Modelling time evolving interactions in networks through a non stationary extension of stochastic block models

Marco Corneli, Pierre Latouche and Fabrice Rossi

Université Paris 1 Panthéon-Sorbonne - Laboratoire SAMM
90 rue de Tolbiac, F-75634 Paris Cedex 13 - France

Résumé. Le modèle à blocs stochastiques (SBM) décrit les interactions entre les sommets d'un graphe selon une approche probabiliste, basée sur des classes latentes. SBM fait l'hypothèse implicite que le graphe est stationnaire. Par conséquence, les interactions entre deux classes sont supposées avoir la même intensité pendant toute la période d'activité. Pour relaxer l'hypothèse de stationnarité, nous proposons une partition de l'horizon temporel en sous intervalles disjoints, chacun de même longueur. Ensuite, nous proposons une extension de SBM qui nous permet de classer en même temps les sommets du graphe et les intervalles de temps où les interactions ont lieu. Le nombre de classes latentes (K pour les sommets, D pour les intervalles de temps) est enfin obtenu à travers la maximisation de la vraisemblance intégrée des données complétées (ICL exacte). Après avoir testé le modèle sur des données simulées, nous traitons un cas réel. Pendant une journée, les interactions parmi les participants de la conférence HCM Hypertext (Turin, 29 Juin 1er Juillet 2009) ont été traitées. Notre méthodologie nous a permis d'obtenir une classifications intéressante des 24 heures: les moments de rencontre tels que les pauses café ou buffets ont bien été détectés. La complexité de l'algorithme de recherche, linéaire en fonction du nombre initial de clusters (K_{max} et D_{max} respectivement), nous oriente vers l'utilisation d'instruments avancés de classification, pour réduire le nombre attendu de classes latentes et ainsi pouvoir utiliser le modèle pour des réseaux de grand dimension.

Mots-clés. Graphes aléatoires, classification temporelle, modèles à blocs stochastiques, vraisemblance classifiante intégrée

Keywords. Random graphs, time event clustering, stochastic block models, integrated classification likelihood.

1 Introduction

Since the interactions between nodes of a network generally have a time varying intensity, the network has a non trivial time structure that we wish to infer. An example of this complexity can be observed in Figure (1). On the vertical axis the aggregated number of proximity face-to-face interactions (less than 1.5 meter) between attendees of the *HCM*

Hypertext conference (Turin, June 29th - July 1st, 2009) is given. On the horizontal axis, a time line is reported, corresponding to the 24 hours of the first day of conference. The

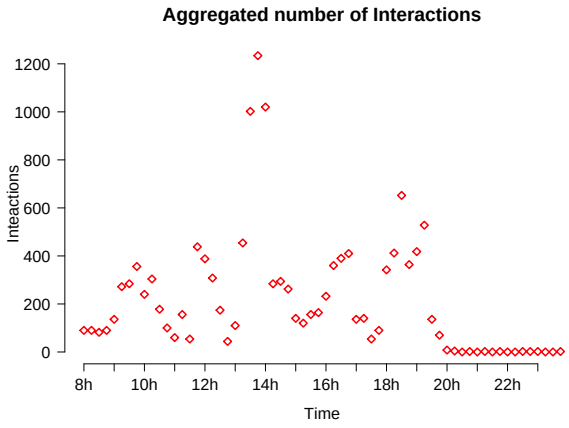

Figure 1: Aggregated number of interactions between conference attendees per quarter-hour.

day was partitioned in small time intervals of 20 seconds in the original data frame. We considered 15 minutes time aggregations leading to a partition of the day made of 96 consecutive quarter-hours: each red point in the figure corresponds to one of them. If we associate each attendee to a node and each interaction to an edge, the non stationarity of such a graph is clear. In the following section, after describing the general stochastic block model, we illustrate the temporal evolution we propose.

2 A non stationary stochastic block model

We present here the SBM (Holland et al. [2]), using the same notations as in Wyse, Friel, Latouche (2014). The set of nodes $A = \{a_1, \ldots, a_N\}$ is introduced. Undirected links between nodes i and j from A are counted by the observed variable X_{ij}, being the component (i, j) of the $N \times N$ adjacency matrix $X = \{X_{ij}\}_{i \leq N, j \leq N}$. Nodes in A are clustered in K disjointed subgroups respectively:

$$A = \cup_{k \leq K} A_k, \qquad A_l \cap A_g = \emptyset, \quad \forall l \neq g$$

Nodes in the same cluster in A have linking attributes of the same nature. We introduce an hidden vector $\mathbf{c} = \{c_1, \ldots, c_N\}$ labelling each node's membership:

$$c_i = k \qquad \textit{iff} \quad a_i \in A_k, \quad \forall k \leq K.$$

In order to introduce a temporal dimension, consider now a sequence of equally spaced, adjacent time steps $\{\Delta_u := t_u - t_{u-1}\}_{u \leq U}$ over the interval $[0, T]$ and a partition $C_1, \ldots, C_D$ of the same interval[1]. We introduce furthermore a random vector $\mathbf{y} = \{y_u\}_{u \leq U}$, such that

[1] T and U are linked by the following relation: $T = \Delta_u U$.

$y_u = d$ if and only if $I_u :=]t_{u-1}, t_u] \in C_d, \forall d \leq D$. We attach to $\mathbf{y}$ a multinomial distribution:

$$p(\mathbf{y}|\boldsymbol{\beta}, D) = \prod_{d \leq D} \beta_d^{|C_d|},$$

where $|C_d| = \#\{I_u : I_u \in C_d\}$. Now we define $N_{ij}^{I_u}$ as the number of observed connections between a_i and a_j, in the time interval I_u and we make the following crucial assumption:

$$p(N_{ij}^{I_u}|c_i = k, c_j = g, y_u = d) \qquad \text{follows a} \qquad \text{Poisson}(\Delta_u \lambda_{kgd}), \tag{1}$$

hence the number of interactions is *conditionally* distributed like a Poisson random variable with parameter depending on k, g, d (Δ_u is constant).

Notation: In the following, for seek of simplicity, we will note:

$$\prod_{k,g,d} := \prod_{k \leq K} \prod_{g \leq K} \prod_{d \leq D} \quad \text{and} \quad \prod_{c_i} := \prod_{i:c_i = k}$$

and similarly for $\prod_{c_j}$ and $\prod_{y_u}$.

The adjacency matrix, noted N^{Δ}, has three dimensions ($N \times N \times U$) and its observed likelihood can be computed explicitly:

$$p(N^{\Delta}|\Lambda, \mathbf{c}, \mathbf{y}, K, D) = \prod_{k,g,d} \frac{\Delta^{S_{kgd}}}{\prod_{c_i} \prod_{c_j} \prod_{y_u} N_{ij}^{I_u}!} e^{-\Delta \lambda_{kgd} R_{kgd}} \lambda_{kgd}^{S_{kgd}}, \tag{2}$$

where we noted $S_{kgd} := \sum_{c_i} \sum_{c_j} \sum_{y_u} N_{ij}^{I_u}$ and $R_{kgd} := |A_k||A_g||C_d|^2$. The subscript u was removed from Δ_u to emphasize that time steps are equally spaced for every u.

Since $\mathbf{c}$ and $\mathbf{y}$ are not known, a multinomial factorizing probability density $p(\mathbf{c}, \mathbf{y}|\Phi, K, D)$, depending on a hyperparameter Φ, is introduced. The joint distribution of labels looks finally as follows:

$$p(\mathbf{c}, \mathbf{y}|\Phi, K, D) = \left(\prod_{k \leq K} \omega_k^{|A_k|} \right) \left(\prod_{d \leq D} \beta_d^{|C_d|} \right), \tag{3}$$

where $\Phi = \{\boldsymbol{\omega}, \boldsymbol{\beta}\}$.

2.1 Exact ICL for non stationary SBM

The integrated classification criterion (ICL) was introduced as a model selection criterion in the context of Gaussian mixture models by Biernacky et al. [5]. Côme and Latouche [6] proposed an exact version of the ICL based on a Bayesian approach for the stochastic

[2]Self loops are considered here for seek of simplicity. The model can easily be extended to graphs with undirected links and/or no self loops.

block model and this is the approach we follow here. The quantity we focus on is the *complete data* log-likelihood, integrated with respect to the model parameters Φ and $\Lambda = \{\lambda_{kgd}\}_{k\leq K, g\leq K, d\leq D}$:

$$\mathcal{ICL} = \log\left(\int p(N^{\Delta}, \mathbf{c}, \mathbf{y}, \Lambda, \Phi | K, D) d\Lambda d\Phi\right). \tag{4}$$

Introducing a prior distribution $\nu(\Lambda, \Phi | K, D)$ over the pair Φ, Λ and thanks to ad hoc independence assumptions, the ICL can be rewritten as follows:

$$\mathcal{ICL} = \log\left(\nu(N^{\Delta} | \mathbf{c}, \mathbf{y}, K, D)\right) + \log\left(\nu(\mathbf{c}, \mathbf{y} | K, D)\right). \tag{5}$$

The choice of prior distributions over the model parameters is crucial to have an explicit form of the ICL.

3 ICL maximization and experiments

In order to maximize the integrated complete likelihood (ICL) in equation (5) with respect to the four unknowns $\mathbf{c}, \mathbf{y}, K, D$, we rely on a greedy search over labels and the number of nodes and time clusters. This approach is described in Wyse, Frial and Latouche [4] for a stationary latent block model.

Experiments were conducted on both simulated and real data. A detailed illustration of results will be provided during the conference.

References

[1] A.N. Randriamanamihaga, E. Côme, L. Oukhellou and G. Govaert, Clustering the Vélib' dynamic Origin/Destination flows using a family of Poisson mixture models, *Neurocomputing*, vol. 141, pp. 124-138, 2014.

[2] P.W.Holland, K.B. Laskey and S. Leinhardt, Stochastic blockmodels: first steps, *Social networks*, vol.5, pp.109-137, 1983.

[3] R. Guigourès, M. Boullé and F. Rossi, A Triclustering Approach for Time Evolving Graphs, in *Co-clustering and Applications, IEEE 12th International Conference on Data Mining Workshops (ICDMW 2012)*, pages 115-122, Brussels, Belgium, December 2012.

[4] J. Wyse, N. Friel and P. Latouche, Inferring structure in bipartite networks using the latent block model and exact ICL, *arXiv pre-print* arXiv: 1404.2911, 2014.

[5] C. Biernacky, G. Celeux and G. Govaert, Assessing a mixture model for clustering with the integrated completed likelihood, *IEEE Trans. Pattern Anal. Machine Intel*, vol.7, pp. 719-725, 2000.

[6] E. Côme and P. Latouche, Model selection and clustering in stochastic block models with the exact integrated complete data likelihood, *arXiv pre-print*, arXiv:1303.2962, 2013.

[7] L. Isella, J. Sthelé, A. Barrat, C.Cattuto, J.F. Pinton, W. Van den Broeck, What's in a crowd? Analysis of face-to-face behavioral networks, *Journal of Theoretical Biology*, vol. 271, pp. 166-180, 2011.