}$.
$\hat{U}$ is close to U if the degrees in the graph are large enough, which motivates

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

The CSC algorithm

Combinatorial Spectral Clustering (CSC) proceeds as follows:

1. **Spectral embedding:** compute the matrix $\hat{U}$ of K eigenvectors of $\hat{A}$ associated to the largest eigenvalues (in absolute value).
2. **Community reconstruction:** compute an approximation of

$$(\mathcal{P})' : \quad (\hat{Z}, \hat{X}) \in \underset{Z' \in \{0,1\}^{n\times K}, X' \in \mathbb{R}^{K\times K}}{\text{argmin}} ||Z'X' - \hat{U}||_F^2$$

using alternate minimization and a suitable initialization.

An adaptive version: If K is unknown, we let $\hat{K}$ be the number of eigenvalues (with multiplicity) satisfying

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

for some constants r and η.

Consistency properties

Let $\mathcal{Z}_\epsilon$ be the set of membership matrices for which the proportion of pure nodes in each community is larger than ϵ:

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

Theorem 2 *Let $\eta \in]0,1/2[$ and $r > 0$. Let $\hat{U}$ be a matrix whose columns are orthogonal eigenvectors of $\hat{A}$ associated to an eigenvalue $\hat{\lambda}$ satisfying*

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

Let $\hat{K}$ be the number of such eigenvectors. Let

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

Assume that $\frac{\alpha_n}{\log n} \to \infty$ and $\epsilon < \min_z \beta_z$. There exists a positive constant C_1 such that, for n large enough, with probability larger than $1 - n^{-r}$, $\hat{K} = K$ and

$$\text{Err}(\hat{Z}, Z) \leq \frac{K^2 C_1}{d_0^2 \mu_0^2} \frac{\log(4n^{1+r})}{\alpha_n},$$

where μ_0 and d_0 depend on B and $O = \lim_{n\to\infty} \frac{1}{n}(Z^T Z)$, two $K \times K$ matrices that are independent of n.

Practical implementation

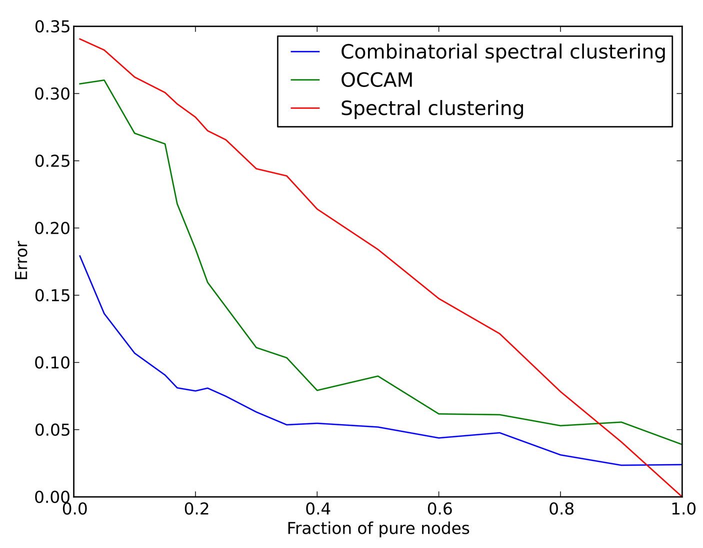
Initialization: K-means++ procedure with first centroid chosen at random among nodes whose degree is smaller than the median degree
Alternate minimization:

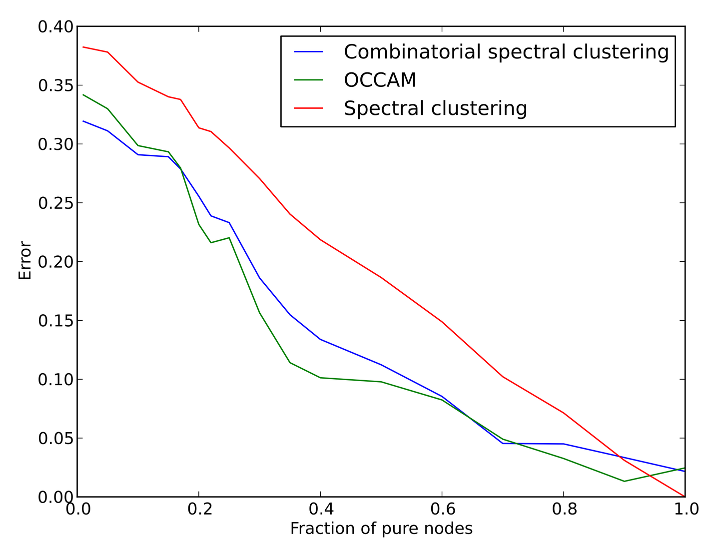
- X' fixed. $\forall i$, $Z'_i = \underset{z \in \{0,1\}^{1\times K} : 1 \leq ||z|| \leq O_{\max}}{\text{argmin}} ||zX' - \hat{U}_{i,.}||$
- Z' fixed. If $Z'^T Z'$ is invertible, $X' = (Z'^T Z')^{-1} Z'^T \hat{U}$ (else, re-initialize X')

Empirical performance

We compare empirically the performance of three spectral algorithms: Spectral Clustering (SC), designed for non-overlapping communities, CSC and another spectral algorithm recently proposed by [2] and inspired by a random graph model called OCCAM.

- Simulated data





Comparison of the algorithms under instances of the SBMO (left) and the OCCAM (right) $n = 500$, $K = 5$, $O_{\max} = 3$, average over 100 networks

- Real-world networks

	n	K	c	$O_{\max}$	Error	NVI
SC	190 (173)	3.17 (1.07)	1.09 (0.06)	2.17 (0.37)	0.120 (0.083)	0.556 (0.256)
OCCAM	190 (173)	3.17 (1.07)	1.09 (0.06)	2.17 (0.37)	0.127 (0.102)	0.556 (0.280)
CSC	190 (173)	3.17 (1.07)	1.09 (0.06)	2.17 (0.37)	0.102 (0.049)	0.544 (0.217)

Performance of the three algorithms averaged over 6 Facebook ego-networks in terms of error and normalized variation of information.

References

[1] E. Kaufmann, T. Bonald, M. Lelarge An Adaptive Spectral Algorithm for the Recovery of Overlapping Communities in Networks *arXiv:1506.04158*, 2015

[2] Zhang, Y., Levina, E., and Zhu, J. Detecting Overlapping Communities in Networks with Spectral Methods. *arXiv:1412.3432* , 2014