A State-Space Model for the Dynamic Random Subgraph Model

Rawya Zreik, Pierre Latouche, Charles Bouveyron

A State-Space Model for the Dynamic Random Subgraph Model

Rawya ZREIK[1,2], Pierre LATOUCHE[1] and Charles BOUVEYRON[2]

1- Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne

2- Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes

Abstract. In recent years, many random graph models have been proposed to extract information from networks. The principle is to look for groups of vertices with homogenous connection profiles. Most of these models are suitable for static networks and can handle different types of edges. This work is motivated by the need of analyzing an evolving network describing email communications between employees of the Enron compagny where social positions play an important role. Therefore, in this paper, we consider the random subgraph model (RSM) which was proposed recently to model networks through latent clusters built within known partitions. Using a state space model to characterize the cluster proportions, RSM is then extended in order to deal with dynamic networks. We call the latter the dynamic random subgraph model (dRSM).

1 Introduction

Network analysis has become an independent discipline which is no longer limited to sociology and is now applied in many areas such as biology, geography or history. The most recent statistical methods for the modeling and processing of the data are generally based on the stochastic block model (SBM) [1]. The SBM model assumes that each vertex belongs to a latent group, and that the probability of connection between a pair of vertices depends exclusively on their group. Among the recent extensions of the SBM model, we consider the random subgraph model (RSM) proposed by Jernite *et al.* [2]. The RSM model aims at modeling categorical edges using prior knowledge of a partition of the network into subgraphs. The subgraphs are assumed to be made of latent clusters which have to be inferred from the data in practice. The vertices are then connected with a probability depending only on the subgraphs whereas the edge type is assumed to be sampled conditionaly on the latent groups. In this work, we propose to extend the RSM model in order to deal with dynamic networks. The proposed model is called the dynamic random subgraph model (dRSM). A *state-space model* (SSM) is considered to characterize the temporal evolution of the cluster mixing proportions. The inference of this model is made through a variational EM (VEM) algorithm. The methodology is eventually applied to the famous Enron dataset describing the evolution of electronic communications between employees for two years (2001-2002). Figure 1 presents the evolution of the e-mail communication network for the four months before the financial collapse of the company in December 2001.

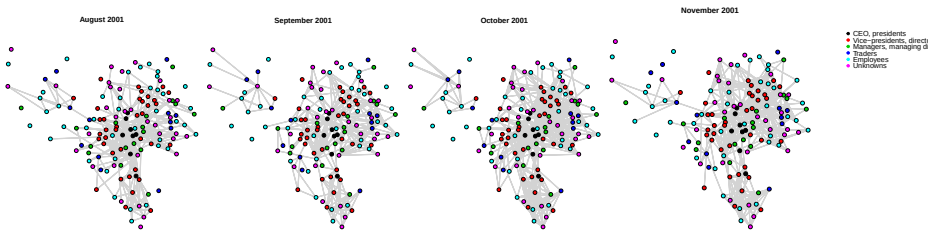

Fig. 1: Electronic communication network between 148 Enron employees during the 4 months (August-December 2001) before the bankruptcy of the company.

2 The dynamic random subgraph model

We consider a set of T networks $\{\mathcal{G}^{(t)}\}_{t=1}^{T}$, where $\mathcal{G}^{(t)}$ is a directed graph observed at time t and for which a partition $\mathcal{P}^{(t)}$ into S subgraphs is also known. Each $\mathcal{G}^{(t)}$ is represented by its $N \times N$ adjacency matrix $X^{(t)}$ where N denotes the number of nodes (assumed constant over time). No self loops are considered. The edge $X_{ij}^{(t)}$, describing the relationship between nodes i and j, is assumed to take its values in $\{0, \dots C\}$ such that $X_{ij}^{(t)} = c$ means that nodes i and j are linked by a relationship of type c at time t and $X_{ij}^{(t)} = 0$ indicates the absence of relationship. Our goal is to cluster at each time t the N nodes into K latent groups with homogeneous connection profiles, *i.e.* find an estimate at each time t of the binary matrix Z which is such that $Z_{ik}^{(t)} = 1$, if at time t, the node i belongs to the class k, and 0 otherwise.

2.1 *The model at each time* t

The network is assumed to be generated at each time t as follows. Each vertex i is first associated to a latent class k with a probability depending on the subgraph which it belongs to. We assume, that for a given number K of latent groups, the variable $Z_i^{(t)}$ is drawn from a multinomial distribution of parameter $\alpha_{s_i}^{(t)}$:

$$Z_i^{(t)} \sim \mathcal{M}(1, \alpha_{s_i}^{(t)}),$$

where $\alpha_s^{(t)} = (\alpha_{s_1}^{(t)}, \dots, \alpha_{s_K}^{(t)})$ is the vector of prior probabilities of the K latent groups in the subgraph s at time t and is such that $\sum_{k=1}^{K} \alpha_{s_k}^{(t)} = 1, \forall s \in 1, \dots, S$. On the other hand, we assume that the type of link between nodes i and j is sampled from a multinomial distribution depending on the latent vectors $Z_i^{(t)}$ and $Z_j^{(t)}$ as follows:

$$X_{i,j}^{(t)} | Z_{ik}^{(t)} Z_{jl}^{(t)} = 1 \sim \mathcal{M}(1, \Pi_{kl}),$$

with $\Pi_{kl} \in [0,1]^{C+1}$ and $\sum_{c=0}^{C} \Pi_{klc} = 1$.

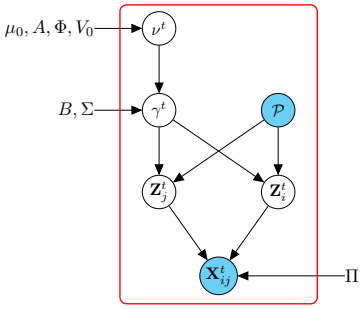

Fig. 2: *The graphical model for dRSM.*

2.2 *Modeling of the evolution of the random subgraphs*

We introduce now a hidden state, in the form of a *state-space model* [3, chapter 13] to capture the dynamic behavior of the proportions of the latent classes within the subgraphs over time. In order to apply this model, we introduce a new latent variable $\gamma_s^{(t)}$ which is assumed to be distributed according to a normal distribution with mean $B\nu^{(t)}$ and covariance matrix Σ, where:

$$\begin{cases} \nu^{(t)} = A\nu^{(t-1)} + \omega \\ \gamma_s^{(t)} = B\nu^{(t)} + v \\ \nu^{(1)} = \mu_0 + u. \end{cases}$$

The noise terms ω, u and v are supposed to be Gaussian: $\omega \sim \mathcal{N}(0, \Phi)$, $v \sim \mathcal{N}(0, \Sigma)$, $u \sim \mathcal{N}(0, V_0)$. A and B are two transition matrices of size $(K-1) \times (K-1)$. We finally assume a logistic link between the latent group proportions $\alpha_s^{(t)}$ and the hidden variable $\gamma_s^{(t)}$ in the form of $\alpha_s = f(\gamma_s)$ where,

$$\alpha_{sk}^{(t)} = \exp(\gamma_{sk}^{(t)} - C(\gamma_s^{(t)})), \ \forall\, k = 1, \ldots, K, \tag{1}$$

and $C(\gamma_s^{(t)}) = \sum_{\ell=1}^{K} \exp(\gamma_{s\ell}^{(t)})$. Thus, the state-space model of $\gamma^{(t)}$ allows us to characterize the temporal process of the latent group proportions. Due to the bijectivity constrain of this logistic transformation, $\alpha_s^{(t)}$ only has $K-1$ degree of freedom. Thus, we only need to draw the first $K-1$ components of $\gamma_s^{(t)}$ in which the last component is arbitrarily set to zero, to generate $\alpha_s^{(t)}$. Finally, our model has three latent variables (ν, γ, Z) and is parameterized by $\theta = (\mu_0, A, B, \Phi, V_0, \Sigma, \Pi)$ where the vector μ_0 has $(K-1)$ components, whereas $(A, B, \Phi, V_0, \Sigma)$ are $(K-1) \times (K-1)$ matrices. This model is called the *dynamic random subgraph model* (dRSM). Figure 2 presents the graphical model for dRSM.

2.3 *Inference with VEM algorithm for the dRSM model*

We aim at maximizing the log-likelihood $\log p(X|\theta)$ associated with the model. To achieve this maximization, a common approach consists in using an EM algo-

rithm. However, such an algorithm connot be derived here since $p(Z|X,\theta)$ is intractable. Therefore, we propose to use a variational EM-type algorithm (VEM), which locally optimizes the model parameters with respect to a lower bound of the log-likelihood. Traditionally, the VEM algorithm focuses on optimizing a lower bound of the form $\mathcal{L}(q,\theta) = \sum_z \int_\gamma \int_\nu q(Z,\gamma,\nu) \log \frac{p(X,Z,\gamma,\nu|\theta)}{q(Z,\gamma,\nu)} d\gamma \, d\nu$. Unfortunately, because $\log p(Z|\alpha = f(\gamma))$ involves here a non linear transformation, additional approximations are required. Making use of a Taylor expansion for the term $\log C(\gamma_s^{(t)})$, we derive the following inequality:

$$\begin{aligned} \log p(Z|\alpha)) &\geq \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{i=1}^{N} Y_{is} Z_{ik}^{(t)} \Big(\gamma_{sk}^{(t)} - \big(\xi_s^{-1(t)} \sum_{l=1}^{K} \exp(\gamma_{sl}^{(t)}) - 1 + \log(\xi_s^{(t)})\big)\Big) \quad (2) \\ &= \log h(Z,\gamma,\xi), \quad (3) \end{aligned}$$

where $\xi_s^t \in \mathbb{R}^{*+}$ is a new variational parameter. Replacing $\log p(Z|f(\gamma))$ by $\log h(Z,\gamma,\xi)$ in $\mathcal{L}(q,\theta)$, a new lower bound $\tilde{\mathcal{L}}(q,\theta)$ for $\log p(X|\theta)$ is obtained. We finally assume that $q(Z,\gamma,\nu)$ can be factorized, that is:

$$q(Z,\gamma,\nu) = q(Z)q(\gamma)q(\nu) = \Big(\prod_{t=1}^{T}\prod_{i=1}^{N} q(Z_i^{(t)})\Big)\Big(\prod_{t=1}^{T} q(\gamma^{(t)})\Big)\Big(\prod_{t=1}^{T} q(\nu^{(t)})\Big),$$

and $q(\gamma)$ is a product of normal distributions of parameters $\hat{\gamma}_{sk}^{(t)}$,$\hat{\sigma}_{sk}^{(t)}$:

$$q(\gamma) = \prod_{t=1}^{T}\prod_{s=1}^{S}\prod_{k=1}^{K-1} \mathcal{N}(\gamma_{sk}^{(t)}; \hat{\gamma}_{sk}^{(t)}, \hat{\sigma}_{sk}^{(t)^2}).$$

The VEM update step (E-step) for the distribution $q(Z_i^{(t)})$ is given by:

$$q(Z_i^{(t)}) \sim \mathcal{M}(Z_i^{(t)}; 1, \tau_i^{(t)}) \ \forall\, i,t.$$

However, for the distribution of $q(\nu)$, it was not possible to identify a usual probability distribution, but we recognized the distribution associated with a *state-space model*. Thus, the corresponding parameters can be estimated using the standard Kalman filter and Rauch-Tung-Striebel (RTS) smoother equations.

$$q(\nu) \propto p(\nu^{(1)}|\mu_0, V_0)\Big[\prod_{t=2}^{T} p(\nu^{(t)}|\nu^{(t-1)}, A, \Phi)\Big]\Big[\prod_{t=1}^{T} p(\frac{\sum_{s=1}^{S}\hat{\gamma}_s^{(t)}}{S}|\nu^{(t)}, \frac{\Sigma}{S}, B)\Big],$$

where $\nu^{(t)}$ is a latent variable and $x^{(t)} = \dfrac{\sum_{s=1}^{S}\hat{\gamma}_s^{(t)}}{S}$ seen as an observed variable such that $x^{(t)} = C\nu^{(t)} + \tilde{v}$ and $\tilde{v} \sim \mathcal{N}(0, \dfrac{\Sigma}{S})$.

In the M-step of the VEM algorithm updating formulas for the parameters Π and ξ can be obtained by maximizing the lower bound $\tilde{\mathcal{L}}(q,\theta)$. Updates for $\hat{\gamma}_{s_{ik}}^{(t)}$ and $\hat{\sigma}_{sl}^{(t)^2}$ have however to be found by numerically maximizing $\tilde{\mathcal{L}}(q,\theta)$ using a quasi-Newton algorithm.

t	periods
t_1	from 01/01/2000 to 01/12/2000
t_2	from 01/01/2001 to 01/03/2001
t_3	from 01/04/2001 to 01/06/2001
t_4	from 01/07/2001 to 01/09/2001
t_5	from 01/10/2001 to 01/03/2002

Table 1: The time periods for the study.

3 Application to the Enron network

In this section, we apply our methodology to the Enron data set, which describes the exchange of emails among 148 individuals who have worked for the Enron company at each time t. We are interested here in five time periods noted $t_1, t_2, \ldots, t_5$ (Table 1). A partition of the employees into three subgraphs depending on their status in the company (Managers, Employees, Others) is also available. The network is a directed and binary network without self loops, *i.e.* $C = 1$ and $X_{ij}^t = 1$ if i and j exchanged at least one email during the period t, 0 otherwise, with $t \in \{1, \ldots, 5\}$ and $S = 3$. We used the variational EM algorithm introduced in the previous section in order to look for $K = 4$ latent groups in the data.

3.1 Results

Since the network is binary, the dRSM model used here is a special case of the model we proposed with $C = 1$. By construction, the matrix Π verifies $\Pi_{kl0} + \Pi_{kl1} = 1, \forall (k, l)$ therefore, only the Π_{kl1} terms describing the connections probabilities are given in Table 2. Table 2 shows that the three clusters (1, 2 and 4) correspond to the communities where the probability of connection between two nodes of the same community is stronger than between different communities nodes. Thus, these clusters are mainly distinguished that they have different intra-cluster probabilities of connection, where the cluster 1 has the highest density (0.478), followed by the clusters 2 and 4 . Finally, the cluster 3 is built from low probabilities of connection (0.001). It gathers in fact at all individuals participating in non structured exchange in the network.

Applying our dRSM model on the Enron network allows the characterization of the subgraphs evolution with latent clusters according to the time. Figure 3 presents all estimated proportions of three subgraphs. We observe a drop in the proportion of cluster 3, in all subgraphs between t_4 and t_5, *i.e.* just before and after the opening of the investigation by the US federal agency. This specific network structure is here a reaction to the crisis of October 2001. The employees exchange emails on the subject and contact people preferentially. At this period, the proportion of cluster 4 (lower intra-cluster density), such as the one of cluster 3, increase. It is worth noticing that the structuring of the network starts earlier (at t_3) among managers than among employees. The subgraphs 2 and 3 have a

	cluster 1	cluster 2	cluster 3	cluster 4
cluster 1	0.478	0.037	0.005	0.023
cluster 2	0.020	0.181	0.006	0.012
cluster 3	0.001	0.002	0.001	0.003
cluster 4	0.012	0.012	0.024	0.119

Table 2: Terms Π_{kl1} of the matrix Π estimated using the VEM algorithm

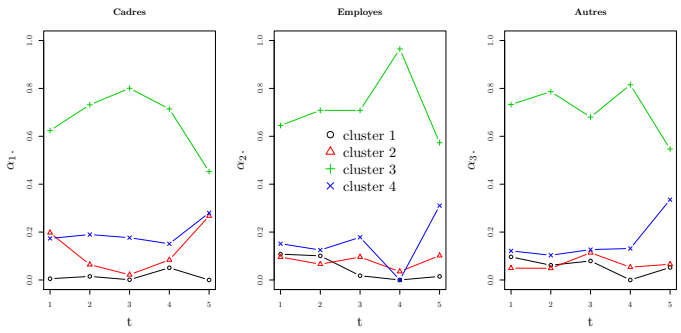

Fig. 3: Proportions of the $K = 4$ clusters, at each time. Subgraph 1 (Managers), left figure; subgraph 2 (employees), middle figure; subraph 3 (other), right figure.

mild reaction to others, but it disappears at t_4. This observation suggests that managers were aware of the arrival of the crisis before other employees. Finally, the cluster 1 with high intra-cluster density allows to see that the managers are the only individuals in the network for which we observe a diminution in the proportion of the cluster 1 at the time (t_4, t_5) unlike the subgraphs 2 and 3. Finally, through these observations and by taking into consideration the high position of managers where the exchange of emails is very preferable, allows to be separated the managers from the rest of the network.

References

[1] Y.J. Wang and G.Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82:8–19, 1987.

[2] Y. Jernite, P. Latouche, C. Bouveyron, P. Rivera, L. Jegou, and S. Lamassé. The random subgraph model for the abalysis of an acclesiasrical network in merovingian gaul. *Annals of Applied Statistics*, 2013.

[3] C.M. Bishop. *Pattern recognition and machine learning*. Springer-Verlag, 2006.