Stochastic block model for multiple networks

Tabea Rebafka

Sorbonne Université, Université de Paris, CNRS,
Laboratoire de Probabilités, Statistique et Modélisation, Paris, France.
e-mail: tabea.rebafka@sorbonne-universite.fr

Abstract: A model-based approach for the analysis of a collection of observed networks is considered. We propose to fit a stochastic block model to the data. The novelty consists in the analysis of not a single, but multiple networks. The major challenge resides in the development of a computationally efficient algorithm. Our method is an agglomerative algorithm based on the integrated classification likelihood criterion that performs simultaneously model selection and node clustering. Compared to the single-network context, an additional difficulty resides in the necessity to compare networks one to another and aggregate partial solutions. We propose a distance measure to compare stochastic block models and solve the label switching problem among graphs in a computationally efficient way.

Keywords and phrases: multiple networks, stochastic block model, integrated classification likelihood, agglomerative algorithm.

1. Introduction

Entire collections of networks are more and more available in various fields of application. In medical research, for instance, each patient may be described by his or her personal metabolic network. In ecology the interactions of species in different ecosystems are represented by so-called ecological networks. In social sciences or communication social interactions of individuals are observed for many different communities. Though there is a lack of statistical methods to analyze those data, as the literature mainly deals with the analysis of a single graph, which is already a challenging problem notably due to the complex dependencies inherent to relational data.

The most successful random graph model is the stochastic block model (SBM) (Nowicki and Snijders, 2001). It is a highly flexible model that offers a huge variety of graph topologies and is especially appropriate to model heterogeneous networks. A further advantage is interpretability of model parameters. Today numerous variants of the SBM exist (binary, valued, including covariates, degree-corrected, multipartite, dynamic versions, overlapping, mixed membership) emphasizing the relevance of the model.

The SBM is a discrete latent variable model and parameter estimation is challenging. Several inference algorithms have been proposed as a variational EM-algorithm (Daudin et al., 2008), MCMC methods (Peixoto, 2014), a pseudo-likelihood approach (Amini et al., 2013), or more recently a variational autoencoder based on neural networks (Mehta et al., 2019). These algorithms are time-consuming and not scalable to graphs with a very large number of nodes. While some of them are fast for a single run, they do not provide stable solutions and many runs are required. The problem of model selection, that is the choice of the number of possible values of the latent variables, increases the difficulty considerably.

Côme and Latouche (2015) propose an alternative approach based on the so-called integrated classification likelihood (ICL). Their hierarchical algorithm performs both model selection and clustering simultaneously. Recently Côme et al. (2020) proposed a genetic algorithm for the initialization that brings considerable improvement and makes the algorithm fast and scalable to large networks.

In this paper we consider a new problem: the analysis of a collection of networks that can be considered as independent realizations of a common model. While the analysis of a single network is already computationally challenging, increasing the number of observed networks makes the problem much harder and any inference algorithm must carefully pay attention to this issue.

To draw on the benefits of the SBM, we aim at fitting a common SBM to the collection of observed networks. That is, the networks are considered as independent realizations of a SBM with

fixed model parameter. We build on the ICL approach to develop a fast agglomerative inference algorithm performing parameter estimation, node clustering and model selection at a time. A major issue in our method is the comparison of networks one to another and to identify group of nodes in different networks that play the same role. To this end, we introduce a distance measure to compare two SBMs, match their blocks and thus solve the label switching problem among graphs in a computationally efficient way.

1.1. Stochastic block model for a collection of networks

Denote $\theta = (\boldsymbol{\pi}, \boldsymbol{\gamma})$ the parameter of a binary SBM with K blocks, group proportions $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K) \in (0,1)^K$ such that $\sum_k \pi_k = 1$ and connectivity matrix $\boldsymbol{\gamma} = (\gamma_{k,l})_{k,l} \in (0,1)^{K \times K}$. Throughout the paper we consider directed graphs without loops, but extensions are straightforward. Let $\mathcal{A} = (A^m)_{m \in [\![M]\!]}$ be a collection of M independent realizations of the SBM with model parameter θ. Here $A^m = (A^m_{i,j})_{1 \le i,j \le n_m} \in \{0,1\}^{n_m \times n_m}$ denotes the observed adjacency matrix of the m-th network. The number of vertices n_m may vary from one network to another.

More precisely, let $\mathbf{Z}^m = (Z^m_1, \dots, Z^m_{n_m}) \in [\![K]\!]$ be a vector of discrete latent variables for the nodes of the m-th network. We suppose $Z^m_i, i \in [\![n_m]\!], m \in [\![M]\!]$ independent and identically distributed with $\mathbb{P}(Z^m_i = k) = \pi_k$ for $k \in [\![K]\!]$. When convient, we use the one-hot encoding $Z^m_i = (Z^m_{i,1}, \dots, Z^m_{i,K}) \in \{0,1\}^K$ with $Z^m_i \sim \mathcal{M}(1, \boldsymbol{\pi})$. Then, conditionally on $\mathcal{Z} = (Z^m)_{m \in [\![M]\!]}$,

$$\mathcal{A}|\mathcal{Z} = \bigotimes_{m=1}^{M} A^m | Z^m = \bigotimes_{m=1}^{M} \bigotimes_{i \neq j} A^m_{i,j} | Z^m_i, Z^m_j = \bigotimes_{m=1}^{M} \bigotimes_{i \neq j} \mathcal{B}(\gamma_{Z^m_i, Z^m_j}),$$

where $\mathcal{B}(\gamma)$ denotes the Bernoulli distribution with parameter γ.

2. Integrated classification likelihood

To define the integrated classification likelihood (ICL) a Bayesian point of view is adopted. We consider classical conjugate priors $p(\theta) = p(\boldsymbol{\pi}, \boldsymbol{\gamma}) = p(\boldsymbol{\pi})p(\boldsymbol{\gamma})$ with

$$p(\boldsymbol{\pi}) = \text{Dirichlet}(\boldsymbol{\pi}; \delta^0_1, \dots, \delta^0_K), \qquad p(\boldsymbol{\gamma}) = \prod_{k,l} \text{Beta}(\gamma_{k,l}; \eta^0_{k,l}, \zeta^0_{k,l}).$$

Now the ICL criterion is defined as

$$\text{ICL}(\mathcal{A}, \mathcal{Z}) = \log(p(\mathcal{A}, \mathcal{Z})) = \int \log(p(\mathcal{A}, \mathcal{Z}|\theta; K)) p(\theta) \mathrm{d}\theta. \tag{1}$$

The ICL is well defined in various latent variable models like mixture models or the simple SBM. Traditionally, an asymptotic approximation of the ICL is used to perform model selection, that is, the estimation of the number of latent groups (Biernacki et al., 2000). Côme and Latouche (2015) showed that the exact ICL, as defined in (1), has closed-form expression for a simple SBM. Interestingly, by integrating out the model parameters, the criterion only depends on the observations $\mathcal{A}$ and the unknown latent variables $\mathcal{Z}$. The value of $\mathcal{Z}$ that optimizes the ICL is considered as a could estimate of the latent variables. For the SBM Côme and Latouche (2015) showed that the maximization of the ICL with respect to $\mathcal{Z}$ is feasible.

In our model of multiple networks generated from a common SBM, the independence of the networks yields that the ICL is the sum of the ICL of the single networks, that is,

$$\text{ICL}(\mathcal{A}, \mathcal{Z}) = \sum_{m=1}^{M} \log(A^m, Z^m) = \sum_{m=1}^{M} \text{ICL}^{\text{simple}}(A^m, Z^m).$$

2.1. ICL maximization

To maximize the ICL for a single SBM, Côme and Latouche (2015) propose a greedy hill-climbing algorithm. Starting from an initial node clustering, choose a vertice and search the best move to another block or keep the current block assignment if no move increases the ICL. Repeat this procedure until no more moves improve the ICL.More formally, one iteration is as follows:

1. Select a vertice i^*. Denote g the current block assignment of i^*, i.e. $Z_{i,g} = 1$.
2. For any block $h \neq g$ compute the impact on the ICL of moving node i^* to block h. That is, compute the difference $\Delta_{i^*}^{g\to h}$ of the current value of the ICL and the ICL with $Z_{i,h} = 1$. Set $\Delta_{i^*}^{h\to h} = 0$.
3. Choose the best block assignment as

$$h^* = \arg\max_{h\in[\![K]\!]} \Delta_{i^*}^{g\to h},$$

and move node i^* to block h^*, that is, set $Z_{i,h^*} = 1$ (and $Z_{i,g} = 0$ if $g \neq h^*$).

In the simple SBM, these changes of the ICL $\Delta_{i^*}^{g\to h}$ are easy to compute. In our model, as the ICL is the sum of the ICL of simple SBMs, $\Delta_{i^*}^{g\to h}$ can be computed rapidly in a similar way.

2.2. Model selection

When the algorithm is initialized with a clustering containing a large number of blocks K^0, it occurs that the algorithm empties blocks by transferring all nodes of a block to other blocks. Hence, the number of blocks diminishes and the final node clustering indirectly provides an estimate of the number of blocks K. That is, the algorithm performs model selection automatically.

2.3. Parameter estimation

In the ICL approach first model selection and node clustering are performed. Then, the model parameter θ can be estimated by the maximum likelihood estimator (MLE) on the complete data, that is,

$$\hat{\theta} = \arg\max_{\theta} p(\mathcal{A}, \mathcal{Z}^{\mathrm{ICL}}|\theta),$$

where $\mathcal{Z}^{\mathrm{ICL}}$ denotes the solution of the ICL maximization algorithm. In our model, the MLE has simple closed-form expressions.

3. Agglomerative algorithm

While the ICL maximization is a straightforward adaptation of the simple SBM to the model with multiple networks, the initialization or combination of partial solutions is a far more challenging issue. We propose the following agglomerative procedure: first a SBM is fitted to every network independently following Côme et al. (2020), then solutions are aggregated successively. We propose to merge networks two by two to ever larger groups of networks that share the same blocks until all networks are merged together. The principle is illustrated in Figure 1.

3.1. Merging networks or aggregation of SBMs

Let θ^1 and θ^2 be the parameters of two SBMs estimated on two sets of networks $\mathcal{A}^1$ and $\mathcal{A}^2$, resp. Aggregating these models refers to searching a common SBM that explains both sets of networks at a time, some kind of average of θ^1 and θ^2. The problem is that model parameters in the SBM are identifiable only up to label switching of the blocks. So the question is to match the blocks of the two SBMs that play the same role in the respective models. All permutations of the block

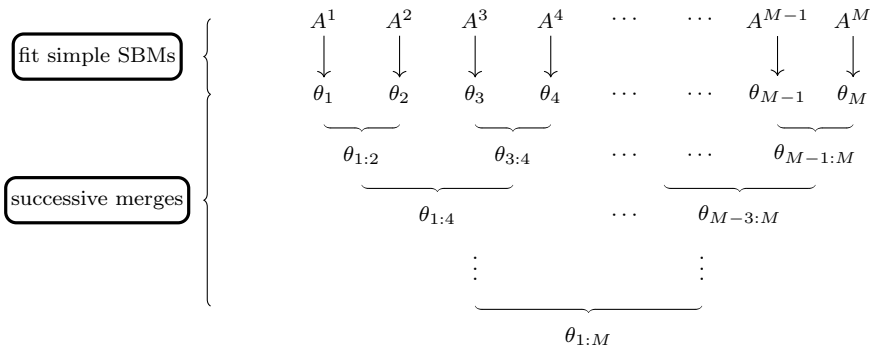

Figure 1: Agglomerative algorithm. First a SBM is fitted to each network independently following Côme et al. (2020), then merges are performed by matching SBM blocks and performing ICL maximization to the new groups of networks.

labels have to be considered to find the best match. And in case that the numbers of blocks of the two SBMs differ, we further have to identify pairs (or groups) of blocks of one model that correspond to a single block in the other SBM. For this task, we introduce a criterion, namely a distance measure for SBMs, to evaluate the correspondance of two SBMs for given block labelings.

Once the block correspondance identified, we relabel the node clusterings of the networks accordingly. Then the ICL maximization algorithm is applied to the union of the two sets of networks in order to update the node clusterings under the new common SBM.

4. Conclusion

In this paper a novel approach for the analysis of multiple networks is proposed. A computationally efficient agglomerative algorithm is developed to fit a SBM jointly to all networks. In the talk we will give more details on computational issues and the method to match the blocks of two SBMs. Moreover, we will illustrate the performance of the algorithm in a numerical study.

References

Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

Côme, E., Jouvin, N., Latouche, P., and Bouveyron, C. (2020). Hierarchical clustering with discrete latent variable models and the integrated classification likelihood.

Côme, E. and Latouche, P. (2015). Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling*, 15(6):564–589.

Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183.

Mehta, N., Duke, L. C., and Rai, P. (2019). Stochastic blockmodels meet graph neural networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4466–4474. PMLR.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Peixoto, T. (2014). Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E*, 89(1).