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 A satisfies E[A] = A, with

\[ A = \frac{\alpha}{n} ZBZ^T, \]

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
  \[ \forall z \in S, \frac{|\{i : Z_i = z\}|}{n} \to \beta_z. \]
- \( \alpha_n \) is a degree parameter

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

\[ \text{Err}(\hat{Z}, Z) = \frac{1}{nk} \inf_{\hat{Z}P_z} |\hat{Z}P_z - Z||_F. \]

Identification: (needed to perform estimation) If B, B' are invertible and Z, Z' have at least one pure node degree, i.e. belong to \( Z = \{Z \in \{0,1\}^{n \times K} \mid \forall k \in \{1,\ldots,K\}, \exists i \in \{1,\ldots,n\}, Z_{i,k} = \sum_{\ell} Z_{i,\ell} = 1\}, \)

then

\[ \alpha_n ZBZ^T = \alpha_n Z'B'(Z')^T \Rightarrow \text{Err}(\hat{Z}, Z) = 0. \]

Motivation: spectral analysis

- Spectral analysis of the expected adjacency matrix

Let U = [u_1 | ... | u_n] 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 \( U = ZX', \) then \( \exists \in \mathbb{R}_{\geq 0} \) such that \( Z = ZX'P_z. \)

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 \( \hat{A}. \)

U is close to \( \hat{U} \) if the degrees in the graph are large enough, which motivates

\[ \min_{Z} (\tilde{Z}, \tilde{X}) = \arg \min_{Z \in \{0,1\}^{n \times K}, X \in \mathbb{R}^{K \times K}} ||ZX' - \hat{U}||_F. \]

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 \( Z \)

\[ (\tilde{Z}, \tilde{X}) = \arg \min_{Z \in \{0,1\}^{n \times K}, X \in \mathbb{R}^{K \times K}} ||ZX' - \hat{U}||_F \]

using alternate minimization and a suitable initialization.

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

\[ |\l| > \sqrt{2(1 + n)\delta_{max}(n) \log(4n^{1+r})}, \]

for some constants \( r \) and \( \eta. \]

Consistency properties

Let \( Z \) be the set of membership matrices for which the proportion of pure nodes in each community is larger than

\[ Z = \left\{ Z \in \{0,1\}^{n \times K}, \forall k \in \{1,\ldots,K\}, \frac{|\{i : Z_i = 1(k)\}|}{n} > \right\}. \]

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

\[ |\hat{\lambda}| \geq \sqrt{2(1 + n)\delta_{max}(n) \log(4n^{1+r})}. \]

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

\[ (\hat{P}, \hat{Z}) = \arg \min_{Z \in \{0,1\}^{n \times K}} ||ZX' - \hat{U}||_F. \]

Assume that \( \frac{\alpha_n}{\delta_{max}(n)} \to \infty \) and \( \min \beta_z > 0. \) There exists a positive constant \( C_1 \) such that, for \( n \) large enough, with probability larger than \( 1 - n^{-r}, \)

\[ \hat{K} = K \imply \text{Err}(\hat{Z}, Z) \leq \frac{2C_1 \log(n^{1+r})}{\delta_{max}(n)} \cdot \frac{\max_{\beta_z}}{\alpha_n}. \]

Practical implementation

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

A iterate among nodes whose degree is larger than the median degree

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.

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).

\[ n = 500, K = 5, \delta_{max} = 3, \text{ average over 100 networks } \]

Real-world networks

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

References