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

Marco Corneli
SAMM laboratory
Université Paris 1 Pantheon-Sorbonne
Paris, France 75634 Paris Cedex 13
Email: Marco.Corneli@malix.univ-paris1.fr

Pierre Latouche
SAMM laboratory
Université Paris 1 Pantheon-Sorbonne
Paris, France 75634 Paris Cedex 13
Email: pierre.latouche@univ-paris1.fr

Fabrice Rossi
SAMM laboratory
Université Paris 1 Pantheon-Sorbonne
Paris, France 75634 Paris Cedex 13
Email: Fabrice.Rossi@univ-paris1.fr

***Abstract*—The stochastic block model (SBM) [1] describes interactions between nodes of a network following a probabilistic approach. Nodes belong to hidden clusters and the probabilities of interactions only depend on these clusters. Interactions of time varying intensity are not taken into account. By partitioning the whole time horizon, in which interactions are observed, we develop a non stationary extension of the SBM, allowing us to simultaneously cluster the nodes of a network and the fixed time intervals in which interactions take place. The number of clusters as well as memberships to clusters are finally obtained through the maximization of the complete-data integrated likelihood relying on a greedy search approach. Experiments are carried out in order to assess the proposed methodology.**

I. A NON STATIONARY STOCHASTIC BLOCK MODEL

We consider a set of N nodes $A = \{a_1, \dots, a_N\}$. Directed links between these nodes are observed over the time interval $[0, T]$ and self loops are not considered. Nodes in A belong to K disjoint clusters:

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

We introduce an hidden vector $\mathbf{c} = \{c_1, \dots, c_N\}$ such that:

$$c_i = k \qquad \text{iff} \quad i \in A_k, \quad \forall k \leq K$$

and assume that $\mathbf{c}$ follows a multinomial distribution:

$$\mathbf{P}\{c_i = k\} = \omega_k \qquad \text{with} \qquad \sum_{k \leq K} \omega_k = 1.$$

As a consequence, modulo an independence assumption, the joint density for vector $\mathbf{c}$ is:

$$p(\mathbf{c}|\boldsymbol{\omega}, K) = \prod_{k \leq K} \omega_k^{|A_k|},$$

where $|A_k| = \#\{a_i : a_i \in A_k\}$ is the number of nodes inside cluster A_k.

In order to introduce a temporal structure, consider a sequence of equally spaced, adjacent time steps:

$$\Delta_u := t_u - t_{u-1}, \qquad u \in \{1, \dots, U\}$$

over the interval $[0, T]$ and a partition $C_1, \dots, C_D$ of the same interval[1]. We introduce a random vector $\mathbf{y} = \{y_u\}_{u \leq U}$, not observed, such that:

$$y_u = d \qquad \text{iff} \qquad I_u :=]t_{u-1}, t_u] \in C_d, \forall d \leq D.$$

We attach to $\mathbf{y}$ a multinomial distribution:

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

where $|C_d| = \#\{I_u : I_u \in C_d\}$. Assuming that labels $\mathbf{c}$ and $\mathbf{y}$ are independently distributed, their joint distribution is:

$$p(\mathbf{c}, \mathbf{y}|\Phi, K, D) = \left(\prod_{k \leq K} \omega_k^{|A_k|} \right) \left(\prod_{d \leq D} \beta_d^{|C_d|} \right), \quad (1)$$

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

We now define $N_{ij}^{I_u}$ as the number of observed connections from i to j, in the time interval I_u. The following crucial assumption is made:

$$p(N_{ij}^{I_u}|c_i = k, c_j = g, y_u = d) = p(N_{ij}^{I_u}|\lambda_{kgd}),$$

where:

$$p(N_{ij}^{I_u}|\lambda_{kgd}) = \frac{(\Delta_u \lambda_{kgd})^{N_{ij}^{I_u}}}{N_{ij}^{I_u}!} e^{-\Delta_u \lambda_{kgd}}. \quad (2)$$

We note Λ the set of all parameters λ_{kgd}.

Remarks: For i and j fixed and $\mathbf{c}$ known, the sequence:

$$N_{ij}^{I_1}, \dots, N_{ij}^{I_U}$$

is independently but not identically distributed, therefore this model is a *non-stationary* extension of the SBM. Moreover, this is the sequence of increments of a *non-homogeneous* Poisson process, counting interactions from cluster c_i to cluster c_j.

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

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

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

and similarly for $\prod_{c_j=g}$ and $\prod_{y_u=d}$.

The adjacency matrix, noted N^{Δ}, has three dimensions $(N \times N \times U)$:

$$N^{\Delta} = \{N_{ij}^{I_u}\}_{i \leq N, j \leq N, u \leq U}$$

and its observed likelihood can be computed explicitly:

$$p(N^{\Delta}|\Lambda, \mathbf{c}, \mathbf{y}, K, D) = \prod_{k,g,d} \prod_{c_i=k} \prod_{c_j=g} \prod_{y_u=d} p(N_{ij}^{I_u}|\lambda_{kgd}) = \prod_{k,g,d} \frac{(\Delta\lambda_{kgd})^{S_{kgd}}}{P_{kgd}} e^{-\Delta\lambda_{kgd} R_{kgd}}, \quad (3)$$

where we noted:

$$S_{kgd} := \sum_{c_i=k} \sum_{c_j=g} \sum_{y_u=d} N_{ij}^{I_u}, \quad P_{kgd} = \prod_{c_i=k} \prod_{c_j=g} \prod_{y_u=d} N_{ij}^{I_u}!$$

$$R_{kgd} := \begin{cases} |A_k||A_g||C_d| & \text{if} \quad g \neq k \\ |A_k||A_{k-1}||C_d| & \text{if} \quad g = k \end{cases}$$

Note that the subscript u was removed from Δ_u to emphasize that time steps are equally spaced for every u.

II. Exact ICL for non stationary SBM

In order to estimate the label vectors $\mathbf{c}$ and $\mathbf{y}$ and the number of clusters K and D, we directly maximize an exact version of the Integrated Classification Criterion (ICL) [2] by means of a greedy search [3]. The quantity we focus on is the *complete data* log-likelihood, integrated with respect to the model parameters Φ and Λ:

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

Introducing a prior distribution $\nu(\Lambda, \Phi|K, D)$ over the pair (Φ, Λ) and making an independence assumption, allow us to write the ICL as follows:

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

The choice of conjugated prior distributions is crucial to obtain an explicit form of the ICL. We impose a Gamma a priori over Λ:

$$\nu(\lambda_{kgd}|a, b) = \frac{b^a}{\Gamma(a)} \lambda_{kgd}^{a-1} e^{-b\lambda_{kgd}}$$

and a factorizing Dirichlet a priori distribution to Φ:

$$\nu(\Phi|K, D) = \text{Dir}_K(\boldsymbol{\omega}; \alpha, \ldots, \alpha) \times \text{Dir}_D(\boldsymbol{\beta}; \gamma, \ldots, \gamma),$$

where the parameters of each distribution have been set constant for simplicity.

III. Experiments

The dataset we used was collected during the **ACM Hypertext** conference held in Turin, June 29th - July 1st, 2009 and represents the dynamical network of face-to-face proximity interactions of 113 conference attendees over about 2.5 days[2].

[2]More informations can be found at: http://www.sociopatterns.org/datasets/hypertext-2009-dynamic-contact-network/.

We focused on the first conference day, namely the twenty four hours going from 8am of June 29th to 7.59am of June 30th. The day was partitioned in small time intervals of 20 seconds in the original data frame and face-to-face proximity interactions (less than 1.5 meters) were monitored by radio badges that conference attendees volunteered to wear. Further details can be found in [4]. We considered 15 minutes time aggregations, thus leading to a partition of the day made of 96 consecutive quarter-hours ($U = 96$ with previous notation).

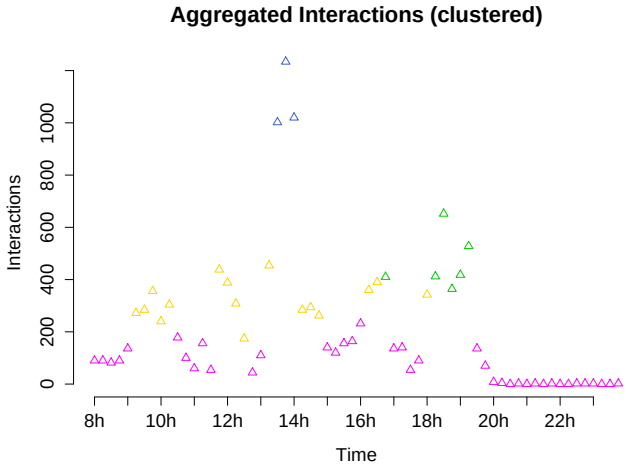

Fig. 1: Aggregated interactions for each time interval

The greedy search algorithm converged to a final ICL of -53217.4, corresponding to 20 clusters for nodes (people) and 4 time clusters. In Figure 1, the total number of interactions for each quarter-hour can be observed. Different colors correspond to the time clusters found by the model, in particular time intervals corresponding to the highest number of interactions have been placed in cluster C_4 (blue), those corresponding to an intermediate interaction intensity, in C_2 (green) and C_3 (yellow). Cluster C_1 (magenta) contains those quarter-hours where the frequency of interactions is very low or null. It is interesting to note how the model closely recovers times of social gathering:

- 9.00-10.30 - set-up time for posters and demos.
- 13.00-15.00 - lunch break.
- 18.00-19.00 - wine and cheese reception.

A complete program of the day can be found at: http://www.ht2009.org/program.php.

References

[1] P. Holland, K. Laskey, and S. Leinhardt, "Stochastic blockmodels: first steps," *Social Networks*, vol. 5, pp. 109–137, 1983.

[2] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 7, pp. 719–725, 2000.

[3] E. Côme and P. Latouche, "Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood," *Statistical Modelling*, 2015, to appear.

[4] L. Isella, J. Stehl, A. Barrat, C. Cattuto, J. Pinton, and W. Van den Broeck, "What's in a crowd? analysis of face-to-face behavioral networks," *Journal of Theoretical Biology*, vol. 271, no. 1, pp. 166–180, 2011.