



HAL
open science

Clustering on multilayer graphs with missing values

Guillaume Braun, Christophe Biernacki, Hemant Tyagi

► **To cite this version:**

Guillaume Braun, Christophe Biernacki, Hemant Tyagi. Clustering on multilayer graphs with missing values. Journée de Statistique de la SFdS, May 2020, Nice, France. hal-03122104

HAL Id: hal-03122104

<https://inria.hal.science/hal-03122104>

Submitted on 26 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CLUSTERING ON MULTILAYER GRAPHS WITH MISSING VALUES

Guillaume Braun ¹ & Christophe Biernacki ² & Hemant Tyagi ³

¹ *Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé
59650 Villeneuve d'Ascq, France, guillaume.braun@inria.fr*

² *Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé
59650 Villeneuve d'Ascq, France, christophe.biernacki@inria.fr*

³ *Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé
59650 Villeneuve d'Ascq, France, hemant.tyagi@inria.fr*

Résumé. La classification sur les graphes à couches multiples rencontre un intérêt croissant depuis une décennie en raison de leurs nombreux champs d'applications. Différentes méthodes ont été proposées, mais elles reposent toutes sur l'hypothèse que la totalité des interactions entre les noeuds du réseau sont observées. Nous proposons un cadre statistique afin d'étudier le cas souvent plus réaliste dans lequel des noeuds ne sont pas observés sur certaines couches. Une méthode spécifique permettant d'estimer les paramètres du modèle et d'imputer les valeurs d'interactions manquantes est également proposée.

Mots-clés. Graphes à couches multiples, Classification, Stochastic Block Model, données manquantes.

Abstract. Multilayer graphs clustering have gained increasing interest this last decade due to numerous applications in various fields. Several clustering methods have been proposed, but they rely all on the assumption that the network is fully observed. We propose a statistical framework to handle nodes that are missing on some layers as well as a method to estimate the model parameters and to impute missing edge values.

Keywords. Multilayer graph, Clustering, Stochastic Block Model, missing data.

1 Introduction

Simple graphs are often used to model relationships between different agents: each agent is associated to a node and the link between two agents is represented by an edge between the corresponding nodes. Weighted and directed edges can be used to express the intensity and the direction of a link.

However, relationships between agents can have multiple aspects that are better represented by multilayer graphs where each layer corresponds to one aspect of the relationship. For example, the social network of a person involves different relationship types such as emails exchanges, telephone calls, offline personal interactions, professional links and so

on. All these layers are dependent to some extent since a modification in one layer can impact the others. For example a change of employer could have repercussions on both online and offline networks. A layer might also correspond to a snapshot of the entire network at a sampled time instant. These networks are sometimes referred to as temporal or dynamic networks to emphasise that they are time dependent.

Because of their broad field of applications including Biology, Sociology, Genetics and Ecology, multilayer graph analysis has gained increasing attention in the last decade, especially for clustering purposes (see [Kim and Lee, 2015] for a survey). Although various clustering methods have been proposed, they rely on the common assumption that the network is fully observed. However missing values often occur in practice. For instance, various missing schemes for edges have been already studied for unilayer graphs in [Tabouy et al., 2019].

We propose a new sampling design to deal with missing nodes in some layers. Until now only missing edges have been considered in the literature. In the unilayer case, since there is no information about missing nodes, it is impossible to make an inference concerning their links with the other nodes. But in the multilayer case, a missing node in one layer can be present in other layers and the additional information provided by the layers where the node is observed can be used. Moreover, efficiently clustering missing nodes in some layer would have interesting applications in link prediction.

Our originality is to propose some algorithms for estimating the model parameters and the latent partition when some nodes are missing between layers. Our proposal fully relies on the existing Multi-Layer Stochastic Block Model [Lei et al., 2019] embedded in an explicit missing data mechanism.

In Section 2 we introduce notations, present the classical Stochastic Block Model for unilayer graphs, its extension to the multilayer case and also some non-response mechanisms. Section 3 describes the estimation procedure based on approximations of the maximum likelihood estimator. Finally, Section 4 exposes our future work directions which include ongoing numerical experiments.

2 A model for multilayer networks with missing nodes

2.1 Multilayer networks notations

A network can be recorded by an adjacency matrix. Thus a multilayer network can be represented by a collection of L adjacency matrices: each layer $l = 1, \dots, L$ is associated to an adjacency matrix $Y^l = (Y_{ij}^l)_{1 \leq i, j \leq n}$ and the whole multilayer network is denoted by $Y = (Y^l)_{1 \leq l \leq L}$. A multilayer graph is said to be *pillar*, if the set of nodes $\mathcal{N} = \{1, \dots, n\}$ is the same in each layer.

We restrict our study to undirected graphs that don't have loops. This implies that all the adjacency matrices are symmetric and have zero diagonal.

Let $R_{ij}^l = 1$ if the edge between the nodes i and j in the layer l is observed and 0 otherwise. We denote $\mathcal{D}^{o,l} = \{(i, j) \text{ such that } R_{ij}^l = 1\}$ to be the set of observed edges in layer l , $Y^{o,l} = (Y_{ij}^l)_{(i,j) \in \mathcal{D}^{o,l}}$ to be the corresponding values of the observed edges in layer l and $Y^o = (Y^{o,l})_{1 \leq l \leq L}$. Symmetrically, we denote $\mathcal{D}^{m,l} = \{(i, j) \text{ such that } R_{ij}^l = 0\}$ to be the set of missing edges in layer l , $Y^{m,l} = (Y_{ij}^l)_{(i,j) \in \mathcal{D}^{m,l}}$ to be the corresponding values of the observed edges in layer l and $Y^m = (Y^{m,l})_{1 \leq l \leq L}$. We say that a node i is not observed in the layer l if any of the edges between i and the other nodes are unobserved in this layer. We denote $\mathcal{N}^{o,l}$ to be the corresponding set of observed nodes. The set of nodes that appear in at least one layer is denoted by $\mathcal{N}^o = \cup_l \mathcal{N}^{o,l}$.

2.2 The Multi-Layer Stochastic Block Model (ML-SBM)

The ML-SBM is an extension to multilayer networks of the Stochastic Block Model (SBM) devoted to unilayer networks. In the ML-SBM each layer is generated from a SBM, with eventually different connectivity parameters, and every node has the same block membership in every layer. A first introduction of this model can be founded in [Lei et al., 2019].

Let $\mathcal{K} = \{1, \dots, K\}$ the different block (or cluster) numbers. The block-membership of the node i is encoded by Z_i , and $(Z_i)_{i \in \mathcal{N}}$ are i.i.d. random variables distributed over \mathcal{K} , independently of the layer number. The whole partition is now denoted by $Z = (Z_i)_{i \in \mathcal{N}}$. The distribution of Z_i is completely determined by $\mathbb{P}(Z_i = k) = \alpha_k$ for $k = 1, \dots, K$. It is sometimes convenient to associate Z_i to a binary vector (Z_{i1}, \dots, Z_{iK}) where $Z_{ik} = 1$ if $Z_i = k$ and $Z_{ik'} = 0$ for $k' \neq k$. The realization of the random variable Z_i is denoted in lowercase by z_i , and the whole mixing proportion parameter is denoted by $\alpha = (\alpha_1, \dots, \alpha_K)$.

Considering now the adjacency matrices Y^l ($l = 1, \dots, L$), each adjacency matrix is symmetric ($Y_{ij} = Y_{ji}$ for all i and j), and the conditional distribution of edges depends on the layer. The latter is given by

$$Y_{ij}^l | (Z_i = k, Z_j = k') \stackrel{\text{i.i.d.}}{\sim} \mathcal{B}(\pi_{kk'}^l), \forall i < j$$

where $\mathcal{B}(p)$ denotes a Bernoulli distribution of parameter p and $\pi_{kk'}^l$ is the probability for node in community k to be linked with a node in community k' . Let us finally denote $\pi^l = (\pi_{kk'}^l)_{1 \leq k, k' \leq K}$ to be a symmetric matrix of probabilities and $\theta = (\alpha, \pi^1, \dots, \pi^L)$ to be the mixture model parameters.

2.3 Missing values

The mechanisms for missingness can be traditionally classified as MCAR (missing completely at random), MAR (missing at random) or MNAR (missing not at random). We assume that nodes are MAR and that the missingness mechanism is ignorable. Such a

classical assumption avoids defining (and then estimating) the related distribution of the R_{ij}^l 's since it will have no consequence on the estimation of the parameter of interest, namely θ , in the clustering context [Little and Rubin, 2002].

3 Maximum likelihood estimation (MLE)

3.1 Observed and complete log-likelihood

The complete data log-likelihood of the model is given by

$$\mathcal{L}_c(\theta; z, Y) = \sum_i \log(\alpha_{z_i}) + \sum_{\substack{i,j,l \\ i < j}} (Y_{ij}^l \log \pi_{z_i, z_j}^l + (1 - Y_{ij}^l) \log(1 - \pi_{z_i, z_j}^l)).$$

However the block membership is usually not observed and some edge values are also missing. So we rather use the observed log-likelihood $\mathcal{L}(\theta; Y^o)$ obtained by integration over the latent variables Z and the missing edges Y^m , given by

$$\mathcal{L}(\theta; Y^o) = \log \left(\sum_z \int_{Y^m} \exp(\mathcal{L}_c(\theta; z, Y)) dY^m \right).$$

The associated MLE is computationally intractable in general, even with a classical EM algorithm [Dempster et al., 1977]. But there exists approximate EM algorithms and we explore one of them below. Once we have an estimate $\hat{\theta}$ for θ , the partition and the missing edges can be respectively estimated by

$$\hat{Z} = \arg \max_z \mathbb{P}_{\hat{\theta}}(z|Y^o) \quad \text{and} \quad \hat{Y}^m = \arg \max_{Y^m} \mathbb{P}_{\hat{\theta}}(Y^m|Y^o). \quad (1)$$

3.2 Variational EM (VEM)

The VEM algorithm has been developed in [Daudin et al., 2008] for avoiding computational difficulties implied by the EM algorithm when estimating the SBM. Instead of trying to maximize $\mathcal{L}(\theta; Y^o)$, VEM intends to maximize a lower bound $J_{\theta, \tau}(Y^o) := \mathcal{L}(\theta; Y^o) - KL(\mathbb{P}_{\tau}(Z) || \mathbb{P}_{\theta}(Z|Y^o)) = \mathbb{E}_{\mathbb{P}_{\tau}(Z)}(\log(\mathbb{P}_{\theta}(Z, Y^o)))$ of this quantity. If we consider the set of all possible distributions \mathbb{P}_{τ} for Z , the maximum is attained when $\mathbb{P}_{\tau}(Z) = \mathbb{P}_{\theta}(Z|Y^o)$ but this last quantity is difficult to compute. So we restrict the set in which \mathbb{P}_{τ} belongs to. We are only looking on probability that can be factorized such that $\mathbb{P}_{\tau}(Z) = \prod_i \mathcal{M}_{\tau_i}(Z_i)$ where \mathcal{M}_{τ_i} is the multinomial distribution with parameters $\tau_i = (\tau_{i1}, \dots, \tau_{iK})$ and $\tau = (\tau_1, \dots, \tau_n)$.

After having chosen an initial parameter $\theta^{(0)}$, we construct iteratively a sequence $(\theta_{t \geq 1}^{(t)})$ by repeating the following steps until $\|\theta^{(t+1)} - \theta^{(t)}\| < \epsilon$, where ϵ is a parameter fixed by the user.

- **VE step:** compute $\tau^{(t)} := \arg \max_{\tau} J_{\theta^{(t)}, \tau}(Y^o)$;
- **M step:** update $\theta^{(t)}$ by $\theta^{(t+1)} := \arg \max_{\theta} J_{\theta, \tau^{(t)}}(Y^o)$.

These two maximization problems are solved straightforwardly:

1. The variational parameters $\tau^{(t)}$ maximizing $J_{\theta^{(t)}, \tau}(Y^o)$ when $\theta^{(t)}$ is held fixed are obtained with the following fixed point relation:

$$\tau_{ik}^{(t)} \propto \alpha_k^{(t)} \prod_{\substack{l \leq L \\ (i,j) \in \mathcal{D}^{o,l} \\ i < j}} \prod_{k'=1}^K b(Y_{ij}^l; \pi_{kk'}^{l(t)})^{\tau_{jk'}^{(t)}},$$

where $b(y, \pi) = \pi^y (1 - \pi)^{1-y}$.

2. The parameters $\theta^{(t+1)}$ maximizing $J_{\theta, \tau^{(t)}}(Y^o)$ when $\tau^{(t)}$ is held fixed are:

$$\alpha_k^{(t+1)} = \frac{\sum_{i \in \mathcal{N}^o} \tau_{ik}^{(t)}}{|\mathcal{N}^o|}, \quad \pi_{kk'}^{l(t)} = \frac{\sum_{\substack{(i,j) \in \mathcal{D}^{o,l} \\ i < j}} \tau_{ik}^{(t)} \tau_{jk'}^{(t)} Y_{ij}^l}{\sum_{\substack{(i,j) \in \mathcal{D}^{o,l} \\ i < j}} \tau_{ik}^{(t)} \tau_{jk'}^{(t)}}.$$

3.3 Missing value estimation: partition and edges

The VEM algorithm provides an estimation of the ML-SBM parameters that can be used to solve the maximization problems (1) in order to get an estimator for the partition and missing edges. Unfortunately, these optimization problems are computationally intractable. We propose an approximated solution based on Gibbs sampling. Let's describe the algorithm.

1. **Initialisation.** Let $z^{(0)} = (z_1^{(0)}, \dots, z_n^{(0)})$ and $Y^{m(0)}$ random initial choices for the partition and missing values.
2. **Sampling partition.** Generate the partition with:

$$\mathbb{P}_{\hat{\theta}}(z_i^{(t)} = k | Y^o, Y^{m, (t-1)}, z^{-i, (t-1)}) \propto \hat{\alpha}_k \times \prod_{l,j,k'} b(y_{ij}^l; \hat{\pi}_{kk'}^l)^{z_{jk'}^{(t-1)}}$$

where $z^{-i, (t-1)}$ correspond to $z^{(t-1)}$ with the i th component removed.

3. **Missing values imputation.** Generate the missing data $Y^{l^{m, (t)}} = (y_{ij}^{l(t)})_{\substack{l \in [L] \\ i, j \in Y^m \\ i < j}}$

according to:

$$\mathbb{P}_{\hat{\theta}}(y_{ij}^{l(t)} | Y^o, z^{(t)}) = \prod_{k,k'} b(y_{ij}^{l(t)}; \hat{\pi}_{kk'}^l)^{z_{ik}^{(t)} z_{jk'}^{(t)}}.$$

Gibbs algorithm is iterated several times. The samples obtained during the burn-in period are disregarded. Finally, the final partition and missing observations are estimated using the mode of their marginal sampled distribution.

4 Directions for future works

We intend to perform numerical experiments on simulated and real data. We are also aiming to extend the proposed model and develop a clustering algorithm for multilayer graphs based on spectral methods because these methods have good performance on unilayer graphs and are more scalable than those based on maximum likelihood approximations.

References

- [Daudin et al., 2008] Daudin, J. J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183.
- [Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.
- [Kim and Lee, 2015] Kim, J. and Lee, J.-G. (2015). Community detection in multi-layer graphs: A survey. *ACM SIGMOD Record*, 44:37–48.
- [Lei et al., 2019] Lei, J., Chen, K., and Lynch, B. (2019). Consistent community detection in multi-layer network data. *Biometrika*, 107(1):61–73.
- [Little and Rubin, 2002] Little, R. and Rubin, D. (2002). *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley.
- [Tabouy et al., 2019] Tabouy, T., Barbillon, P., and Chiquet, J. (2019). Variational inference for stochastic block models from sampled data. *Journal of the American Statistical Association*, pages 1–20.