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

Friends

Colleagues

Sports team

Emilie Kaufmann (Inria)

Identification of overlapping communities in networks
Overlapping communities: examples

- Co-authorship network

E. Kaufmann (Inria)

Machine learning (NIPS)

T. Bonald

M. Lelarge

Networks (SIGMETRICS)
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)
1. The non-overlapping case
2. The stochastic-blockmodel with overlaps (SBMO)
3. An estimation procedure in the SBMO
4. Theoretical analysis
5. Implementation and results
Outline

1. The non-overlapping case
2. The stochastic-blockmodel with overlaps (SBMO)
3. An estimation procedure in the SBMO
4. Theoretical analysis
5. Implementation and results
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

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

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 \( k : \{1, \ldots, n\} \rightarrow \{1, \ldots, K\} \)
- 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 $p > q$,

$$B = \begin{pmatrix} p & q \\ q & p \end{pmatrix}$$
$A_{i,j} = B_{k(i),k(j)}$

**Observation 1:** $A$ is constant on communities:

$$A_{i,:} = A_{j,:} \iff k(i) = k(j)$$

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

**Observation 2:** this property is preserved for the matrix

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

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

$$U_{i,:} = U_{j,:} \iff k(i) = k(j)$$

(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 $$\hat{U} = [\hat{u}_1 | \ldots | \hat{u}_K] \in \mathbb{R}^{n \times K}$$, matrix of $$K$$ eigenvectors of $$\hat{A}$$ associated to largest eigenvalues

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

**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]
1. The non-overlapping case

2. The stochastic-blockmodel with overlaps (SBMO)

3. An estimation procedure in the SBMO

4. Theoretical analysis

5. Implementation and results
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 \( Z \in \{0, 1\}^{n \times K} \).

\[ Z_i := Z_{i,:} \in \{0, 1\}^{1 \times K} \]: 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**:

  \[
  \text{MisC}(\hat{Z}, Z) = \min_{\sigma \in \mathcal{S}_K} \left| \left\{ i \in \{1, \ldots, n\} : \exists k \in \{1, \ldots, K\}, \hat{Z}_{i,\sigma(k)} \neq Z_{i,k} \right\} \right|
  \]

- **estimation error**:

  \[
  \text{Error}(\hat{Z}, Z) = \frac{1}{nK} \inf_{\sigma \in \mathcal{S}_K} \| \hat{Z} P_\sigma - Z \|_F^2
  \]

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

To perform estimation, the model needs to be identifiable:

\[ Z'B'Z'^T = ZBZ^T \quad \Rightarrow \quad \text{MisC}(Z', Z) = 0. \]

Not always the case! \( ZBZ^T = Z'B'Z'^T = Z''B''Z''^T \), with

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

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

\[ B'' = \begin{pmatrix} a + b - c & b - c & a - c & 0 \\ b - c & b & 0 & 0 \\ a - c & 0 & a & 0 \\ 0 & 0 & 0 & c \end{pmatrix} \quad Z'' = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \]
To perform estimation, the model needs to be identifiable:

\[ Z' B' Z'^T = Z B Z^T \Rightarrow \text{MisC}(Z', Z) = 0. \]

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

1. The non-overlapping case
2. The stochastic-blockmodel with overlaps (SBMO)
3. An estimation procedure in the SBMO
4. Theoretical analysis
5. Implementation and results
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.
A = \( ZBZ^T \) the expected adjacency matrix of an identifiable SBMO:

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

\( U = [u_1 | \ldots | u_K] \) 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 \( \sigma \in \mathcal{G}_k : Z = Z'P_\sigma \)

\((u_1, \ldots, u_K \text{ form a basis of } \text{Im}(A) \text{ and } \text{Im}(A) \subset \text{Im}(Z))\)
This motivates the following estimation procedure:

\[(\mathcal{P}) : (\hat{Z}, \hat{X}) \in \arg\min_{Z', X' \in \mathbb{R}^{K \times K}} ||Z'X' - \hat{U}||_F^2,\]

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

\[(\mathcal{P})' : (\hat{Z}, \hat{X}) \in \arg\min_{Z' \in \{0,1\}^{n \times K} : \forall i, Z'_i \neq 0, X' \in \mathbb{R}^{K \times K}} ||Z'X' - \hat{U}||_F^2.\]

If \(K\) is unknown, let \(\hat{K}\) be the number of eigenvalues \(\lambda\) of \(\hat{A}\) satisfying \(|\lambda| \geq \sqrt{2 \,(1 + \eta) \, \hat{d}_{\max} \,(n) \, \log(4 \, n^{1+r})}\).
1. The non-overlapping case
2. The stochastic-blockmodel with overlaps (SBMO)
3. An estimation procedure in the SBMO
4. Theoretical analysis
5. Implementation and results
Under which conditions is $(\mathcal{P}) : \ (\hat{Z}, \hat{X}) \in \text{argmin}_{Z', X' \in \mathbb{R}^{K \times K}} \|Z'X' - \hat{U}\|_F^2$, a good estimation procedure? 

We present the analysis of a slight variant : 

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

where 

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

for $\epsilon$ smaller than the smallest proportion of pure nodes.
To analyze the solution of \((P_\epsilon)\) when the network grows,

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

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

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

**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 $M_0 := O^{1/2}BO^{1/2}$ associated to $\mu$
2. $u = ZO^{-1/2}x$ 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 $A$

![Diagram showing the spectrum of $A$]

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

![Diagram showing the spectrum of $\hat{A}$]

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

Emilie Kaufmann (Inria)  
Identification of overlapping communities in networks
Step 1: Why is $\hat{U}$ close to $U$?

An adaptive eigenvectors perturbation result

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

$$d_{\text{max}}(n) \geq C_1(\eta) \log \left( \frac{n}{\delta} \right),$$

$$\frac{\lambda_{\text{min}}(A)^2}{d_{\text{max}}(n)} > C_2(\eta) \log \left( \frac{n}{\delta} \right),$$

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

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

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

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

Lemma

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

1. $\forall i \in \mathcal{N}$, $\|Z'_iX' - U_i\| \leq \frac{d_0}{4K\sqrt{n}}$
2. there exists $(i_1, \ldots, i_K) \in \mathcal{N}^K$ and $(j_1, \ldots, j_K) \in \mathcal{N}^K$:

$$\forall k \in [1, K], \quad Z_{i_k} = Z'_{j_k} = 1_{\{k\}}$$

Then there exists a permutation matrix $P_\sigma$ such that

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

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

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

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.

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

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

1. The non-overlapping case
2. The stochastic-blockmodel with overlaps (SBMO)
3. An estimation procedure in the SBMO
4. Theoretical analysis
5. Implementation and results

Emilie Kaufmann (Inria)
Identification of overlapping communities in networks
**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 $(P')$

\[(P') : (\hat{Z}, \hat{X}) \in \arg\min_{Z' \in \{0,1\}^{n \times K} : \forall i, Z'_i \neq 0} \min_{X' \in \mathbb{R}^{K \times K}} ||Z'X' - \hat{U}||_F^2\]

using alternate minimization.

\[||Z'X' - \hat{U}||_F^2 = \sum_{i=1}^{n} ||Z'_iX'_i - \hat{U}_i||^2\]
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 $O_{\text{max}}$.

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

1: $\hat{U}$ Selection of the eigenvectors
2: Form $\hat{U}$ a matrix whose columns are $K$ eigenvectors of $\hat{A}$ associated to eigenvalues $\lambda$ satisfying

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

3: $\hat{Z} = 0 \in \mathbb{R}^{n \times \hat{K}}$
4: $\hat{X} \in \mathbb{R}^{K \times \hat{K}}$ initialized with $k$-means++ applied to $\hat{U}$, the first centroid being chosen at random among nodes with degree smaller than the median degree

6: $\text{Loss} = +\infty$
7: $\text{while (Loss} - \|\hat{Z} \hat{X} - \hat{U}\|_F^2 > \epsilon) \text{ do}$

8: $\text{Update membership vectors: } \forall i, \hat{Z}_{i,:} = \arg\min_{z \in \{0,1\}^{1 \times \hat{K}}} \|\hat{U}_{i,:} - z \hat{X}\|_1$

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

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

11: $\text{end while}$
Experiments on simulated data

CSC versus two spectral algorithms:
- Normalized Spectral Clustering (SC)
- the OCCAM spectral algorithm (OCCAM) [Zhang et al. 14]

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

under SBMO

![Graph showing performance comparison]

under OCCAM

![Graph showing performance comparison]
**Ego-networks** from the ego-networks dataset (SNAP, [Mc Auley, Leskovec 12])

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>K</th>
<th>c</th>
<th>$O_{\text{max}}$</th>
<th>FP</th>
<th>FN</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>190</td>
<td>3.17</td>
<td>1.09</td>
<td>2.17 (0.37)</td>
<td>0.200 (0.110)</td>
<td>0.139 (0.107)</td>
<td>0.120 (0.083)</td>
</tr>
<tr>
<td>OCC.</td>
<td>190</td>
<td>3.17</td>
<td>1.09</td>
<td>2.17 (0.37)</td>
<td>0.176 (0.176)</td>
<td>0.113 (0.084)</td>
<td>0.127 (0.102)</td>
</tr>
<tr>
<td>CSC</td>
<td>190</td>
<td>3.17</td>
<td>1.09</td>
<td>2.17 (0.37)</td>
<td>0.125 (0.067)</td>
<td>0.101 (0.062)</td>
<td>0.102 (0.049)</td>
</tr>
</tbody>
</table>

**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

$C_1 = \{\text{NIPS}\}, \ C_2 = \{\text{ICML}\}, \ C_3 = \{\text{COLT, ALT}\}$

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

<table>
<thead>
<tr>
<th></th>
<th>c</th>
<th>(\hat{c})</th>
<th>FP</th>
<th>FN</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>1.22</td>
<td>1.</td>
<td>0.38</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>OCCAM</td>
<td>1.22</td>
<td>1.02</td>
<td>0.43</td>
<td>0.41</td>
<td>0.42</td>
</tr>
<tr>
<td>CSC</td>
<td>1.22</td>
<td>1.04</td>
<td>0.26</td>
<td>0.28</td>
<td>0.27</td>
</tr>
</tbody>
</table>

$C_1 = \{\text{ICML}\}, \ C_2 = \{\text{COLT, ALT}\}$.

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

<table>
<thead>
<tr>
<th></th>
<th>c</th>
<th>(\hat{c})</th>
<th>FP</th>
<th>FN</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>1.09</td>
<td>1.</td>
<td>0.39</td>
<td>0.55</td>
<td>0.46</td>
</tr>
<tr>
<td>OCCAM</td>
<td>1.09</td>
<td>1.01</td>
<td>0.29</td>
<td>0.44</td>
<td>0.36</td>
</tr>
<tr>
<td>CSC</td>
<td>1.09</td>
<td>1.03</td>
<td>0.21</td>
<td>0.31</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Emilie Kaufmann (Inria) Identification of overlapping communities in networks
Experiments in the sparse case

A simple SBMO: \( A = \alpha_n ZBZ^T \)

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

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

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

![Graph showing the fraction of correct entries in Z vs. fraction of each kind of pure nodes for different values of a = 10, 20, 30, 50.](image)
Experiments in the sparse case

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

\[
a(2 - 3s) > sa
\]

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

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

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

**Conjecture:**

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

- Adding the threshold

![Graph showing the fraction of correct entries in Z vs. s, the fraction of each kind of pure nodes, for different values of a (10, 20, 30, 50).]
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 ($P'$) more efficiently
- are other spectral embeddings possible?
- can the pure nodes assumption be relaxed?


Y. Zhang, E. Levina, J. Zhu, *Detecting Overlapping Communities in Networks with Spectral Methods*, 2014