Dynamic stochastic topic block model for time evolving networks with textual edges

Charles Bouveyron [1], Marco Corneli [2], Pierre Latouche [2] & Fabrice Rossi [2]

[1] *Laboratoire MAP5, UMR CNRS 8145,*
Université Paris Descartes et Sorobonne Paris Cité.
E-mail: charles.bouveyron@parisdescartes.fr
[2] *Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne.*
E-mail: Marco.Corneli@malix.univ-paris1.fr

Résumé. Dans ce papier, nous développons une approche probabiliste qui prend en compte non seulement la fréquence des interactions parmi les acteurs d'un réseau mais aussi le contenu des arêtes textuelles. L'objectif est de grouper les nœuds du réseau en classes qui soient homogènes en terme de fréquence d'interactions mais aussi de sujets traités. Ainsi, on dira qu'un réseau est stationnaire sur un sous-intervalle de temps quand les proportions de sujets traités entre chaque paire de groupes de nœuds ne changent pas dans le sous-intervalle. Les paramètres du modèle sont estimés à l'aide d'un algorithme de classification variationnelle de type EM, des expériences sur des données à la fois simulées et réelles montrent l'intérêt de l'approche proposée.

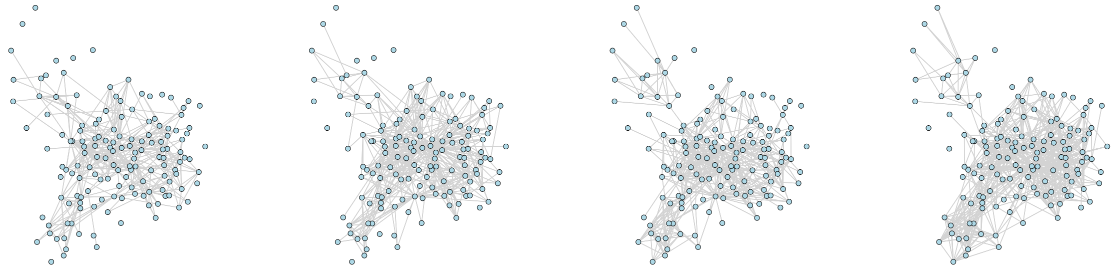
Mots-clés. Graphe aléatoire dynamique, Stochastic Block Model, processus de Poisson non homogène, modélisation des topics, Latent Dirichlet Allocation.

Abstract. In the present paper we develop a probabilistic approach accounting for the content of textual edges in a network as well as their frequency. The goal is to cluster the vertices into groups which not only are homogeneous in terms of amount of interactions but also in terms of discussed topics. Similarly, the network will be considered stationary on a time subinterval when the proportions of discussed topics between each pair of groups of nodes do not change in the sub interval. A classification variational expectation-maximization (C-VEM) algorithm is adopted to perform inference and experiments on both simulated and real data are used to assess the proposed methodology.

Keywords. Dynamic random graph, clustering, Stochastic Block Model, non homogeneous Poisson process, topic modelling, Latent Dirichlet Allocation.

1 Introduction

One of the main goals in network analysis is to cluster the vertices of a graph into groups/clusters of homogeneous interactivity patterns. This task can be accomplished by relying on non parametric techniques (e.g. modularity maximization, Newman and Girvan, 2004) as well as on probabilistic methods like the stochastic block model (SBM, Nowicki and Snijders, 2001). When dealing with dynamic (i.e. time varying) graphs, one approach consists in fixing the clusters of vertices in time to look for a partition of the whole time horizon such that interactions between groups are stationary on each sub-interval (Corneli et al. 2016). Alternatively, a time series of static graphs is built by aggregating interactions on sub-intervals and nodes are allowed to change groups at each step (Matias and Miele, 2016). Although the majority of the existing methods rely solely on the presence/absence of links between nodes, the analysis of the text content, in networks made out of textual edges, provides a deeper understanding of the network structure. The latent Dirichlet allocation model (LDA, Blei et al., 2003) represents documents as random mixtures over latent topics. Bouveyron et al. (2016) the STBM model which generalises both the SBM and the LDA and allows to handle (static) networks with textual edges. The basic ideas are summarised in the following steps: i) interactions between each pair of nodes (i, j) are sampled according to the SBM, ii) an interaction corresponds to a document (or a sequence of documents) sent from i to j, iii) the proportion of topics discussed in such messages only depends on the clusters of i and j. We propose an extension of the STBM, dealing with dynamic graphs. The dynamic network motivating this work is the Enron data set, containing all the email communications between the 149 employees of the company, from 1999 to 2001. In this paper, only the year 2001 is considered[1] and the approach we adopted, consists in partitioning the whole time horizon in sub-periods (months, weeks, etc.) and aggregating interactions/e-mails between employees on each period to obtain a sequence of static graphs. In Figure 1, e.g. each graph corresponds to a quarter of the year. The aim is to cluster vertices in groups *not* varying in time and to estimate the number and the composition of such groups, like in STBM. Moreover, time sub-periods are clustered in *time* clusters, in which topic proportions between each pair of *node* clusters are constant (they do not change in time).



(a) First quarter. (b) Second quarter. (c) Third quarter. (d) Fourth quarter.

Figure 1: The Enron e-mail data set viewed as a dynamic graph. Each graph corresponds to a quarter of 2001.

[1] The company failed bankruptcy on December, 2nd, 2001.

The paper is organized as follows: in Section 2 we introduce the model and obtain its complete-data likelihood, in Section 3 we discuss the estimation procedure and the final section concludes the paper and outlines the future works and perspectives.

2 The dynamic STBM (DSTBM)

2.1 Dynamic modelling of edges

A dynamic network consisting in time evolving instantaneous interactions between N nodes is considered. Interactions, observed over the time interval $[0, T]$ are directed and self loops are not allowed. In a block modelling perspective, nodes are clustered in K hidden groups $\mathcal{A}_1, \ldots, \mathcal{A}_K$ whose number has to be estimated. An hidden N-vector Z labels node memberships ($Z_i = k$ iff node i is in cluster $\mathcal{A}_k$). A multinomial probability distribution is attached to Z

$$p(Z|\pi) = \prod_{k=1}^{K} \pi_k^{|\mathcal{A}_k|},$$

where $\sum_{k=1}^{K} \pi_k = 1$ and $|\mathcal{A}_k|$ is the number of nodes in cluster $\mathcal{A}_k$[2]. Interactions from node i to node j are counted by a non homogeneous Poisson process (NHPP) $\{N_{ij}(t)\}_{t \leq T}$ whose intensity function, $\lambda_{ij}(t)$, positive and integrable on $[0, T]$ only depends on the clusters of the two nodes

$$N_{ij}(t)|Z_{ik}Z_{jg} = 1 \sim \mathcal{P}\left(\int_0^t \lambda_{kg}(u)du\right),$$

for $t \leq T$. The $N \times (N-1)$ NHPPs, associated with all different pairs (i, j) are assumed to be independent conditionally on Z to be known. The interval $[0, T]$ is now partitioned in U subintervals, $I_u := [t_{u-1}, t_u[$, where

$$0 = t_0 < t_1 < \cdots < t_U = T.$$

The size of I_u is denoted by Δ_u and it is set constant and unitary, without loss of generality Henceforth, we only focus on the increments of each process on the considered time partition

$$X_{iju} := N_{ij}(t_u) - N_{ij}(t_{u-1}), \qquad \forall(i, j, u).$$

In words, we focus on the number of interactions from i to j taking place over the time interval I_u. An $N \times N \times U$ tensor $X = \{X_{iju}\}_{i,j,u}$ is naturally defined. Furthermore the time intervals $I_1, \ldots, I_U$ are assigned to D disjoint hidden time clusters $\mathcal{C}_1, \ldots, \mathcal{C}_D$ whose number has to be estimated. Each cluster contains a certain number of time intervals,

[2] In the following, the zero-one notation ($Z_{ik} = 1$ if node i is in cluster A_k, it is null otherwise) will be used interchangeably, when no confusion arises.

not necessarily adjacent and an hidden U-vector Y is introduced to label memberships to time clusters ($Y_u = d$ iff I_u belongs to cluster $\mathcal{C}_d$). We assume that Y is sampled form a multinomial distribution

$$p(Y|\delta) = \prod_{d=1}^{D} \delta_d^{|\mathcal{C}_d|},$$

where $\sum_{d=1}^{D} \delta_d = 1$ and $|\mathcal{C}_d|$ denotes the number of time intervals in $\mathcal{C}_d$. The intensity functions are assumed to be stepwise constant on each time cluster $\mathcal{C}_d$, such that

$$X_{iju}|Z_{ik}Z_{jg}Y_{ud} = 1 \sim \mathcal{P}(\lambda_{kgd}).$$

A $K \times K \times D$ tensor $\Lambda = \{\lambda_{kgd}\}_{k,g,d}$ is introduced. The complete-data likelihood of the model can easily be obtained

$$p(X, Z, Y|\Lambda, \pi, \delta) = p(X|Z, Y, \Lambda)p(Z|\pi)p(Y|\delta),$$

where the random vectors Z and Y are independent by assumption.

2.2 Dynamic modelling of documents

The model described in the previous section can easily be extended by assuming that a directed interaction characterizing the pair (i, j) corresponds to a document sent from i to j. With the previous notations, X_{iju} the number of documents sent from i to j over the time interval I_u. The documents counted by X_{iju} are considered as a single one, obtained by concatenation. In the following, a dictionary is denoted by $\mathcal{W}$, it contains $|\mathcal{W}|$ words and W_n^{iju} is the n-th word (in the aggregate document) sent from i to j during the time interval I_u. The number of such words is L_{iju}. A list of Q *topics*, whose number has to be estimated, is introduced and each word is associated with one topic through a latent L_{iju}-vector noted V^{iju}. In formulas, $V_n^{iju} = q$ iff the word W_n^{iju} is associated with the q-th topic[3]. For each pair of clusters $\mathcal{A}_k$, $\mathcal{A}_g$ and a time cluster $\mathcal{C}_d$, a vector of topic proportions $\theta_{kgd} := (\theta_{kgdq})_{q \leq Q}$ is sampled from a Dirichlet distribution

$$\theta_{kgd} \sim \text{Dir}(\alpha = (\alpha_1, \ldots, \alpha_Q)),$$

such that for a pair of nodes (i, j) interacting in I_u

$$\mathbf{P}(V_{nq}^{iju} = 1|X, Z, Y, \theta) = \prod_{k,g}^{K} \prod_{d}^{D} \theta_{kgdq}^{Z_{ik}Z_{jg}Y_{ud}}$$

Given V, the word W_n^{iju} is assumed to be drawn from a multinomial distribution

$$W_n^{iju}|V_{nq}^{iju} = 1 \sim \mathcal{M}(1, \beta = (\beta_{q1}, \ldots, \beta_{q|\mathcal{W}|})).$$

[3] Once more, we emphasize that the zero-one notation $V_{nq}^{iju} = 1/0$ can be used alternatively.

Remark. The $Q \times |\mathcal{W}|$ matrix $\beta = \{\beta_{qv}\}_{q,v}$, of multinomial of probabilities, does not depend on the cluster assignments.

The complete-data conditional distribution of the LDA model, described so far, can be obtain straightforward

$$p(W, V, \theta | X, Z, Y, \beta) = p(W|V, X, \beta)p(V|X, Z, Y, \theta)p(\theta)$$

and the joint distribution of the DSTBM is

$$p(X, Z, Y, W, V, \theta|\Lambda, \pi, \delta, \beta) = p(W, V, \theta|X, Z, Y, \beta)p(X, Z, Y|\Lambda, \pi, \delta).$$

A graphical representation can be seen in the following figure

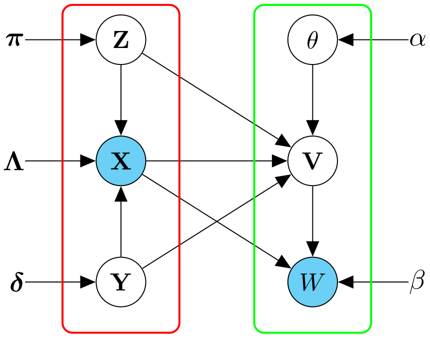


Figure 1: Graphical representation of the dynamic STBM.

3 Estimation Strategy

First, we assume that the number of clusters (K), time clusters (D) and the number of topics (Q) are known. Then, we proceed in two steps

1. We focus on the complete-data integrated log-likelihood

$$\log p(X, Z, Y, W|\Lambda, \pi, \delta, \beta) = \log \sum_V \int_\theta p(X, Z, Y, W, V, \theta|\Lambda, \pi, \delta, \beta)d\theta$$

for a given value of the pair (Z, Y) (e.g. random initialization). Since the above quantity is not directly tractable, a joint distribution $R(V, \theta)$ is introduced over the pair (V, θ) and a variational decomposition of the above log-likelihood is employed to obtain the following inequality

$$\log p(X, Z, Y, W|\Lambda, \pi, \delta, \beta) \geq \tilde{\mathcal{L}}(R(\cdot); Z, Y, \beta)) + \log p(X, Z, Y|\Lambda, \pi, \delta), \qquad (1)$$

where

$$\tilde{\mathcal{L}}(R(\cdot); Z, Y, \beta)) := \mathbf{E}_{R(V,\theta)}\left(\log \frac{p(W, V, \theta|X, Z, Y, \beta)}{R(V, \theta)}\right).$$

The lower bound on the right hand side of equation (1) is maximized with respect to $R(\cdot), \Lambda, \pi, \delta$ and β.

2. A greedy search approach is used to maximize the lower bound with respect to Z and Y, while holding the estimates, obtained in step 1, fixed.

The two steps above are repeated until no further increase in the lower bound is possible.

4 Conclusion and further work

We proposed a dynamic extension of the STBM, a probabilistic approach allowing to cluster nodes of a network with textual edges. The clustering algorithm takes into account the frequency of the interactions between pairs of nodes as well as the text contents. In addition, our methodology allows to track the evolution of the topic proportions in the communications between each pair of node clusters. The next steps consist in i)developing a model selection criterion to select K, D and Q, ii)testing the proposed methodology on simulated and real datasets to assess its potential.

Bibliography

[1] Newman, M. E. J. and Girvan, M. (2004), Finding and evaluating community structure in networks, *American Physical Society*.
[2] Nowicki, K. and Snijders, T. (2001). Estimation and prediction for stochas- tic block-structures. *Journal of the American Statistical Association*.
[3] Matias, C. and Miele, V. (2016). Statistical clustering of temporal networks through a dynamic stochastic block model *The Journal of the Royal Statistical Society: Series B*, to appear.
[4] Corneli, M., Latouche, P. and Rossi, F. (2016), Block modelling in dynamic networks with non-homogeneous Poisson processes and exact ICL, *Social Network Analysis and Mining*, Springer.
[5] Bouveyron, C., Latouche, P. and Zreik, R. (2017), The stochastic topic block model for the clustering of vertices in networks with textual edges, *Statistics and Computing*.
[6] Blei, D. M., NG, A. Y. and Jordan, M. I. (2003), Latent Dirichlet allocation, *Journal of Machine Learning Research*.