Clustering multilayer graphs with missing nodes

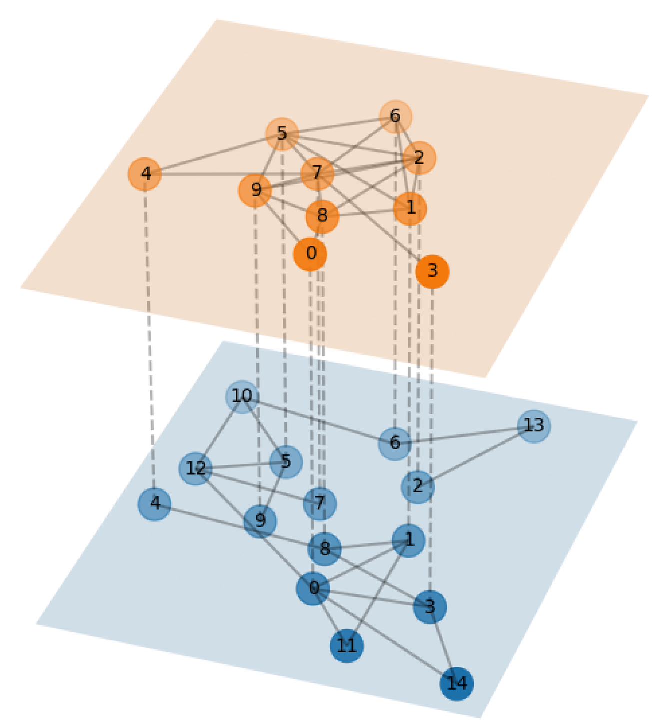
Guillaume Braun Hemant Tyagi Christophe Biernacki

{guillaume.braun, hemant.tyagi,christophe.biernacki}@inria.fr

Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000

Motivations

Clustering multilayer graphs defined on **different sets of nodes**.



$\implies$ Consider absent nodes as **missing**.

The problem setting

- Symmetric **adjacency matrices** $A^{(l)}$ with $A^{(l)}_{ii} = 0$ for each layer l.
- **Mask matrices** $\Omega^{(l)} = (w_i^{(l)} w_j^{(l)})_{i,j \le n}$ where $w_i^{(l)} = 1$, if i is observed on layer l, else 0 .
- **Missing nodes** generation: $w_i^{(l)} \overset{\text{ind.}}{\sim} \mathcal{B}(\rho)$.
- Observed nodes on layer l: $J_l = \{i : w_i^{(l)} = 1\}$.
- $A_{J_l} \in \mathbb{R}^{|J_l| \times |J_l|}$ the submatrix of $A^{(l)}$ restricted to the observed nodes.
- **MLSBM**: a generative model of the multilayer graph with global community structure.

Three clustering strategies

- Cluster each individual layer and then find a consensus partition (**late aggregation method**).
- Aggregate the layer and then apply an unilayer clustering algorithm (**early fusion method**).
- Estimate a low rank subspace common to all layers and perform clustering on this space (**intermediate fusion method**).

Late fusion method

Algorithm: `k-pod`

1. Let $\hat{U}_{J_l} \in \mathbb{R}^{|J_l| \times K}$ be the matrix formed by the top K eigenvalues (in absolute value) of $A_{J_l} \in \mathbb{R}^{|J_l| \times |J_l|}$.
2. Transform $\hat{U}_{J_l}$ to a matrix $\hat{U}^{(l)} \in \mathbb{R}^{n \times K}$ by completing with 0 the rows of missing nodes.
3. Stack the matrices $\hat{U}^{(l)} \implies \hat{U} \in \mathbb{R}^{n \times KL}$.
4. Solve
$$\min_{\substack{Z \in \mathcal{M}_{n,K} \\ C \in \mathbb{R}^{K \times KL}}} ||(\hat{U} - ZC) \odot \Omega_U||_F^2 \qquad (1)$$
where $\Omega_U = (w^{(1)} \otimes \mathbf{1}_K \ \cdots \ w^{(L)} \otimes \mathbf{1}_K)$.
5. Apply k-means on $\hat{Z}$ solution of (1)

Consistency of `k-pod`

Under mild conditions (unbalanced communities, sparsity, ...)

$$\texttt{misclust} \underset{n \to +\infty}{\longrightarrow} 0.$$

Intermediate fusion method

Algorithm: `OLMFm`

1. Find
$$\hat{Q} \in \underset{\substack{Q^T Q = I_k \\ B^{(1)}, \dots, B^{(l)}}}{\operatorname{argmin}} \sum_l ||A_{J_l} - Q_{J_l} B^{(l)} Q_{J_l}^T||_F^2. \qquad (2)$$
where $Q_{J_l} \in \mathbb{R}^{|J_l| \times K}$ is obtained from Q by removing rows corresponding to missing nodes in layer l.
2. Apply k−means on $\hat{Q}$.

Advantage: taking simultaneously the information provided by each allows to cluster sparser graphs than with `k-pod`.

Early fusion methods

Direct aggregation requires **missing values imputation** $\implies$ fill missing entries with **zeros**.

Algorithm: `sumAdj0`

1. Compute $A = L^{-1} \sum_l A^{(l)} \odot \Omega^{(l)}$.
2. Compute $U_k \in \mathbb{R}^{n \times K}$ the matrix formed by the top K eigenvalues of A.
3. Apply K-means on the rows of U_K.

Consistency of `sumAdj0`

Under mild conditions (unbalanced communities, sparsity, ...)

$$\texttt{misclust} \underset{n \to +\infty}{\longrightarrow} 0.$$

⚠ Filling missing nodes with zeros can lead to an **important bias**.

$\implies$ a more clever way to impute these missings values:

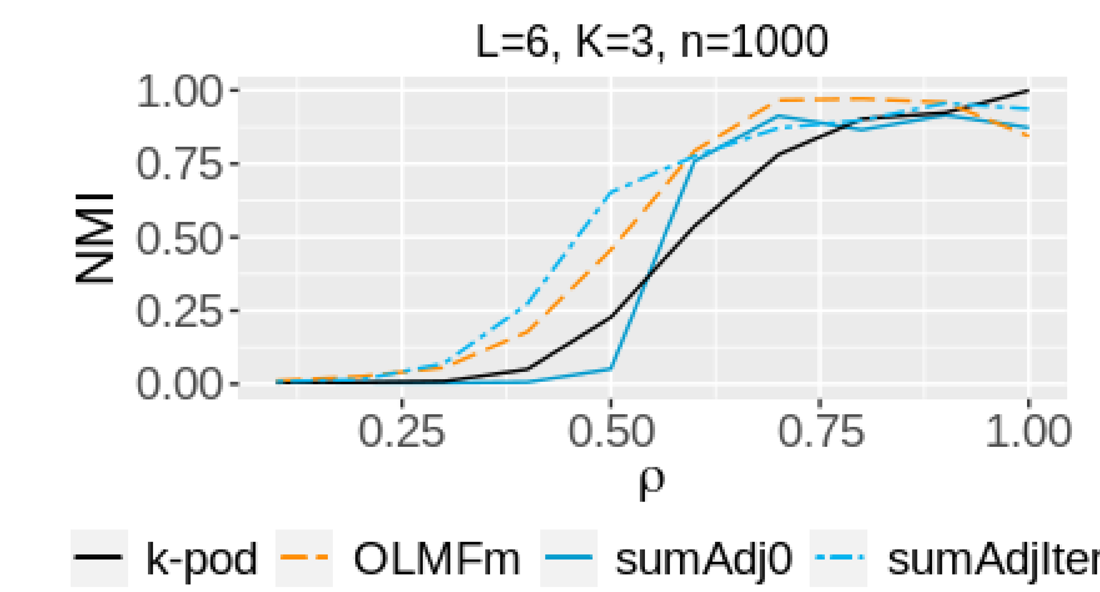
Algorithm: `sumAdjIter`

1. At iteration t, given an initial estimate $\hat{U}_K^t \in \mathbb{R}^{n \times K}$ of the common subspace, estimate the membership matrix $\hat{Z}^t$ by applying k-means on $\hat{U}_K^t$. Then, estimate the connectivity matrix $\hat{\Pi}^{(l),t}$ for each l as
$$\hat{\Pi}^{(l),t} = ((\hat{Z}^t)^T \hat{Z}^t)^{-1} (\hat{Z}^t)^T A^{(l),t} \hat{Z}^t ((\hat{Z}^t)^T \hat{Z}^t)^{-1}.$$
2. Given $\hat{Z}^t$ and $\hat{\Pi}^{(l),t}$ estimate the rows and columns corresponding to missing nodes by computing $\hat{Z}^t \hat{\Pi}^{(l),t} (\hat{Z}^t)^T$.
3. Update the imputed matrices $A^{(l),t+1}$ by replacing the rows and columns of missing nodes by their estimated profiles.
4. Repeat the previous steps using $\hat{U}_K^{t+1}$ and $A^{(l),t+1}$.

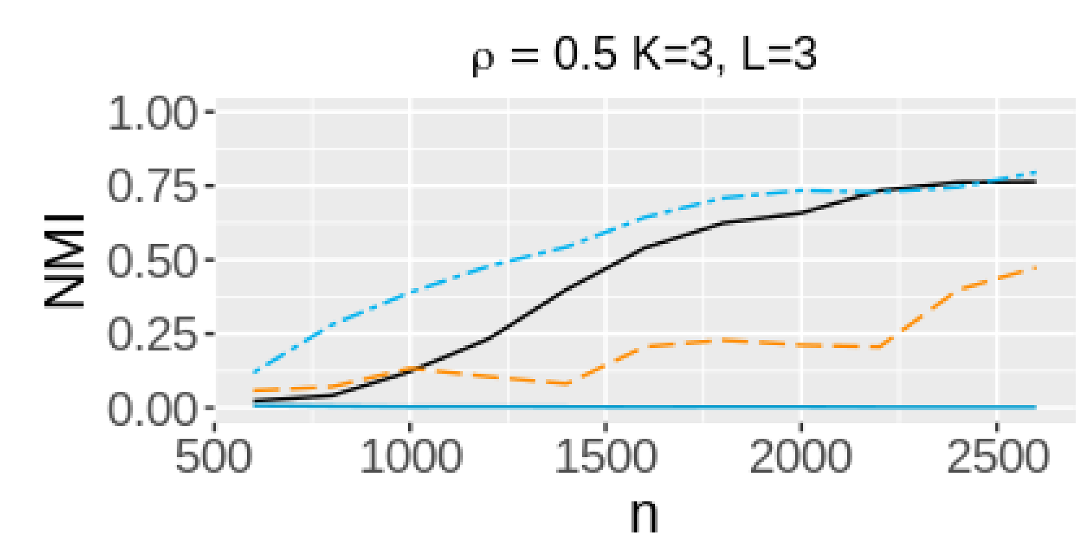
Advantage: best performances in challenging sparsity regimes.

Numerical experiments

- Synthetic data: simulate MLSBM with L layers, n nodes and K communities and delete nodes with probability ρ.
- NMI measures clustering quality and is average over 20 repetitions.
- When ρ is small, early and intermediate fusion methods usually outperform the late aggregation method `k-pod`.



- When the number of node increases, the performance of `sumAdjIter` and `k-pod` improve faster than `OLMFm` and `sumAdj0` that are less sensitive.



- These methods, except for `k-pod`, also works on real datasets such like the MIT Reality Mining dataset with synthetic node deletion.

Conclusion

- We proved consistency of two estimators for clustering multilayer graphs with missing nodes.