Clustering multilayer graphs with missing nodes

Based on a joint work with Hemant Tyagi and Christophe Biernacki

Guillaume Braun

Modal Seminar

24 November 2020, Inria Lille-Nord Europe

Outline

1 Introduction

2 Statistical framework

3 Algorithms and theoretical guarantees

4 Numerical experiments

5 Conclusion

Outline

1 Introduction

2 Statistical framework

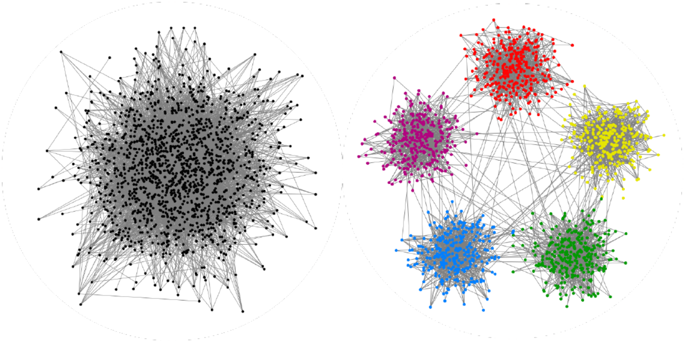
3 Algorithms and theoretical guarantees

4 Numerical experiments

5 Conclusion

Clustering on networks: motivations

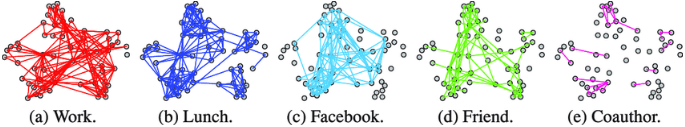
- Network datasets arise naturally in many fields including sociology (social networks) and biology (protein-protein interactions).
- Such networks are complex and not directly analyzable. Clustering is a task that aims to gather nodes having similar connectivity properties.



Incorporate more information with multilayer networks

- Simple networks only consider one kind of relationship between the agents.
- Often relationships are defined on different modalities.
- These multiple aspects of relationships can be represented by a multilayer graph.

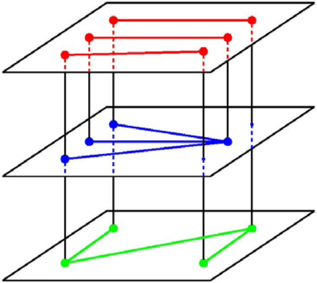
Example: the AUCS dataset



(a) Work. (b) Lunch. (c) Facebook. (d) Friend. (e) Coauthor.

What happens when the layers don't share the same set of nodes ?

Problem: existing clustering methods for multilayer graphs don't work in this case.



→ Consider the nodes that don't appear on some layers as missing.

⚠ Nodes are missing if none of their connections with other nodes have been observed. It is different from having no connections with other nodes.

Outline

1 Introduction

2 Statistical framework

3 Algorithms and theoretical guarantees

4 Numerical experiments

5 Conclusion

Multilayer graphs with missing nodes

- A (pillar) multilayer graph is a sequence of graphs $\mathcal{G} = (\mathcal{G}^{(1)}, \dots, \mathcal{G}^{(L)})$ on the same set of nodes $[n]$.
- Each graph $\mathcal{G}^{(l)}$ is undirected and has no self-loop $\Leftrightarrow$ its associated adjacency matrix $A^{(l)} \in \{0,1\}^{n \times n}$ is symmetric with $A_{ii}^{(l)} = 0$.
- When a node i is observed on layer l, $w_i^{(l)} = 1$ and 0 else.
- Missing nodes MCAR generating process: $w_i^{(l)} \overset{\text{ind.}}{\sim} \mathcal{B}(\rho)$ for $0 < \rho \leq 1$.
- Set of observed nodes on layer l: $J_l = \{i : w_i^{(l)} = 1\}$.
- Mask matrix $\Omega^{(l)} = (w_i^{(l)} w_j^{(l)})_{i,j \leq n}$.

Stochastic Block Model (SBM)

• The SBM (see Abbe (2018)) is a popular model for random graph with a community structure. It is parametrized by:

- the number of nodes n;
- the number of communities K;
- a membership matrix $Z \in \{0,1\}^{n \times K}$ such that $Z_{ik} = 1$ if node i belongs to $\mathcal{C}_k$, 0 otherwise;
- a full-rank symmetric connectivity matrix of probabilities

$$\Pi = (\pi_{kk'})_{k,k' \in [K]} \in [0,1]^{K \times K}.$$

• A graph $\mathcal{G}$ is distributed according to a stochastic block model SBM(Z, n, K, Π) if the adjacency matrix A associated to $\mathcal{G}$ has zero diagonal entries and

$$A_{ij} | (Z_{ik} = 1, Z_{ik'} = 1) \overset{\text{ind.}}{\sim} \mathcal{B}(\pi_{kk'}), \quad 1 \leq i < j \leq n.$$

• The sparsity of such graphs is measured by $p_{max} = \max_{ij} p_{ij}$. The sizes of the largest and smallest community are denoted by n_{max} and n_{min}.

Multi-Layer SBM (MLSBM)

- The MLSBM is an extension of SBM to the multilayer setting (Paul and Chen (2020)). It is parametrized by the number of layers L, a common block membership matrix $Z \in \mathcal{M}_{n,K}$, and connectivity matrices $\Pi^{(1)}, \ldots, \Pi^{(L)} \in [0,1]^{K \times K}$.

- Each layer $\mathcal{G}^{(l)} \overset{\text{ind.}}{\sim} SBM(Z, n, K, \Pi^{(l)})$.

- The sparsity of each layer is denoted $p_{max}^{(l)}$ and $p_{max} := \max_l p_{max}^{(l)}$.

Outline

1 Introduction

2 Statistical framework

3 Algorithms and theoretical guarantees

4 Numerical experiments

5 Conclusion

Three different strategies

Several algorithms have been proposed to cluster multilayer graphs in the complete setting. They can roughly be categorized as follows.

- Late fusion methods: Each layer is analyzed individually and the final partition is obtained by aggregating these individual results.
- Early fusion methods: Clustering is performed on a aggregation of all layers.
- Intermediate fusion methods: Clustering is applied on a factor common to all layers.

Our contributions:

- provide new algorithms for the missing nodes setting based on these strategies;
- show consistency of two of them;
- numerical comparisons between algorithms.

Late fusion method 1/3

Input: The number of communities K, the sets J_l and the adjacency matrices $A_{J_l} \in \mathbb{R}^{|J_l| \times |J_l|}$.

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. The matrix $\hat{U}_{J_l}$ can be transformed to a matrix $\hat{U}^{(l)}$ of size $n \times K$ by completing with 0 the rows of the nodes that haven't been observed.
3. Let $\hat{U} \in \mathbb{R}^{n \times KL}$ matrix obtained by stacking $\hat{U}^{(l)}$.
4. Solve
$$\min_{\substack{Z \in \mathcal{M}_{n,K} \\ C \in \mathbb{R}^{K \times KL}}} ||(\hat{U} - ZC) \odot \Omega_{\mathrm{U}}\,||_F^2 \tag{1}$$
where $\Omega_{\mathrm{U}} = (w^{(1)} \otimes \mathbf{1}_K \ \cdots \ w^{(L)} \otimes \mathbf{1}_K)$.
5. Apply k-means on $\hat{Z}$ solution of (1)

This algorithm will be referred to as `k-pod` .

Late fusion method 2/3

The optimization problem (1) is NP-hard and we used the following heuristic borrowed from Chi et al. (2015).

1. Initialize randomly the partition $\hat{Z}$ and the centroid matrix $\hat{C}$.
2. Replace $\hat{U}$ by $\hat{U} \odot \Omega_{\text{U}} + (\hat{Z}\hat{C}) \odot (11^T - \Omega_{\text{U}})$.
3. Apply K-means on the complete matrix $\hat{U}$ and update $\hat{C}$ and $\hat{Z}$.
4. Repeat the previous two steps until convergence.

Late fusion method 3/3

Theorem

Consider the missing nodes MLSBM and suppose that $\rho L \geq 1$, $KL \leq Cn$, $\rho n_{min} \geq C_1 K^2 \max(\log^2 n, \sqrt{np_{max}})$ and $np_{max}^{(l)} \geq c\rho^{-1} \log n$ for a constant $C_1 > 0$ large enough. Let $\lambda_K^{(l)}$ be the K-th largest singular value of $\Pi^{(l)}$ and recall that $\beta = n_{max}/n_{min}$. There exists a value $c_1(\beta, K) > 0$ depending on β and K such that if

$$\sum_l \frac{np_{max}^{(l)}}{\rho L(n_{min}\lambda_K^{(l)})^2} \leq c_1(\beta, K)$$

then there exists $n_0(\rho, K)$ such that for all $n \geq n_0(\rho, K)$, with probability at least $1 - O(n^{-1})$, it holds that the solution $\hat{Z} \in \mathcal{M}_{n,K}$ of the optimization problem (1) satisfies

$$r(\hat{Z}, Z) \leq C\exp(-c'\rho L) + \frac{C(\beta, K)}{\rho L} \sum_l \frac{p_{max}^{(l)}}{n_{min}(\lambda_K^{(l)})^2},$$

where

$$r(\hat{Z}, Z) = r(\hat{z}, z) = \frac{1}{n} \min_{\sigma \in \mathfrak{S}_K} \sum_i 1_{\{\hat{z}(i) \neq \sigma(z(i))\}}.$$

Proof sketch

- Let $\mathcal{N}_u = \{i : |L_i| \geq \rho L/c\}$ where $c > 1$ is a constant. We have $|\mathcal{N}_u^c| = O(n \exp(-c'\rho L)$ for $c' > 0$.
- Let's define the set of 'bad nodes' as

$$\mathcal{S}_k := \{i \in \mathcal{C}_k \cap \mathcal{N}_u : \forall l \in L_i, ||U_{i*}^{(l)} O_l - \bar{U}_{i*}^{(l)}|| \geq \delta_k^{(l)}/2\}$$

where $\delta_k^{(l)}$ is the smallest distance between two rows of $U^{(l)}$ corresponding to different communities and $O_l \in \mathbb{R}^{K \times K}$ is an orthogonal matrix. We can show that the nodes belonging to $\mathcal{T}_k := (\mathcal{C}_k \setminus \mathcal{S}_k) \cap \mathcal{N}_u$ are well clustered.

- We can bound the sizes of the $\mathcal{S}_k$ by using similar ideas as in Lei and Rinaldo (2015).
  - Bound $\sum_k |\mathcal{S}_k| \delta_k^2$ by $\frac{2C}{\rho L} ||(\hat{U} - U') \odot \Omega_U||_F^2$ where $\delta_k = \min_l \delta_k^{(l)}$ and U' is obtained by stacking the matrices $U^{(l)} O_l$
  - Notice that $||(\hat{U} - U') \odot \Omega_U||_F^2 = \sum_l ||\hat{U}_{J_l} - U_{J_l} O_l||_F^2$ and use Weding's bounds + concentration inequalities.
- **Conclusion.** The misclustering rate $r(\hat{Z}, Z)$ is bounded by $\frac{|\mathcal{N}_u^c| + \sum_k |\mathcal{S}_k|}{n}$.

Early fusion methods 1/4

Main drawback of late fusion methods : rely heavily on the quality of each layer.

→ Alternative: first aggregate the data then apply a clustering method. A popular way to aggregate the layer is to take the mean of adjacency matrices.

Problem: the sum is not defined when there are missing values.

→ We can impute them.

Early fusion methods 2/4

Input: The number of communities K, the matrices $A^{(l)}$ and $\Omega^{(l)}$.

1. Compute $A = L^{-1} \sum_l A^{(l)} \odot \Omega^{(l)}$.
2. Compute the eigenvectors $u_1, \ldots, u_K$ associated with the K largest eigenvalues of A (ordered in absolute values) and form $U_K = [u_1 \; u_2 \; \cdots \; u_K]$.
3. Apply K-means on the rows of U_K to obtain a partition of $\mathcal{N}$ into K communities.

Output: A partition of the nodes $\mathcal{N} = \cup_{i=1}^{K} \mathcal{C}_i$.

This algorithm will be referred to as `sumAdj0` .

Early fusion methods 3/4

Theorem

Under the missing nodes MLSBM , there exist constants $c, C > 0$ such that with probability at least $1 - O(n^{-1})$, the solution $\hat{Z} \in \mathcal{M}_{n,K}$ obtained from the previously described algorithm satisfies

$$r(\hat{Z}, Z) \leq \underbrace{\frac{c}{\rho^2 \lambda_K}\left(\sqrt{\frac{np_{max}}{L}} + \sqrt{\frac{\log n}{L}}\right)}_{\textit{noise error}} +$$

$$\underbrace{C\frac{(\rho^{-2}-1)}{\lambda_K}\left(np_{max}\sqrt{\frac{\log(n)}{L}} + \frac{np_{max}\log n}{L}\right)}_{\textit{missing data error}}$$

Proof sketch

The proof is obtained in three steps. Let's denote $\tilde{A} = \rho^{-2} L^{-1} \sum_l A^{(l)} \odot \Omega^{(l)}$.

- First we show by using concentration inequality from Bandeira and van Handel (2016) that w.h.p.

$$||\tilde{A} - \mathbb{E}(\tilde{A}|\Omega)|| \leq c_1 \rho^{-2} \left(\sqrt{\frac{np_{max}}{L}} + \sqrt{\frac{\log(n)}{L}} \right)$$

- Then we use Bernstein's inequality to show that w.h.p.

$$||\mathbb{E}(\tilde{A}|\Omega) - \mathbb{E}(\tilde{A})|| \leq c_2(\rho^{-2} - 1) \left[np_{max} \sqrt{\frac{\log(n)}{L}} + \frac{\log(n) np_{max}}{L} \right]$$

- Conclude by using Weding's bound.

Early fusion methods 4/4

• Imputing missing values with zeros $\implies$ bias. Could we reduce this bias by using a different imputation method ?

1. At iteration t, given an initial estimate $\hat{U}_K^t \in \mathbb{R}^{n \times K}$ of the common subspace we can estimate the membership matrix $\hat{Z}^t$ by applying k-means on $\hat{U}_K^t$. Then, we can 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 $\hat{\Pi}^{(l),t}$ we estimate the rows and columns corresponding to missing nodes by computing $\hat{Z}^t \hat{\Pi}^{(l),t} (\hat{Z}^t)^T$.
3. We obtain the updated imputed matrix $A^{(l),t+1}$ by replacing the rows and columns of missing nodes by their estimated profiles.

Repeat te previous steps using $\hat{U}_K^{t+1}$ and $A^{(l),t+1}$.

This algorithm will be referred to as `sumAdjIter`

Intermediate fusion method: OLMF in the complete case

Main drawback of sum of adjacency matrix : we can loose information when some layers are dissortatives and other assortatives.

➔ Alternative: try to extract a factor common to each layer.

• The orthogonal linked matrix factorization (OLMF)(Paul and Chen (2020)) estimator is a solution of the following optimization problem

$$(\hat{Q}, \hat{B}^{(1)}, \dots, \hat{B}^{(L)}) \in \underset{\substack{Q^T Q = I_k \\ B^{(1)}, \dots, B^{(L)}}}{\operatorname{argmin}} \sum_{l} ||A^{(l)} - QB^{(l)}Q^T||_F^2, \tag{2}$$

where $Q \in \mathbb{R}^{n \times K}$, $B^{(l)} \in \mathbb{R}^{K \times K}$. The community estimation is then obtained by applying K-means on the rows of $\hat{Q}$.

Adaptation of OLMF to the missing case

- Replace the matrices $A^{(l)}$, Q in the objective function in (2) with $A_{J_l} \in \mathbb{R}^{n \times n}$, $Q_{J_l} \in \mathbb{R}^{n \times K}$ (entries corresponding to missing nodes are filled with zeros), and solve:

$$(\hat{Q}, \hat{B}^{(1)}, \ldots, \hat{B}^{(L)}) \in \underset{\substack{Q^T Q = I_k \\ B^{(1)}, \ldots, B^{(l)}}}{\operatorname{argmin}} \sum_{l} ||A_{J_l} - Q_{J_l} B^{(l)} Q_{J_l}^T||_F^2. \tag{3}$$

In our experiments, we employ a BFGS algorithm for solving (3). This algorithm will be referred to as `OLMFm` .

Outline

1 Introduction

2 Statistical framework

3 Algorithms and theoretical guarantees

4 Numerical experiments

5 Conclusion

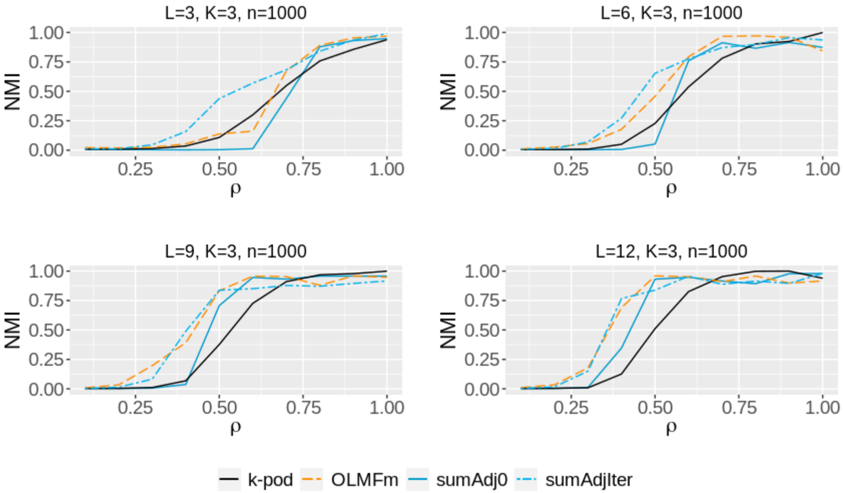
Experimental design

The multilayer graphs are generated from a missing MLSBM where the parameters n, L and ρ are varying and :

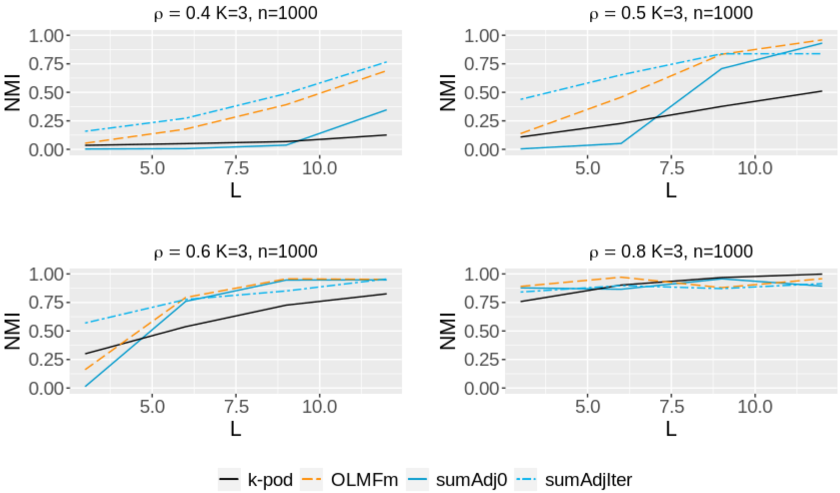
- the diagonal (resp. off-diagonal) entries of the connectivity matrices are generated uniformly at random over $[0.18, 0.19]$ (resp. $0.7 * [0.18, 0.19]$);
- the ground truth partition is generated from a multinomial law with parameters $1/K$ and $K = 3$.

The normalized mutual information (NMI) criterion is used to compare the estimated community to the ground truth partition.

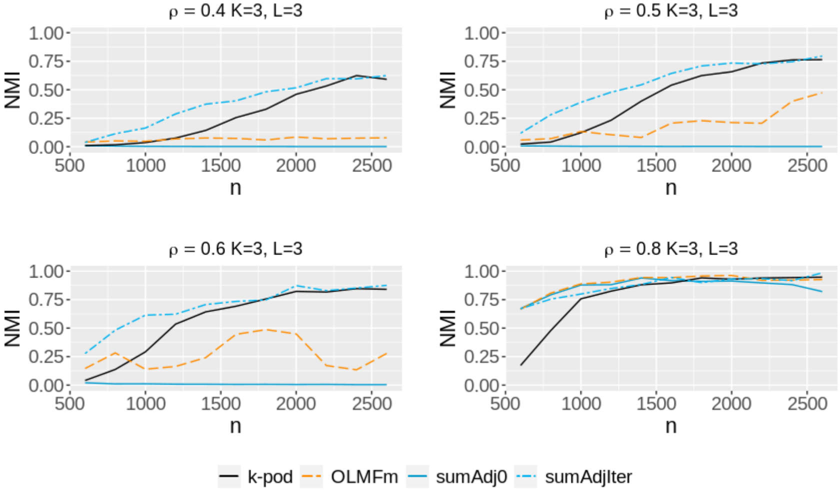
Numerical experiments with varying ρ



Numerical experiments with varying L



Numerical experiments with varying n



MIT Reality Mining dataset

This dataset records interactions (measured by cell phones activities) between 96 students and staff at MIT over one year. We discretized the time into one week intervals. Nodes are randomly removed from each layer.

ρ	sumAdj0	OLMFm	sumAdjIter
1	1.00	1.00	1.00
0.9	0.99	0.96	0.99
0.8	0.97	0.86	0.97
0.7	0.96	0.93	0.96
0.6	0.94	0.79	0.94
0.5	0.89	0.91	0.90
0.4	0.76	0.73	0.78
0.3	0.56	0.57	0.62
0.2	0.26	0.41	0.36
0.1	0.09	0.10	0.11

Outline

1 Introduction

2 Statistical framework

3 Algorithms and theoretical guarantees

4 Numerical experiments

5 Conclusion

Conclusion and perspectives

- We proved consistency of two estimators for clustering multilayer graphs with missing nodes. Nevertheless as shown in the experiments these estimators are unlikely to be optimal.
- The assumption that all the layer share exactly the same structure is strong and it would be interesting to relax it, in particular in the context of dynamic graphs.
- Other generating processes for missing nodes could be considered.
- Model-based approaches with variational methods or Stochastic EM could also be considered.

References

Abbe, E. (2018). Community detection and stochastic block models. *Foundations and Trends® in Communications and Information Theory*, 14(1-2):1–162.

Bandeira, A. S. and van Handel, R. (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.*, 44(4):2479–2506.

Chi, J., Chi, E., and Baraniuk, R. (2015). k -pod a method for k -means clustering of missing data. *The American Statistician*, 70:1–29.

Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237.

Paul, S. and Chen, Y. (2020). Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *Ann. Statist.*, 48(1):230–250.