Clustering of directed graphs: differentiating source and targets nodes or not?

Gilles Celeux\textsuperscript{2} and Yann Vasseur\textsuperscript{1}

\textsuperscript{1}LMO, Université Paris-Saclay

\textsuperscript{2}Inria Saclay-Île-de-France
Biological setting of the case study

*Arabidopsis thaliana (At)*:
Well documented Chlorophyllian plant

*Arabidopsis thaliana* has:
- $\sim 25\,000$ genes

**Case study data**:
- from DNA chips
- $p = 1937$ genes
- $n = 2670$ expression intensities
- $n \sim p$
Biological assumption

Genes of $At$ are supposed to be organized as follows

1. A regulator gene has an effect on the same regularized genes. These genes are called **co-regulators** genes.

2. A group of genes has an effect at the same time on several genes. These genes are called **co-régularized** genes.

**Biological Aim**: make this organisation evident.
Statistical work

1. Designing a regulation network
   → This network is a directed graph.
   ← This graph has been built through penalized regressions using resampling schemes (cf. Vasseur 2017).

2. Clustering of the resulting directed graph
   The adjacency matrix of this graph is $A_{jj'} = 1$ if gene $j$ is regularized by gene $j'$ and 0 otherwise.
   Co-clustering models are especially relevant to highlight the links between groups of regularized genes and groups of regulator genes.
The Latent Block Model (LBM)

It is a co-clustering model of an adjacency matrix $A_{jj'}$ realisation of a random variable $C_{jj'}$.

- the LBM assumes that, knowing the latent categorical variables $v$ and $w$ defining the partitions, the random variables $C_{jj'}$ are independent.
- The conditional density $\phi(A_{jj'}; \alpha_{h\ell})$ of rv $C_{jj'}$ knowing $v_{jh} = 1$ and $w_{j'l} = 1$ is a Bernoulli distribution $\mathcal{B}(\alpha_{h\ell})$ where $\alpha_{h\ell}$ is the link inside block $h\ell$:

$$\phi(A_{jj'}; \alpha_{h\ell}) = \alpha_{h\ell}^{A_{jj'}} \times (1 - \alpha_{h\ell})^{1-A_{jj'}}.$$  

- Thus the marginal density of $A$ is a mixture density

$$p(A; \theta) = \sum_{(v,w) \in V \times W} p(v; \theta)p(w; \theta)p(A|v, w; \theta)$$

$$= \sum_{(v,w) \in V \times W} \prod_{j,h} \rho_v^{v_{jh}} \prod_{j',\ell} \tau_{w_{j'\ell}} \prod_{h,j,j',\ell} \phi(A_{jj'}; \alpha_{h\ell})^{v_{jh}w_{j'\ell}}.$$
The Stochastic Block Model (SBM)

The SBM is a constrained LBM which designs a single partition of the graph nodes. It incidence matrix is

- \( \forall j \in J, z_{jk} = 1 \iff \text{node } j \text{ belongs to } k \).
- \( p(z) = \prod_{j,k} \pi_k^{z_{jk}} \) with \( \pi_k = \mathbb{P}(z_{jk} = 1), k = 1, \ldots, K \).
- Thus the marginal density of \( A \) is the mixture density

\[
p(A; \theta) = \sum_{z \in \mathcal{Z}} \prod_{j,k} \pi_k^{z_{jk}} \prod_{j',k'} \pi_{k'}^{z_{j'k'}} \prod_{k,j,j',k'} \phi(A_{jj'}; \alpha_{kk'})^{z_{jk}z_{j'k'}} , \tag{1}
\]

where \( \pi_k \) is the probability for a node to belong to cluster \( k \).
Estimation and selection of both models

- The parameters to be estimated for the LBM are the proportions $\rho_h$ and $\tau_\ell = 1$ and the Bernoulli block parameters $\alpha_{h\ell}, h = 1, \ldots, H; \ell = 1, \ldots, L$.

- The parameters to be estimated for the SBM are the proportions $\pi_k$ and the Bernoulli block parameters $\alpha_{kk'}, k = 1, \ldots, K; k' = 1, \ldots, K$.

- The likelihoods of models LBM and SBM are not available.

- Using non informative priors, the model parameters are estimated through Gibbs sampling followed by a variational approximation of the posterior modes.

- A model is selected with the ICL (Integrated Completed Likelihood) criterion.

- In the context of binary data, it is possible to compute the exact value of ICL (no need of asymptotic approximation).
Clustering directed graphs: LBM or SBM?

- Co-clustering a directed graph with the LBM is more natural since this model differentiates the source and the target nodes.

- Co-clustering a directed graph with SBM is restrictive since this model enforces source and target nodes to be in the same cluster.

- Inference with the LBM is more complex and numerical difficulties are more important.

- But the difference of complexity between the two models is not important: the number of parameters with a LBM is $H + L + HL$, and $K + K(K + 1)/2$ with a SBM.

- Many people use the SBM for co-clustering of a graph without making a difference between the undirected and directed cases.
The LBM co-clustering

30 "regularized" clusters and 29 "regulator" clusters
The SBM co-clustering

22 clusters
The ICL values for LBM co-clustering

On left the "regularized" clusters, on right the "regulator" clusters

The ICL value for the optimal couple (30, 29) is -668 523.
The best ICL value for 22 blocks is -669 132
Discussion

- For the presented data set, the LBM wins the match: better ICL value and more relevant "small" blocks...

- It could be useful to run and compare the results of both algorithms...

- For the present data set, The CARI of the blocks get with LBM and SBM is 0.14 (not so good...).

- Our opinion is to favor LBM to SBM when concerned with the clustering of a directed graph.
References