Statistical clustering of temporal networks through a dynamic stochastic block model

Catherine Matias, Vincent Miele

Statistical clustering of temporal networks through a dynamic stochastic block model

Catherine Matias†

Laboratoire de Probabilités et Modèles Aléatoires, UMR CNRS 7599, Université Pierre et Marie Curie, Université Paris Diderot, Paris, France.

Vincent Miele

Université de Lyon, F-69000 Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622 Villeurbanne, France.

Abstract. Statistical node clustering in discrete time dynamic networks is an emerging field that raises many challenges. Here, we explore statistical properties and deterministic inference in a model that combines a stochastic block model (SBM) for its static part with independent Markov chains for the evolution of the nodes groups through time. We model binary data as well as weighted dynamic random graphs (with discrete or continuous edges values). Our approach particularly focuses on the control for label switching issues across the different time steps. We study identifiability of the model parameters, propose an inference procedure based on a variational expectation maximization algorithm as well as a model selection criterion to select for the number of groups. We carefully discuss our initialization strategy which plays an important role in the method and compare our procedure with existing ones on synthetic datasets. We also illustrate our approach on a real data set of encounters among high school students and provide an implementation of the method into a R package called `dynsbm`.

Keywords: dynamic random graph, graph clustering, stochastic block model, variational expectation maximization

1. Introduction

Statistical network analysis has become a major field of research, with applications as diverse as sociology, ecology, biology, internet, etc. General references on statistical modeling of random graphs include the recent book by Kolaczyk (2009) and the reviews (Goldenberg et al., 2010; Snijders, 2011). While static approaches have been developed as early as in the 60's (mostly in the field of sociology), the literature concerning dynamic models is much more recent. Modeling discrete time dynamic networks is an emerging field that raises many challenges.

An important part of the literature on static network analysis is dedicated to clustering methods, with both aims of taking into account the intrinsic heterogeneity of the data and summarizing this data through node classification. Among clustering approaches, community detection methods form a smaller class of methods that aim at finding groups of highly connected nodes. Our focus here is not only on community detection but more generally on node classification based on connectivity behaviours, with a particular interest on model-based approaches. We refer to Matias and Robin (2014) for an overview of recent results on statistical model-based clustering methods for random graphs. When considering a sequence

†Corresponding author.

of snapshots of a network at different time steps, these static clustering approaches will give rise to classifications that are difficult to compare through time and thus difficult to interpret. An important thing to note is that label switching between two successive time steps may not be solved without an extra assumption e.g. that most of the nodes do not change group across two different time steps. However to our knowledge, this kind of assumption has never been discussed in the literature. In this work, we are interested in statistical models for discrete time dynamic random graphs, with the aim of providing a node classification across time, while controlling for label switching issues across the different time steps.

Stochastic block models (SBM) form a widely used class of statistical (and static) random graphs models that provide a clustering of the nodes. SBM introduces latent (i.e. unobserved) random variables on the nodes of the graph, taking values in a finite set. These latent variables represent the nodes groups and interaction between two nodes is governed by these corresponding groups. The model includes (but is not restricted to) the specific case of community detection, where the probability of connection between two nodes is higher when they belong to the same group. Combining SBM with a Markov structure on the latent part of the process (the nodes classification) is a natural way of ensuring a smooth evolution of the groups across time. This has been considered by many authors, in different ways as we discuss now. In Yang et al. (2011), the authors consider undirected, either binary or finitely valued, discrete time dynamic random graphs with no self-loops. As already said, the static aspect of the data is handled through SBM, so that at each time point, nodes belong to (a finite number of) unobserved groups, whose values determine their probability of connection. Now the dynamic aspect of the model is as follows. For each node, its group membership forms a Markov chain, independent of the values of the other nodes memberships. In this work, only the group membership is allowed to vary across time while connectivity parameters among groups stay constant through time. The authors propose a probabilistic simulated annealing algorithm to infer these parameters (either online or offline), based on a combination of Gibbs sampling and simulated annealing. For binary random graphs, Xu and Hero (2014) propose to introduce a state-space model through time on (the logit transform of) the probability of connection between groups. Contrarily to the previous work, both group membership and connectivity parameters per groups pairs may vary through time. Their (online) iterative estimation procedure is based on alternating two steps: a label-switching method to explore the space of node groups configuration, and the (extended) Kalman filter that optimizes the likelihood when the group memberships are known. Note that neither Yang et al. (2011) nor Xu and Hero (2014) propose to infer the number of clusters. Bayesian variants of these dynamic SB models may be found for instance in Ishiguro et al. (2010); Herlau et al. (2013).

Surprisingly, we noticed that the above mentioned methods were evaluated on synthetic datasets in terms of mean value over the time steps of a clustering quality index computed at fixed time step. Naturally, those indexes do not penalize for label switching and two classifications that are identical up to a permutation have the highest quality index value. Computing an index for each time step, the label switching issue between different time steps disappears and the classification task becomes easier. Indeed, such criteria do not control for a smoothed recovery of groups along different time points. It should also be noted that the synthetic experiments from these works were performed under the particular binary affiliation case (where only two connectivity parameters - intra and inter groups - are allowed), while we explain below that those dynamic versions of the affiliation SB model do not have identifiable parameters. In particular, the label switching issue between different

time steps may not be easily removed in this particular case.

Other approaches for model-based clustering of dynamic random graphs do not rely directly on SBM but rather on variants of SBM. When there is an already known partition on the set of graphs and clustering is to be stack on this partition, some authors advocate for the use of the so-called random subgraph model (RSM, see Jernite et al., 2014). The model combines SBM with the known partition by authorizing the groups proportions to differ in the different subgraphs. A dynamic version of RSM that builds upon the approach of Xu and Hero (2014) appears in Zreik et al. (2015). Detection of persistent communities has been proposed in Liu et al. (2014) for directed and dynamic graphs of call counts between individuals. Here the static underlying model is a (time and) degree-corrected SBM (Karrer and Newman, 2011) with Poisson distribution on the call counts. Group memberships are independent through time instead of Markov, but smoothness in the classification is obtained by imposing that intra groups expected call volumes (that represent part of the parameter specifying the distribution of intra groups counts) are constant through time. Inference is performed through a heuristic greedy search in the space of group memberships, as proposed in Karrer and Newman (2011). Note that only real datasets and no synthetic experiments have been explored in this latter work.

Another very popular statistical method for analyzing static networks is based on latent space models. Each node is associated to a point in a latent space and probability of connection is higher for nodes whose latent points are closer (Hoff et al., 2002). In Sarkar and Moore (2005), a dynamic version of the latent space model is proposed, where the latent points follow a (continuous state space) Markov chain, with transition kernel given by a Gaussian perturbation on current position with zero mean and small variance. Latent position inference is performed in two steps: a first initial guess is obtained through multi dimensional scaling. Then, nonlinear optimization is used to maximise the model likelihood. The work by Xu and Zheng (2009) is very similar, adding a clustering step on the nodes. The clustering uses both latent position and a link factor between individuals. Finally, Heaukulani and Ghahramani (2013) rely on Monte Carlo Markov Chain methods to perform a Bayesian inference in a more complicate setup where the latent positions of the nodes are not independent.

Mixed membership models (Airoldi et al., 2008) are also explored in a dynamic context. The work by Xing et al. (2010) relies on a state space model for the evolution of the parameters of the priors of both the mixed membership vector of a node and the connectivity behaviour. Inference is carried out through a variational Bayes expectation maximisation (`VBEM`) algorithm (e.g. Jordan et al., 1999). This concludes our non exhaustive bibliography on model-based clustering methods for dynamic random graphs.

In the present work, we explore statistical properties and deterministic inference in a model that combines SBM for its static part with independent Markov chains for the evolution of the nodes groups through time. Thus our setup is very close to the ones of Yang et al. (2011); Xu and Hero (2014), the main difference being that we allow for both groups and parameters to vary through time and discuss identifiability conditions for valid statistical inference. Moreover, we model binary data as well as weighted random graphs (with discrete or continuous edges) and propose a model selection criterion to choose the number of clusters. For simplicity of notation, we develop our model for undirected random graphs with no self-loops but easy generalizations could be obtained to cover for directed

datasets and/or including self-loops.

Section 2.1 describes the model and sets notation. In Section 2.2, we study parameters identifiability (up to label switching). We first highlight the fact that both groups and parameters values may not freely vary across time without identifiability problems. Then we exhibit a very natural constraint on our parameters and establish that it suffices to obtain identifiability of our parameterization. We note that to our knowledge, it is the first dynamic random graph model where parameters identifiability (up to label switching) is discussed and established. Moreover, we stress that dynamic affiliation SBM does not have identifiable parameters and groups may not be recovered consistently across time (Section 2.3). This is an important point as previous authors have tried to recover groups from this type of synthetic datasets and evaluated their estimated classification in a non natural way. Then, Section 3 describes a variational expectation maximization (`VEM`) procedure for inferring the model parameters and clustering the nodes. The `VEM` procedure works with a fixed number of groups and an Integrated Classification Likelihood (ICL, Biernacki et al., 2000) criterion is proposed for estimating the number of groups. We provide explicit formula in many classical (binary or weighted) examples for the conditional distribution of the edges, given the nodes groups (Section 3.2). We also discuss initialization of the algorithm - an important but rarely discussed step, in Section 3.3. Synthetic experiments are presented in Section 4. There, we discuss classification performances without neglecting the label switching issue that may occur between time steps. We also illustrate our method through the analysis of a real dataset of encounters among high school students in Section 5. Finally, an extension of this work to the case where some nodes are not present at every time point is sketched in Section 6. We mention that the methods are implemented into a R package available at `http://lbbe.univ-lyon1.fr/dynsbm` and will soon be available on the CRAN.

2. Setup and notation

2.1. Model description

We consider weighted interactions between N individuals recorded through time in a set of data matrices $\mathbf{Y} = (Y^t)_{1\leq t\leq T}$. Here T is the number of time points and for each value $t \in \{1,\dots,T\}$, the adjacency matrix $Y^t = (Y^t_{ij})_{1\leq i\neq j\leq N}$ contains real values measuring interactions between individuals $i, j \in \{1,\dots,N\}^2$. Without loss of generality, we consider undirected random graphs without self-loops, so that Y^t is a symmetric matrix with no diagonal elements.

We assume that the N individuals are split into Q latent (unobserved) groups that may vary through time, as encoded by the random variables $\mathbf{Z} = (Z^t_i)_{1\leq t\leq T, 1\leq i\leq N}$ with values in $\mathcal{Q}^{NT} := \{1,\dots,Q\}^{NT}$. This process is modeled as follows. Across individuals, random variables $(Z_i)_{1\leq i\leq N}$ are independent and identically distributed (iid). Now, for each individual $i \in \{1,\dots,N\}$, the process $Z_i = (Z^t_i)_{1\leq t\leq T}$ is an irreducible, aperiodic stationary Markov chain with transition matrix $\boldsymbol{\pi} = (\pi_{qq'})_{1\leq q,q'\leq Q}$ and initial stationary distribution $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_Q)$. When no confusion occurs, we may alternatively consider Z^t_i as a value in $\mathcal{Q}$ or as a random vector $Z^t_i = (Z^t_{i1},\dots,Z^t_{iQ}) \in \{0,1\}^Q$ constrained to $\sum_q Z^t_{iq} = 1$.

Given latent groups $\mathbf{Z}$, the time varying random graphs $\mathbf{Y} = (Y^t)_{1\leq t\leq T}$ are independent, the conditional distribution of each Y^t depending only on Z^t. Then, for fixed $1 \leq t \leq T$, random graph Y^t follows a stochastic block model. In other words, for each time t, conditional on Z^t, random variables $(Y^t_{ij})_{1\leq i<j\leq N}$ are independent and the distribution of

each Y_{ij}^t only depends on Z_i^t, Z_j^t. For now, we assume a very general parametric form for this distribution on $\mathbb{R}$. Following Ambroise and Matias (2012), in order to take into account possible sparse weighted graphs, we explicitly introduce a Dirac mass at 0, denoted by δ_0, as a component of this distribution. More precisely, we assume

$$Y_{ij}^t | \{Z_{iq}^t Z_{jl}^t = 1\} \sim (1 - \beta_{ql}^t)\delta_0(\cdot) + \beta_{ql}^t F(\cdot, \gamma_{ql}^t), \tag{1}$$

where $\{F(\cdot, \gamma), \gamma \in \Gamma\}$ is a parametric family of distributions with no point mass at 0 and densities (with respect to Lebesgue or counting measure) denoted by $f(\cdot, \gamma)$. This could be the Gaussian family with unknown mean and variance, the truncated Poisson family on $\mathbb{N}\backslash\{0\}$ (leading to a 0-inflated or 0-deflated distribution on the edges of the graph), etc. Note that the binary case is encompassed in this setup with $F(\cdot, \gamma) = \delta_1(\cdot)$, namely the parametric family of laws is reduced to a single point, the Dirac mass at 1 and conditional distribution of Y_{ij}^t is simply a Bernoulli $\mathcal{B}(\beta_{ql}^t)$. In the following and by opposition to the 'binary case', we will call 'weighted case' any setup where the set of distributions F is parametrized and not reduced to a single point. Here, the sparsity parameters $\beta^t = (\beta_{ql}^t)_{1 \leq q,l \leq Q}$ satisfy $\beta_{ql}^t \in [0, 1]$, with $\beta^t \equiv 1$ corresponding to the particular case of a complete weighted graph. As a result of considering undirected graphs, the parameters $\beta_{ql}^t, \gamma_{ql}^t$ moreover satisfy $\beta_{ql}^t = \beta_{lq}^t$ and $\gamma_{ql}^t = \gamma_{lq}^t$ for all $1 \leq q, l \leq Q$. Note that for the moment, SBM parameters may be different across time points. We will go back to this point in the next section. The model is thus parameterised by

$$\theta = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = (\boldsymbol{\pi}, \{\beta^t, \gamma^t\}_{1 \leq t \leq T}) = (\{\pi_{qq'}\}_{1 \leq q,q' \leq Q}, \{\beta_{ql}^t, \gamma_{ql}^t\}_{1 \leq t \leq T, 1 \leq q \leq l \leq Q}) \in \Theta,$$

and we let $\mathbb{P}_\theta$ denote the probability distribution on the whole space $\mathcal{Q}^{\mathbb{N}} \times \mathbb{R}^{\mathbb{N}}$. We also let $\phi(\cdot; \beta, \gamma)$ denote the density of the distribution given by (1), namely

$$\forall y \in \mathbb{R}, \quad \phi(y; \beta, \gamma) = (1 - \beta)1\{y = 0\} + \beta f(y, \gamma)1\{y \neq 0\},$$

where $1\{A\}$ is the indicator function of set A. With some abuse of notation and when no confusion occurs, we shorten $\phi(\cdot; \beta_{ql}^t, \gamma_{ql}^t)$ to $\phi_{ql}^t(\cdot)$ or $\phi_{ql}^t(\cdot; \theta)$. Directed acyclic graphs (DAGs) describing the dependency structure of the variables in the model, with different levels of detail, are given in Figure 1.

2.2. Parameters identifiability (general case)

Let us recall that with discrete latent random variables, identifiability can only be obtained up to a label switching on the node groups $\mathcal{Q}$. For any permutation σ in $\mathfrak{S}_Q$ (the set of permutations on $\mathcal{Q}$) and any $\theta \in \Theta$, we define

$$\sigma(\theta) := (\{\pi_{\sigma(q)\sigma(q')}\}_{1 \leq q,q' \leq Q}, \{\beta_{\sigma(q)\sigma(l)}^t, \gamma_{\sigma(q)\sigma(l)}^t\}_{1 \leq t \leq T, 1 \leq q \leq l \leq Q}).$$

It should be noted that here, the permutation σ acts *globally*, meaning that it is the same at each time point t. Now, if we let $\mathbb{P}_\theta^Y$ denote the marginal of $\mathbb{P}_\theta$ on the set of observations $\mathbf{Y}$, identifiability of the parameterisation, up to label switching means

$$\forall \theta, \tilde{\theta} \in \Theta, \quad \mathbb{P}_\theta^Y = \mathbb{P}_{\tilde{\theta}}^Y \implies \exists \sigma \in \mathfrak{S}_Q, \theta = \sigma(\tilde{\theta}).$$

Without additional constraints on the transition matrix $\boldsymbol{\pi}$ or on the parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, the parameters may not be recovered up to label switching. However, it could be that the static

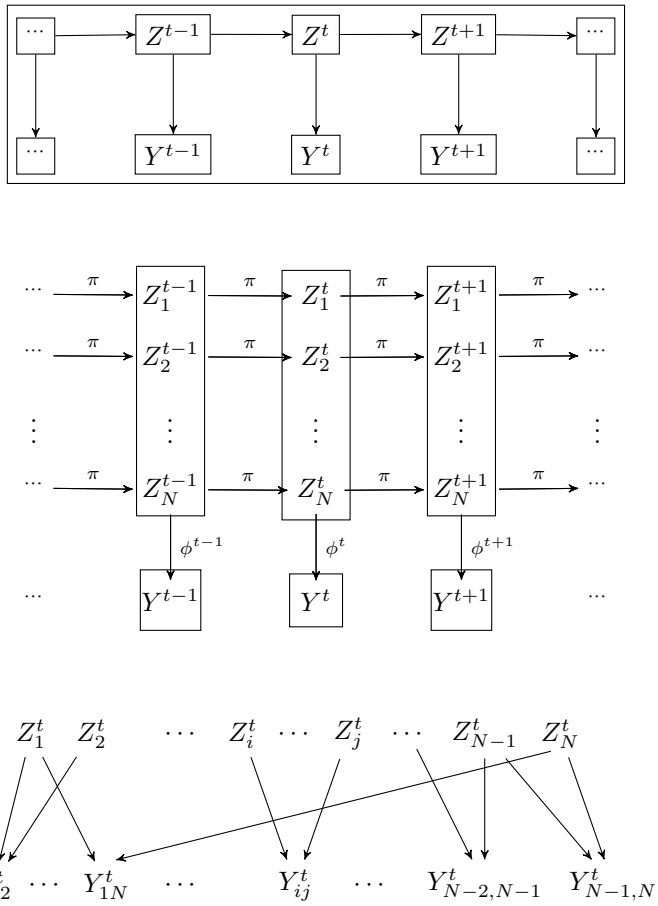

Figure 1. Dependency structures of the model. Top: general view corresponding to hidden Markov model (HMM) structure; Middle: details on latent structure organisation corresponding to N different iid Markov chains $Z_i = (Z_i^t)_{1 \leq t \leq T}$ across individuals; Bottom: details for fixed time point t corresponding to SBM structure.

SBM part of the parameter is recovered up to a *local* label switching. Local label switching on SBM part of the parameter is the weaker following property

$$\forall \theta, \tilde{\theta} \in \Theta, \quad \mathbb{P}_\theta^Y = \mathbb{P}_{\tilde{\theta}}^Y \implies \exists \sigma_1, \dots, \sigma_T \in \mathfrak{S}_Q^T, \forall t, (\beta^t, \gamma^t) = \sigma_t(\tilde{\beta}^t, \tilde{\gamma}^t).$$

This property is not satisfactory since clustering in models that only satisfy a local identifiability of SBM part of the parameter prevents from obtaining a picture of the evolution of the groups across time. We illustrate through a simple example the fact that if both Z^t and (β^t, γ^t) may vary through time, then the parameter can not be identified up to label switching, without additional constraints.

Example 1 (Non identifiability). *Let us consider the case of $Q = 2$ groups and (for simplicity of notation) $2T$ time points. We fix a first parameter value $\theta = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ defined by $\boldsymbol{\pi} = Id$ the size-two identity matrix and $(\beta^t, \gamma^t) = (\beta, \gamma)$ are chosen constant with t. In the following, we let $\phi_{ql}(\cdot)$ denote the constant (with time) conditional density distribution of any Y_{ij}^t given $Z_{iq}^t Z_{jl}^t = 1$, under parameter value θ. The latent process has stationary distribution $\boldsymbol{\alpha} = (1/2, 1/2)$ and since the latent configuration is drawn at the first time point and stays constant ($\boldsymbol{\pi}$ is the identity), it can be seen that the distribution on the set of observations $\mathbf{Y}$ is given by*

$$\begin{aligned}\mathbb{P}_\theta(\mathbf{Y}) &= \frac{1}{2^N} \sum_{q_1 \dots q_N \in \mathcal{Q}^N} \prod_{1 \le i < j \le N} \prod_{t=1}^{2T} \phi(Y_{ij}^t; \beta_{q_i q_j}^t, \gamma_{q_i q_j}^t) \\ &= \frac{1}{2^N} \sum_{q_1 \dots q_N \in \mathcal{Q}^N} \prod_{1 \le i < j \le N} \prod_{t=1}^{2T} \phi_{q_i q_j}(Y_{ij}^t).\end{aligned}$$

Now we consider a second parameter value $\tilde{\theta} = (\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$ such that

$$\tilde{\boldsymbol{\pi}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

which corresponds to the same latent stationary distribution $\boldsymbol{\alpha} = (1/2, 1/2)$ but now the latent configuration is drawn at the first time point and then each node switches group at each following time point. For any $q \in \{1, 2\}$, we let $\bar{q}$ denote the unique value such that $\{q, \bar{q}\} = \{1, 2\}$. Moreover, for any $q \in \{1, 2\}$, we set the intra group parameter at time $t = 1$ to $(\tilde{\beta}_{qq}^1, \tilde{\gamma}_{qq}^1) = (\beta_{qq}, \gamma_{qq})$, or equivalently, we set the conditional distribution $\tilde{\phi}_{qq}^1$ of Y_{ij}^1 given $Z_{iq}^1 Z_{jq}^1 = 1$, under parameter value $\tilde{\theta}$, equal to previous value ϕ_{qq}. Then, we switch the intra group parameters values at each time point by setting

$$\forall t \ge 1, \quad \tilde{\phi}_{11}^{t+1} = \tilde{\phi}_{22}^t \text{ and } \tilde{\phi}_{22}^{t+1} = \tilde{\phi}_{11}^t.$$

Finally, the inter group parameter is not modified through time and we set $\tilde{\phi}_{12}^t = \phi_{12}$. Now, we can write the distribution of $\mathbf{Y}$ under parameter value $\tilde{\theta}$

$$\begin{aligned}\mathbb{P}_{\tilde{\theta}}(\mathbf{Y}) &= \frac{1}{2^N} \sum_{q_1 \dots q_N \in \mathcal{Q}^N} \prod_{1 \le i < j \le N} \tilde{\phi}_{q_i q_j}^1(Y_{ij}^1) \tilde{\phi}_{\bar{q}_i \bar{q}_j}^2(Y_{ij}^2) \dots \tilde{\phi}_{q_i q_j}^{2T-1}(Y_{ij}^{2T-1}) \tilde{\phi}_{\bar{q}_i \bar{q}_j}^{2T}(Y_{ij}^{2T}) \\ &= \frac{1}{2^N} \sum_{q_1 \dots q_N \in \mathcal{Q}^N} \prod_{1 \le i < j \le N} \phi_{q_i q_j}(Y_{ij}^1) \phi_{q_i q_j}(Y_{ij}^2) \dots \phi_{q_i q_j}(Y_{ij}^{2T}),\end{aligned}$$

so that $\mathbb{P}_\theta^Y = \mathbb{P}_{\tilde\theta}^Y$. *To conclude, it suffices to show that there is no global permutation* $\sigma \in \mathfrak{S}_Q$ *such that* $\tilde\theta = \sigma(\theta)$. *This can be seen from the fact that for any* $\sigma \in \mathfrak{S}_Q$, *we have* $\sigma(\tilde{\boldsymbol{\pi}}) = \tilde{\boldsymbol{\pi}} \neq \boldsymbol{\pi}$. *Thus the two parameters* $\theta, \tilde\theta$ *are not equal up to label switching while they produce the same distribution on the observations. It follows that the parameter* θ *is not identifiable up to label switching. Note that the SBM part of the parameter is recovered up to local label switching as choosing the permutations* $\sigma_{2t} = Id$ *and* $\sigma_{2t-1} = (1,2)$ *(the transposition in* $\mathfrak{S}_2$*) for any* $1 \leq t \leq T$, *we obtain that* $\sigma_t(\beta^t, \gamma^t) = (\tilde\beta^t, \tilde\gamma^t)$.

The main problem with the previous example comes from the possibility of arbitrarily relabeling the groups between two time steps. As a consequence, we choose to impose the following constraints on the parameter θ

$$\forall q \in \mathcal{Q}, \forall t, t' \in \{1, \ldots, T\}, \quad \left\{ \begin{array}{ll} \text{Binary case:} & \beta_{qq}^t = \beta_{qq}^{t'}, \\ \text{Weighted case:} & \gamma_{qq}^t = \gamma_{qq}^{t'}. \end{array} \right. \tag{2}$$

In the following and when constrained to be constant (depending on binary or weighted case), these intra group parameters will be simply denoted by β_{qq} and γ_{qq}, respectively. We prove below that these constraints are sufficient to ensure identifiability of the parametrization under natural assumptions.

Assumption 1 (Weighted case). *We assume that*

i) For any $t \geq 1$, *the* $Q(Q+1)/2$ *values* $\{\gamma_{ql}^t, 1 \leq q \leq l \leq Q\}$ *are distinct,*

ii) The family of distributions $\mathcal{F} = \{f(\cdot, \gamma), \gamma \in \Gamma\}$ *is such that all elements* $f(\cdot, \gamma)$ *have no point mass at* 0 *and the parameters of finite mixtures of distributions in* $\mathcal{F}$ *are identifiable, up to label switching.*

Note that Assumption 1 does not impose any constraint on the sparsity parameters β_{ql}^t in the weighted case. In particular and for parsimony reasons, these may be chosen identical (to some β^t or some constant β) or set to two different values, e.g. $\beta_{qq}^t = \beta_{\text{in}}^t$ and $\beta_{ql}^t = \beta_{\text{out}}^t$ whenever $q \neq l$ at each time point (or even constant with time).

Proposition 1. *Considering the distribution* $\mathbb{P}_\theta^Y$ *on the set of observations and assuming the constraint* (2), *the parameter* $\theta = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ *satisfies the following:*

- ***Binary case:*** θ *is generically identified from* $\mathbb{P}_\theta^Y$, *up to label switching, as soon as* N *is not too small with respect to* Q,
- ***Weighted case:*** *Under additional Assumption 1, the parameter* θ *is identified from* $\mathbb{P}_\theta^Y$, *up to label switching, as soon as* $N \geq 3$.

Generic identifiability means 'up to excluding a subset of zero Lebesgue measure of the parameter set'. We refer to Allman et al. (2009, 2011) for more details. In particular, assuming that the Bernoulli parameters β_{ql} are distinct in the binary case is a generic constraint (meaning that it removes a subset of zero Lebesgue measure of the parameter set). As we do not specify the whole generic constraint that is needed here, we do not stress that one either. But the reader should have it in mind in the binary setup. Finally, note that the condition on the number of nodes N being not too small in the binary case is given precisely in Theorem 2 from Allman et al. (2011). The particular affiliation case is not covered by these results and discussed in the next section.

PROOF. The proof combines the approaches of Leroux (1992) for proving identifiability of hidden Markov models (HMM) parameters and Allman et al. (2011) that studies identifiability for (static) SBM.

First, we fix a time point $t \geq 1$ and consider the marginal distribution $\mathbb{P}_\theta(Y^t)$. According to Theorems 1,2 (binary case with $Q = 2$ and $Q \geq 3$, respectively) and Theorem 12 (weighted case) in Allman et al. (2011) on parameters identifiability in static SBM, there exists a permutation σ_t on the group labels $\mathcal{Q}$ such that we can identify (β^t, γ^t) as well as the marginal distribution $\boldsymbol{\alpha}$, up to this permutation. This result stands generically in the binary case only.

Now, for two different time points t, t', we use the constraint (2) and the assumption of distinct parameter values in order to identify the parameters $\{(\beta^t, \gamma^t), t \geq 1\}$ up to a (common) permutation σ on $\mathcal{Q}$. Indeed, in the binary case, assuming that the intra groups Bernoulli parameters satisfy $\beta^t_{qq} = \beta^{t'}_{qq}$ and that the set $\{\beta^t_{qq}; 1 \leq q \leq Q\}$ contains Q distinct values (a generic constraint) suffices to obtain a global permutation σ, not depending on time t, up to which $\{(\beta^t, \gamma^t), t \geq 1\}$ are identified. The same applies in the weighted case, by assuming equality between the parameter $\gamma^t_{qq} = \gamma^{t'}_{qq}$ for any t, t'.

It remains to identify the transition matrix $\boldsymbol{\pi}$ (up to the same permutation σ). We fix an edge (i, j) and following Leroux (1992), consider the bivariate distribution $\mathbb{P}_\theta(Y^t_{ij}, Y^{t+1}_{ij})$. This is given by

$$\mathbb{P}_\theta(Y^t_{ij}, Y^{t+1}_{ij}) = \sum_{q_1,q_2,l_1,l_2 \in \mathcal{Q}} \alpha_{q_1}\alpha_{l_1}\pi_{q_1q_2}\pi_{l_1l_2}\phi^t_{q_1l_1}(Y^t_{ij})\phi^{t+1}_{q_2l_2}(Y^{t+1}_{ij}). \tag{3}$$

Note that Teicher (1967) has proved the equivalence between parameters identifiability of the mixtures of a family of distributions and parameters identifiability of the mixtures of finite products from this same family. For the sake of clarity, we develop his proof adapted to our context. We thus write

$$\mathbb{P}_\theta(Y^t_{ij}, Y^{t+1}_{ij}) = \sum_{q_2,l_2 \in \mathcal{Q}} \Big(\sum_{q_1,l_1 \in \mathcal{Q}} \alpha_{q_1}\alpha_{l_1}\pi_{q_1q_2}\pi_{l_1l_2}\phi^t_{q_1l_1}(Y^t_{ij}) \Big) \phi^{t+1}_{q_2l_2}(Y^{t+1}_{ij}).$$

As the mixtures from the family $\{\phi^{t+1}_{ql}, 1 \leq q \leq l \leq Q\}$ have identifiable parameters (Assumption 1, iii)), we can identify the mixing distribution

$$\sum_{q_2,l_2 \in \mathcal{Q}} \Big(\sum_{q_1,l_1 \in \mathcal{Q}} \alpha_{q_1}\alpha_{l_1}\pi_{q_1q_2}\pi_{l_1l_2}\phi^t_{q_1l_1}(Y^t_{ij}) \Big) \delta_{(\beta^{t+1}_{q_2l_2}, \gamma^{t+1}_{q_2l_2})}.$$

Now, applying again this identifiability at time t and constraint (1), we may identify the whole mixing distribution

$$\sum_{q_2,l_2 \in \mathcal{Q}} \sum_{q_1,l_1 \in \mathcal{Q}} \alpha_{q_1}\alpha_{l_1}\pi_{q_1q_2}\pi_{l_1l_2} \delta_{(\beta^t_{q_1l_1}, \gamma^t_{q_1l_1})} \otimes \delta_{(\beta^{t+1}_{q_2l_2}, \gamma^{t+1}_{q_2,l_2})}.$$

This proves that the mixture given by (3) has identifiable components. From this mixture and the fact that we already identified the parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ up to a global permutation, we may extract the set of coefficients $\{\alpha_q^2\pi^2_{qq'}, 1 \leq q, q' \leq Q\}$ that corresponds to the components $\phi^t_{qq}\phi^{t+1}_{q'q'}$ in (3). As we also already obtained the values $\{\alpha_q, 1 \leq q \leq Q\}$, this now identifies the parameters $\{\pi_{qq'}, 1 \leq q, q' \leq Q\}$. This concludes the proof.

2.3. Discussing identifiability in affiliation case

Note that the affiliation setup is excluded from our previous results. Identifying the whole parameters from a binary affiliation SBM is a difficult task, as may be seen for instance by the many different but always partial results obtained by Allman et al. (2011). In their Corollary 7, the authors establish that *when group proportions are known*, the parameters $\beta_{\text{in}}(:= \beta_{qq} \text{ for all } q)$ and $\beta_{\text{out}}(:= \beta_{ql} \text{ for all } q \neq l)$ of a binary affiliation static SBM are identifiable. In the weighted affiliation case, all parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}^t, \boldsymbol{\gamma}^t)$ of a (static) SBM may be identified (Theorem 13 in Allman et al., 2011). Following the proof of Proposition 1, we could identify $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ in dynamic affiliation SBM under natural assumptions. Now, without an additional constraint on the transition matrix $\boldsymbol{\pi}$, it is hopeless to identify the transition parameters. Indeed, as the groups play similar roles at each time step, label switching between different time steps is free to occur and $\boldsymbol{\pi}$ may not be identified (note that assuming that β^t_{in} or γ^t_{in} does not depend on t is of no help here). This may be seen for instance from Example 1 that remains valid in the affiliation case. In fact, identifying $\boldsymbol{\pi}$ in dynamic affiliation SBM seems to be as hard as identifying the group proportions in static binary affiliation SBM. While static affiliation often relies on an assumption of equal group proportions, there is no simple parallel situation for the transition matrix $\boldsymbol{\pi}$ in the dynamic case (the trivial assumption $\boldsymbol{\pi} = Id$ is far too constrained). Let us now give some intuition on why $\boldsymbol{\pi}$ is difficult to recover. For instance, following the proof of Proposition 1 and looking at the distribution of (Y^t_{ij}, Y^{t+1}_{ij}) enables us to identify a mixing distribution with four components as follows. Let δ^t_{in} (resp. δ^t_{out}) be a shorthand for the Dirac mass at parameter $(\beta^t_{\text{in}}, \gamma^t_{\text{in}})$ (resp. $(\beta^t_{\text{out}}, \gamma^t_{\text{out}})$). From the distribution of (Y^t_{ij}, Y^{t+1}_{ij}), we identify the four following components

$$\Big(\sum_{qq'} \alpha_q^2 \pi_{qq'}^2\Big)\delta^t_{\text{in}} \otimes \delta^{t+1}_{\text{in}}; \Big(\sum_{q}\sum_{l\neq m} \alpha_q^2 \pi_{ql}\pi_{qm}\Big)\delta^t_{\text{in}} \otimes \delta^{t+1}_{\text{out}}; \Big(\sum_{q\neq l}\sum_{m} \alpha_q\alpha_l \pi_{qm}\pi_{lm}\Big)\delta^t_{\text{out}} \otimes \delta^{t+1}_{\text{in}};$$
$$\Big(\sum_{q\neq l}\sum_{q'\neq l'} \alpha_q\alpha_l \pi_{qq'}\pi_{ll'}\Big)\delta^t_{\text{out}} \otimes \delta^{t+1}_{\text{out}}.$$

Now relying on the knowledge of the proportions of each of these four components, it can be seen that it is not easy to identify the individual values of $\boldsymbol{\pi}$. Without a proper identification of the transition matrix $\boldsymbol{\pi}$, we do not recover the behaviour of the group membership through time. Empirical evidence for label switching between time steps in the affiliation setup is given in Section 4.

3. Inference algorithm

3.1. General description

As usual with latent variables, the log-likelihood $\log \mathbb{P}_\theta(\mathbf{Y})$ contains a sum over all possible latent configurations $\mathbf{Z}$ and thus may not be computed except for small values of N and T. A classical solution is to rely on expectation-maximisation (`EM`) algorithm (Dempster et al., 1977), an iterative procedure that finds local maxima of the log-likelihood. The use of `EM` algorithm relies on the computation of the conditional distribution of the latent variables $\mathbf{Z}$ given the observed ones $\mathbf{Y}$. However in the context of stochastic blockmodel, this distribution has not a factored form and thus may not be computed efficiently. We choose to rely here on variational approximations of `EM` algorithm (`VEM`, see for instance Jordan et al., 1999). These approximations have been first proposed in the context of SBM in Daudin

et al. (2008) and later developed in many directions, such as online procedures (Zanghi et al., 2008, 2010) or Bayesian `VEM` (Latouche et al., 2012). We refer to the review by Matias and Robin (2014) for more details about `VEM` algorithm (in particular a presentation of `EM` viewed as a special instance of `VEM`) and its comparison to other estimation procedures in SBM. Note that convergence properties of `VEM` algorithms are discussed in full generality in Gunawardana and Byrne (2005) and in the special case of SBM in Celisse et al. (2012); Bickel et al. (2013).

`VEM` for dynamic SBM. In our context of dynamic random graphs, we start by writing the complete data log-likelihood of the model

$$\begin{aligned}\log \mathbb{P}_\theta(\mathbf{Y},\mathbf{Z}) = &\sum_{i=1}^{N}\sum_{q=1}^{Q} Z_{iq}^1 \log \alpha_q + \sum_{t=2}^{T}\sum_{i=1}^{N}\sum_{1\le q,q'\le Q} Z_{iq}^{t-1} Z_{iq'}^t \log \pi_{qq'} \\ &+ \sum_{t=1}^{T}\sum_{1\le i<j\le N}\sum_{1\le q,l\le Q} Z_{iq}^t Z_{jl}^t \log \phi(Y_{ij}^t;\beta_{ql}^t,\gamma_{ql}^t). \end{aligned} \tag{4}$$

We now explore the dependency structure of the conditional distribution $\mathbb{P}_\theta(\mathbf{Z}|\mathbf{Y})$. First, note that it can be easily deduced from the DAG of the model (Figure 1, top) that

$$\mathbb{P}_\theta(\mathbf{Z}|\mathbf{Y}) = \mathbb{P}_\theta(Z^1|Y^1)\prod_{t=2}^{T}\mathbb{P}_\theta(Z^t|Z^{t-1},Y^t).$$

However, the distribution $\mathbb{P}_\theta(Z^t|Z^{t-1},Y^t) = \mathbb{P}_\theta((Z_i^t)_{1\le i\le N}|Z^{t-1},Y^t)$ can not be further factored. Indeed, for any $i\neq j$, the variables Z_i^t, Z_j^t are not independent when conditioned on Y^t. Our variational approximation naturally considers the following class of probability distributions $\mathbb{Q} := \mathbb{Q}_\tau$ parameterised by τ

$$\begin{aligned}\mathbb{Q}_\tau(\mathbf{Z}) &= \prod_{i=1}^{N}\mathbb{Q}_\tau(Z_i) = \prod_{i=1}^{N}\mathbb{Q}_\tau(Z_i^1)\prod_{t=2}^{T}\mathbb{Q}_\tau(Z_i^t|Z_i^{t-1}) \\ &= \prod_{i=1}^{N}\Big[\prod_{q=1}^{Q}\tau(i,q)^{Z_{iq}^1}\Big] \times \prod_{t=2}^{T}\prod_{1\le q,q'\le Q}\tau(t,i,q,q')^{Z_{iq}^{t-1}Z_{iq'}^t},\end{aligned}$$

where for any values (t,i,q,q'), we have $\tau(i,q)$ and $\tau(t,i,q,q')$ both belong to the set $[0,1]$ and are constrained by $\sum_q \tau(i,q) = 1$ and $\sum_{q'}\tau(t,i,q,q') = 1$. This class of probability distributions $\mathbb{Q}_\tau$ corresponds to considering independent laws through individuals, while for each $i\in\{1,\dots,N\}$, the distribution of Z_i under $\mathbb{Q}_\tau$ is the one of a Markov chain (through time t), with (inhomogeneous) transition $\tau(t,i,q,q') = \mathbb{Q}_\tau(Z_i^t = q'|Z_i^{t-1} = q)$ and initial distribution $\tau(i,q) = \mathbb{Q}_\tau(Z_i^1 = q)$.

We also need to consider marginal components of $\mathbb{Q}_\tau$, namely $\tau_{\mathrm{marg}}(t,i,q) := \mathbb{Q}_\tau(Z_i^t = q)$. These quantities are computed recursively by

$$\tau_{\mathrm{marg}}(1,i,q) = \tau(i,q) \text{ and } \forall t\ge 2,\ \tau_{\mathrm{marg}}(t,i,q) = \sum_{q'=1}^{Q}\tau_{\mathrm{marg}}(t-1,i,q')\tau(t,i,q',q).$$

Note also that all these values $\tau_{\text{marg}}(t,i,q)$ depend on the initial distribution $\tau(i,q)$. Entropy of this distribution is denoted by $\mathcal{H}(\mathbb{Q}_\tau)$. Using this class of probability distributions on $\mathcal{Q}^{\mathbb{N}}$, `VEM` algorithm is an iterative procedure to optimize the following criterion

$$\begin{aligned}
J(\theta,\tau) :=& \mathbb{E}_{\mathbb{Q}_\tau}(\log \mathbb{P}_\theta(\mathbf{Y},\mathbf{Z})) + \mathcal{H}(\mathbb{Q}_\tau) \\
=& \sum_{i=1}^{N}\sum_{q=1}^{Q} \tau(i,q)[\log\alpha_q - \log\tau(i,q)] \\
&+ \sum_{t=2}^{T}\sum_{i=1}^{N}\sum_{1\le q,q'\le Q} \tau_{\text{marg}}(t-1,i,q)\tau(t,i,q,q')[\log\pi_{qq'} - \log\tau(t,i,q,q')] \\
&+ \sum_{t=1}^{T}\sum_{1\le i<j\le N}\sum_{1\le q,l\le Q} \tau_{\text{marg}}(t,i,q)\tau_{\text{marg}}(t,j,l)\log\phi^t_{ql}(Y^t_{ij}). \qquad (5)
\end{aligned}$$

It consists in iterating the following two steps. At k-th iteration, with current parameter value $(\tau^{(k)},\theta^{(k)})$, we do

- `VE`-step: Compute $\tau^{(k+1)} = \text{Argmax}_\tau J(\theta^{(k)},\tau)$,
- `M`-step: Compute $\theta^{(k+1)} = \text{Argmax}_\theta J(\theta,\tau^{(k+1)})$.

PROPOSITION 2. *The value $\hat\tau$ that maximizes in τ the function $J(\theta,\tau)$ satisfies the fixed point equation*

$$\forall t\ge 2, \forall i\ge 1, \forall q,q'\in\mathcal{Q}, \quad \hat\tau(t,i,q,q') \propto \pi_{qq'}\prod_{j,j\neq i}\prod_{l'=1}^{Q}[\phi^t_{q'l'}(Y^t_{ij})]^{\hat\tau_{marg}(t,j,l')},$$

where $\propto$ means 'proportional to' (the constants are obtained by the constraints on τ). Moreover, the value $(\hat{\boldsymbol\pi},\hat{\boldsymbol\beta})$ that maximizes in $(\boldsymbol\pi,\boldsymbol\beta)$ the function $J(\theta,\tau)$ satisfies

$$\begin{aligned}
\forall (q,q')\in\mathcal{Q}^2, \quad \hat\pi_{qq'} &\propto \sum_{t=2}^{T}\sum_{i=1}^{N}\tau_{marg}(t-1,i,q)\tau(t,i,q,q'), \\
\forall t, \forall q\neq l\in\mathcal{Q}^2, \quad \hat\beta^t_{ql} &= \frac{\sum_{1\le i\neq j\le N}\tau_{marg}(t,i,q)\tau_{marg}(t,j,l)1_{Y^t_{ij}\neq 0}}{\sum_{i,j}\tau_{marg}(t,i,q)\tau_{marg}(t,j,l)}, \\
\forall q\in\mathcal{Q}, \quad \hat\beta_{qq} &= \frac{\sum_t\sum_{1\le i\neq j\le N}\tau_{marg}(t,i,q)\tau_{marg}(t,j,q)1_{Y^t_{ij}\neq 0}}{\sum_{t,i,j}\tau_{marg}(t,i,q)\tau_{marg}(t,j,q)}.
\end{aligned}$$

The proof of this result is immediate and omitted. Note that we have given a formula with constant (through time) values β_{qq} for any group $q\in\mathcal{Q}$. While this assumption is an identifiability requirement in the binary setup, it is not necessary in the weighted case. In this latter case, we use it only for parsimony reasons. The corresponding formula when this parameter is not assumed to be constant may be easily obtained.

To complete the algorithm's description, we provide equations to update the parameters $\tau(i,q),\alpha_q$ of initial distributions as well as the connectivity parameter $\boldsymbol\gamma$. First, optimization of $J(\theta,\tau)$ with respect to the initialization parameters $\tau(i,q)$ is a little bit more involved.

By neglecting the dependency on $\tau(i,q)$ of some terms appearing in criterion J, we choose to update this value by solving the fixed point equation

$$\forall i \geq 1, \forall q \in \mathcal{Q}, \quad \hat{\tau}(i,q) \propto \alpha_q \prod_{j,j\neq i} \prod_{l=1}^{Q} \phi_{ql}^1(Y_{ij}^1)^{\hat{\tau}(j,l)}. \tag{6}$$

Our experiments show that this is a reasonable approximation (Section 4). For the sake of completeness, we provide in Appendix A the exact equation satisfied by the solution.

Now parameter $\boldsymbol{\alpha}$ is not obtained from maximising J as it is not a free parameter but rather the stationary distribution associated with transition $\boldsymbol{\pi}$. Thus, $\boldsymbol{\alpha}$ is obtained from the empirical mean of the marginal distribution $\hat{\tau}_{\text{marg}}$ over all data points

$$\forall q \in \mathcal{Q}, \quad \hat{\alpha}_q = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \hat{\tau}_{\text{marg}}(t,i,q).$$

Finally, optimization with respect to $\boldsymbol{\gamma}$ depends on the choice of the parametric family $\{f(\cdot,\gamma), \gamma \in \Gamma\}$. We give some examples below in Section 3.2.

REMARK 1. *Performing* **EM** *algorithm in HMM (Figure 1, top) requires the use of forward-backward equations in order to deal with transition terms $Z_{iq}^{t-1} Z_{iq}^t$ appearing in the complete data log-likelihood* (4)*. In our setup, forward-backward equations are useless and replaced by a variational approximation. Indeed, it can be seen from Figure 1, middle, that the conditional distribution of $Z_{iq}^{t-1} Z_{iq}^t$ given the data can not be computed exactly through such forward-backward equations. This is due to the fact that the set of variables Y^t depend on all hidden ones $Z_1^t, \ldots, Z_N^t$ and focusing only on Z_i^t is not sufficient to determine its distribution.*

REMARK 2. *In Yang et al. (2011), the authors derive a* **VEM** *procedure in a similar (slightly less general) setup, but their variational approximation uses independent marginals (through individuals and also time points). As a consequence, the* **VE***-step that they derive is more involved than ours (see Section 4 in Yang et al., 2011).*

Model selection. Model selection on the number of groups Q is an important step. In case of latent variables, when the true data likelihood may not be easily computed, model selection may be done by maximizing an integrated classification likelihood (ICL) criterion (Biernacki et al., 2000). For any number of groups $Q \geq 1$, let $\hat{\theta}_Q$ be the estimated parameter value with Q groups and $\hat{\mathbf{Z}}$ the corresponding maximum a posteriori (MAP) classification at $\hat{\theta}_Q$. In our case, the general form of ICL is given by

$$ICL(Q) = \log \mathbb{P}_{\hat{\theta}_Q}(\mathbf{Y}, \hat{\mathbf{Z}}) - \frac{1}{2} Q(Q-1) \log(NT) - pen(N, \boldsymbol{\beta}, \boldsymbol{\gamma}), \tag{7}$$

where the first penalization term accounts for transition matrix $\boldsymbol{\pi}$ and $pen(N, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is a penalizing term for the connectivity parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. As the number of parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ depends on the specific form of the family $\{f(\cdot;\gamma), \gamma \in \Gamma\}$, we provide context dependent expressions for ICL in the next section. Note that the first penalization term accounts for NT observations while the number of observations corresponding to SBM part of the parameter in $pen(N, \boldsymbol{\beta}, \boldsymbol{\gamma})$ will be different. We refer to Daudin et al. (2008) for an expression of ICL in static SBM that shows an analogous difference in penalizing groups proportions or connectivity parameters.

3.2. Estimation of γ and model selection: specific examples

As previously said, the M-step equations concerning $\boldsymbol{\gamma}$ differ depending on the specific choice of the parametric family $\{f(\cdot,\gamma),\gamma\in\Gamma\}$. We provide here many examples of classical choices for these parametric families. Remember that the resulting conditional distribution on the observations is a mixture between an element from this family and the Dirac mass at zero. We also provide expressions for ICL criterion in these different setups.

EXAMPLE 2 (BINARY CASE). *This specific case corresponds to a degenerate family with only one element, the Dirac mass at 1, namely $F(y,\gamma)=\delta_1(y)$. The parameter θ reduces to $(\boldsymbol{\pi},\boldsymbol{\beta})$ for which updating expressions at the M-step have already been given (see Proposition 2). Note that we imposed the constraint β^t_{qq} constant with respect to t, for any $q\in\mathcal{Q}$. Now, model selection is performed through (7) where*

$$pen(N,\boldsymbol{\beta},\boldsymbol{\gamma})=pen(N,\boldsymbol{\beta})=\frac{1}{2}Q\log\Big(\frac{N(N-1)T}{2}\Big)+\frac{1}{2}\frac{Q(Q-1)}{2}T\log\Big(\frac{N(N-1)}{2}\Big).$$

EXAMPLE 3 (FINITE CASE). *Let us consider a finite set of $M\geq 2$ known values $\{a_1,\ldots,a_M\}$ not containing 0 and*

$$f(y,\gamma)=\sum_{m=1}^{M}\gamma(m)1_{y=a_m},$$

with $\gamma(m)\geq 0$ and $\sum_m\gamma(m)=1$. The value $\hat{\boldsymbol{\gamma}}$ that maximizes $J(\theta,\tau)$ with respect to $\boldsymbol{\gamma}$ is given by

$$\forall t,\forall q\neq l\in\mathcal{Q}^2,\forall m,\quad \hat{\gamma}^t_{ql}(m)=\frac{\sum_{1\leq i\neq j\leq N}\tau_{marg}(t,i,q)\tau_{marg}(t,j,l)1_{Y^t_{ij}=a_m}}{\sum_{m,i,j}\tau_{marg}(t,i,q)\tau_{marg}(t,j,l)1_{Y^t_{ij}=a_m}},$$

$$\forall q\in\mathcal{Q},\forall m,\quad \hat{\gamma}_{qq}(m)=\frac{\sum_{t=1}^{T}\sum_{1\leq i\neq j\leq N}\tau_{marg}(t,i,q)\tau_{marg}(t,j,q)1_{Y^t_{ij}=a_m}}{\sum_{m,t,i,j}\tau_{marg}(t,i,q)\tau_{marg}(t,j,q)1_{Y^t_{ij}=a_m}}.$$

These equations remain valid when considering a set of disjoint bins $\{I_m\}_m$ instead of pointwise values $\{a_m\}_m$.

In this setup, we do not propose a model selection criterion for selecting the number of groups Q. Indeed, our investigations show that a competition occurs between the number of bins M and the number of groups Q, so that in general we end up selecting only $Q=2$ groups because of a large number of parameters (data not shown). In fact, this finite distribution setup may be viewed as a nonparametric model for which BIC-like criterion (ICL is of that type) are not suited. Section 5 proposes another approach to handle this case, relying on the 'elbow' method applied on the complete data log-likelihood.

EXAMPLE 4 (POISSON CASE). *We consider the truncated Poisson distribution*

$$f(y,\gamma)=(e^{\gamma}-1)^{-1}\frac{\gamma^y}{y!},\quad y\in\mathbb{N}\setminus\{0\},$$

resulting in either a 0-inflated or 0-deflated Poisson when mixed with the Dirac mass at 0. Let

$$\forall x>0,\quad \psi(x)=\frac{xe^x}{e^x-1},$$

which is a strictly increasing function and as such admits a unique inverse function $\psi^{(-1)}$ *on* $(1, +\infty)$. *Note that* $\psi^{(-1)}$ *has no simple analytic expression but is easily found numerically. In this case, the value* $\hat{\boldsymbol{\gamma}}$ *that maximizes* $J(\theta, \tau)$ *with respect to* $\boldsymbol{\gamma}$ *is given by*

$$\forall t, \forall q \neq l \in \mathcal{Q}^2, \quad \hat{\gamma}_{ql}^t = \psi^{(-1)}\Big(\frac{\sum_{1\leq i\neq j\leq N} \tau_{marg}(t,i,q)\tau_{marg}(t,j,l)Y_{ij}^t}{\sum_{i,j} \tau_{marg}(t,i,q)\tau_{marg}(t,j,l)1_{Y_{ij}^t\neq 0}}\Big),$$
$$\forall q \in \mathcal{Q}, \quad \hat{\gamma}_{qq} = \psi^{(-1)}\Big(\frac{\sum_{t=1}^T\sum_{1\leq i\neq j\leq N} \tau_{marg}(t,i,q)\tau_{marg}(t,j,q)Y_{ij}^t}{\sum_{t,i,j} \tau_{marg}(t,i,q)\tau_{marg}(t,j,q)1_{Y_{ij}^t\neq 0}}\Big).$$

Model selection is obtained by maximizing (7) *with*

$$pen(N, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\Big(|\{\beta_{qq}, q \in \mathcal{Q}\}| + Q\Big)\log\Big(\frac{N(N-1)T}{2}\Big) + \frac{1}{2}\Big(|\{\beta_{ql}^t, 1 \leq q < l \leq Q, 1 \leq t \leq T\}| + \frac{Q(Q-1)}{2}T\Big)\log\Big(\frac{N(N-1)}{2}\Big).$$

Moreover, if $\beta_{ql}^t = \beta_{out}$ *does not depend on* t *and* $\beta_{qq} = \beta_{in}$, *the penalty term in ICL becomes*

$$pen(N, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\Big(2 + Q\Big)\log\Big(\frac{N(N-1)T}{2}\Big) + \frac{1}{2}\Big(\frac{Q(Q-1)}{2}T\Big)\log\Big(\frac{N(N-1)}{2}\Big).$$

Example 5 (Gaussian homoscedastic case). *Let us consider the Gaussian distribution*

$$f(y, \gamma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(y-\mu)^2}{2\sigma^2}\Big),$$

where $\gamma = (\mu, \sigma^2) \in \mathbb{R} \times (0, +\infty)$. *For parsimony reasons, we choose to consider the homoscedastic case where the variance is constant across groups and simply denoted by* σ_t^2. *The value* $\hat{\boldsymbol{\gamma}}$ *that maximizes* $J(\theta, \tau)$ *with respect to* $\boldsymbol{\gamma}$ *is given by*

$$\forall t, \forall q \neq l \in \mathcal{Q}^2, \quad \hat{\mu}_{ql}^t = \frac{\sum_{1\leq i\neq j\leq N} \tau_{marg}(t,i,q)\tau_{marg}(t,j,l)Y_{ij}^t}{\sum_{i,j} \tau_{marg}(t,i,q)\tau_{marg}(t,j,l)1_{Y_{ij}^t\neq 0}},$$
$$\forall q \in \mathcal{Q}, \quad \hat{\mu}_{qq} = \frac{\sum_{t=1}^T\sum_{1\leq i\neq j\leq N} \tau_{marg}(t,i,q)\tau_{marg}(t,j,q)Y_{ij}^t}{\sum_{t,i,j} \tau_{marg}(t,i,q)\tau_{marg}(t,j,q)1_{Y_{ij}^t\neq 0}},$$
$$\textit{and } \forall t, \quad \hat{\sigma}_t^2 = \frac{\sum_{1\leq i<j\leq N}\sum_{1\leq q,l\leq Q} \tau_{marg}(t,i,q)\tau_{marg}(t,j,l)[Y_{ij}^t - \hat{\mu}_{ql}^t]^2 1_{Y_{ij}^t\neq 0}}{\sum_{i,j,q,l} \tau_{marg}(t,i,q)\tau_{marg}(t,j,l)1_{Y_{ij}^t\neq 0}}.$$

Here the remaining penalty term in (7) *for ICL criterion writes*

$$\begin{aligned} pen(N, \boldsymbol{\beta}, \boldsymbol{\gamma}) &= \frac{1}{2}(2Q)\log\Big(\frac{N(N-1)T}{2}\Big) + \frac{1}{2}\Big(2\frac{Q(Q-1)}{2}T\Big)\log\Big(\frac{N(N-1)}{2}\Big) \\ &= Q\log\Big(\frac{N(N-1)T}{2}\Big) + \frac{Q(Q-1)}{2}T\log\Big(\frac{N(N-1)}{2}\Big). \end{aligned}$$

3.3. Algorithm initialization

All `EM` based procedures look for local maxima of their objective function and careful initialization is a key in their success. For static SBM, `VEM` procedures often rely on a `k-means`

algorithm on the adjacency matrix to obtain an initial clustering of the individuals. In our context, the dynamic aspect of the data needs to be properly handled. We choose to initialize our `VEM` procedure by running `k-means` on the rows of a concatenated data matrix containing all the adjacency time step matrices Y^t stacked in consecutive column blocks. As a result, our initial clustering of the individuals is constant across time (namely Z_i^t does not depend on t). A consequence of this choice is that this initialization works well when the groups memberships do not vary too much across time (see Section 4 where we explore different values of transition matrix $\boldsymbol{\pi}$). In practice, real datasets will either exhibit nodes that do not change group at all (see Section 5) or nodes that leave a group and then come back to this group. Our initialization is performant in these cases. Another consequence is that while we would expect the performances of the procedure to increase with the number T of time steps, we sometimes observe on the contrary a decrease in these performances. This is due to the fact that increasing T also increases the probability for an individual to change group at some point in time and thus starting with a constant in time clustering of the individuals, it becomes more difficult to correctly infer the groups membership at each time point (see in Section 4 the difference between results for $T = 5$ and $T = 10$).

To conclude this section, we mention that initialization is also a crucial point for other methods and we discuss in the next section its impact on the algorithm proposed in Yang et al. (2011).

4. Synthetic experiments

The methods presented in this manuscript are implemented into a R package and available at `http://lbbe.univ-lyon1.fr/dynsbm`. The package will soon be available on the CRAN.

4.1. Clustering performances

In this section, we explore the performances of our method for clustering the nodes across the different time steps. To this aim, we will consider two different criteria. We rely on the adjusted Rand index (ARI Hubert and Arabie, 1985) to evaluate the agreement between the estimated and the true latent structure. This index is smaller than 1, two identical latent structures (up to label switching) having an ARI equal to 1. Note that it can take negative values and is built on Rand index with a correction for chance. Now there are two different ways of using ARI in a dynamic setup. Following Yang et al. (2011); Xu and Hero (2014), we first consider an averaged value over the different time steps $1 \leq t \leq T$ of ARI_t computed at time t. In this approach the dynamic setup may be viewed as a way of improving the node clustering at each time step over a method that would cluster separately the nodes at each time step. However, this averaged index does not say anything about the smooth recovery of group memberships along time. In particular, it is invariant under local switching on SBM part of the parameter (see Section 2.2). Thus we also consider the global ARI value that compares the clustering of the set of nodes for all time points with the true latent structure. Obviously, good performances for this criteria are more difficult to obtain.

We use synthetic datasets created as follows. We consider binary graphs with $N = 100$ nodes and $T \in \{5; 10\}$ different time steps. We assume $Q = 2$ latent groups with three different values for the transition matrix $\boldsymbol{\pi}$

$$\boldsymbol{\pi}_{low} = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}; \boldsymbol{\pi}_{medium} = \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}; \boldsymbol{\pi}_{high} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}.$$

Table 1. Bernoulli parameter values in 4 different cases, plus an affiliation example.

Difficulty	β_{11}	β_{12}	β_{22}
low-	0.2	0.1	0.15
low+	0.25	0.1	0.2
medium-	0.3	0.1	0.2
medium+	0.4	0.1	0.2
med w/ affiliation	0.3	0.1	0.3

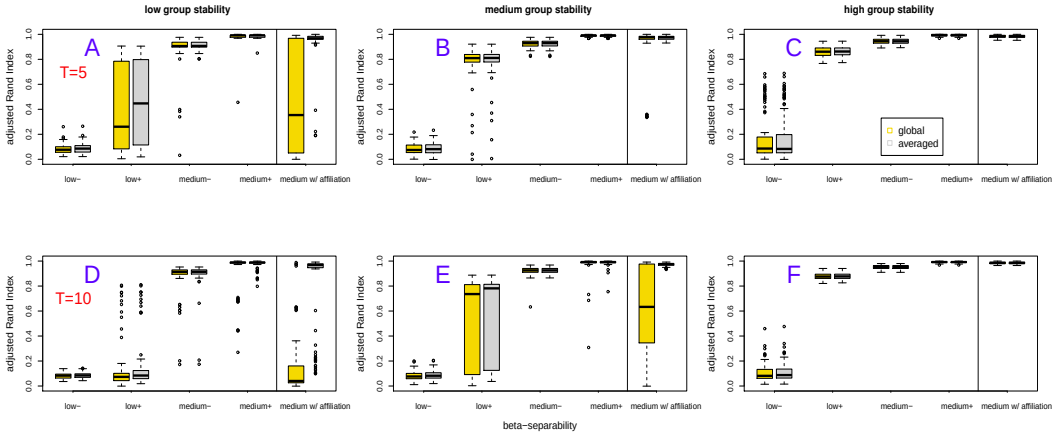

Figure 2. Boxplots of global ARI (gold, left) and averaged ARI (grey, right) in different setups. From left to right: the three panels correspond to $\boldsymbol{\pi} = \boldsymbol{\pi}_{low}$ (panels A,D), $\boldsymbol{\pi}_{medium}$ (panels B,E) and $\boldsymbol{\pi}_{high}$ (panels C,F), respectively. In each panel, from left to right: results corresponding to $\boldsymbol{\beta} = low-, low+, medium-, medium+$ and affiliation case, respectively. First row: $T = 5$ time points, second row: $T = 10$.

These three cases correspond respectively to *low, medium* and *high* group stability. Namely in the first case, individuals are more likely to change group across time, resulting in a more difficult problem from the point of view of the initialization of our algorithm (see Section 3.3). Note that the stationary distribution in those three cases is $\boldsymbol{\alpha} = (1/2, 1/2)$ so that the two groups have equal proportions. As for the Bernoulli parameters $\boldsymbol{\beta}$, we explore 4 different cases representing different difficulty levels, plus a specific example of affiliation for which we recall that parameters are not identifiable (it would otherwise correspond to a medium difficulty). The choice of parameters is given in Table 4.1.

For each combination of $(\boldsymbol{\pi}, \boldsymbol{\beta})$, we generate 100 datasets, estimate their parameters, cluster their nodes and report in Figure 2 boxplots of a global and of an averaged ARI value. Mean squared errors (MSE) for estimation of the transition parameter $\boldsymbol{\pi}$ are given in Figure 3. We only show MSE for $\boldsymbol{\pi}$ as the MSE for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are strongly correlated with the clustering results.

Figure 2 confirms that it is more difficult to obtain a smooth recovery of the groups (measured through global ARI) than a local one (measured through averaged ARI). In particular in the affiliation model, we observe that while the averaged ARI is rather good, the global one can be low. However in the identifiable cases, we obtain rather good performances for this global index when group stability is not too low or when connectivity parameters

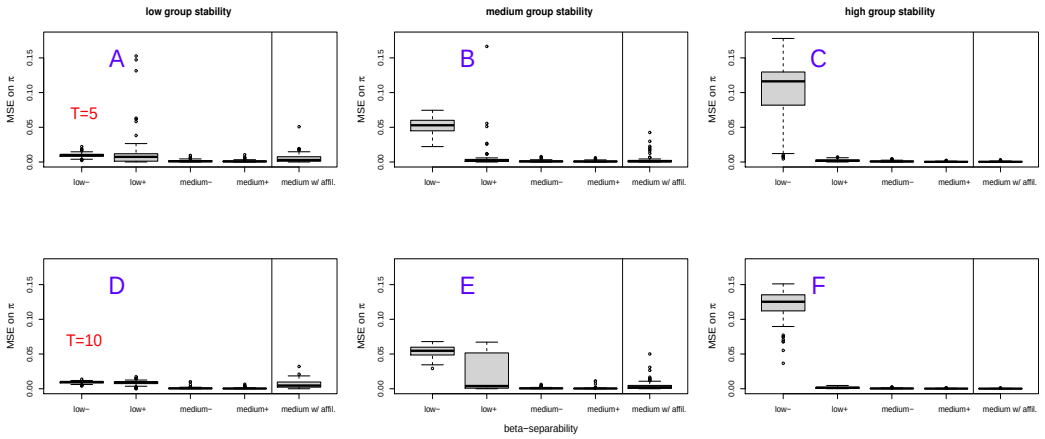

Figure 3. MSE for estimation of transition matrix $\boldsymbol{\pi}$ in different setups. From left to right: the three panels correspond to $\boldsymbol{\pi} = \boldsymbol{\pi}_{low}$ (panels A,D), $\boldsymbol{\pi}_{medium}$ (panels B,E) and $\boldsymbol{\pi}_{high}$ (panels C,F), respectively. In each panel, from left to right: results corresponding to $\boldsymbol{\beta} = low-, low+, medium-, medium+$ and affiliation case, respectively. First row: $T = 5$ time points, second row: $T = 10$.

are well enough separated. As expected, the clustering performances increase with group stability and with a better separation between the groups connectivity behaviors. When increasing the number of time points from 5 to 10, clustering indexes tend to be slightly better, exhibiting a smaller variance. However this is not always the case: for instance with low/medium group stability and $\beta = low+$, we observe that the performances decrease from 5 to 10 time points. We believe that this is due to the initialization of our procedure: with $T = 10$ time points, it is more likely that the groups membership differ from their initial value. As we use as a starting point a constant with time value for these membership, our algorithm is farther from the optimal value. Looking now at the MSE values for estimation of $\boldsymbol{\pi}$ (Figure 3), we observe that when groups are not globally recovered, the MSE values are higher. However in most of the cases, these MSE are rather small so that the dynamics of the groups membership is captured.

Now, we compare our results with other procedures. The models from Yang et al. (2011); Xu and Hero (2014) are the closest to our setup. Since Xu and Hero (2014) obtained comparable performances as the ones from Yang et al. (2011), we focus on the latter here. (In fact, Xu and Hero's method is faster, with slightly lower clustering performances than Yang et al.'s one.) Thus, we use the offline version of the algorithm proposed in Yang et al. (2011) (Matlab code is available on the web site of the first author). We ran their code on the same setup as above. When relying on default values of the algorithm, the results obtained are very poor, with ARI values smaller than 10^{-2} in general (data not shown). We note that the authors do not discuss initialization and simply propose to start with a random partition of the nodes, which proves to be a bad strategy. In order to make fair comparisons, we thus decided to combine their algorithm with our initialization strategy. Results are presented in Figure 4.

From these results, we can see that putting appart our initialization strategy, our procedure outperforms Yang et al.'s one. Indeed, the method obtains good performances only

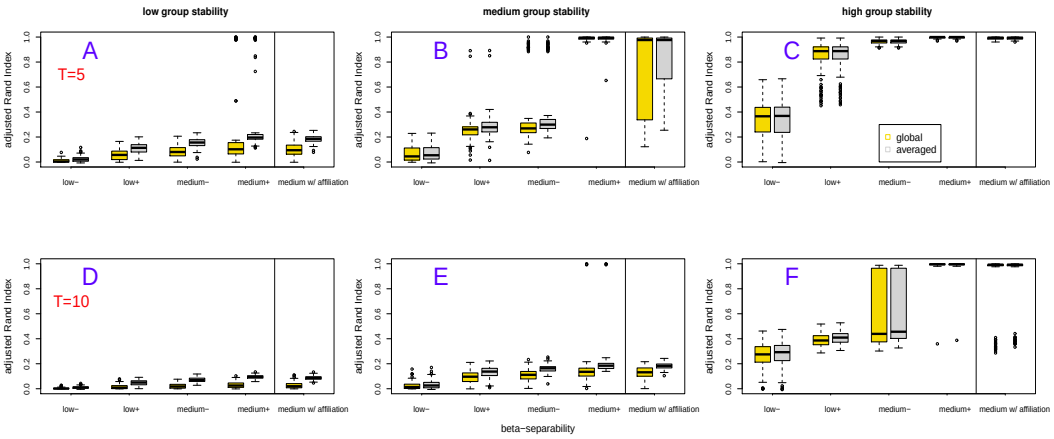

Figure 4. Boxplots of global ARI (gold, left) and averaged ARI (grey, right) in different setups for the combination of our initialization strategy with Yang et al.'s algorithm. From left to right: the three panels correspond to $\boldsymbol{\pi} = \boldsymbol{\pi}_{low}$ (panels A,D), $\boldsymbol{\pi}_{medium}$ (panels B,E) and $\boldsymbol{\pi}_{high}$ (panels C,F), respectively. In each panel, from left to right: results corresponding to $\boldsymbol{\beta} = low-, low+, medium-, medium+$ and affiliation case, respectively. First row: $T = 5$ time points, second row: $T = 10$.

in a few cases: $(\pi_{high}, \beta \in \{medium+;\ med\ w/\ affiliation\}, T \in \{5, 10\})$; $(\pi_{high}, \beta \in \{low+, medium-, T = 5)$ and $(\pi_{medium}, \beta \in \{medium+;\ med\ w/\ affiliation\}, T = 5)$. In all these cases, we can see that the method's performances are due to a very good initialization. Now, when the true classification is farther from initialization, the performances considerably drop. In particular, for intermediate cases (e.g. medium group stability or high group stability with $T = 10$), we can see that our method still succeeds in obtaining a good partition (Figure 2) while this is not the case for Yang et al.'s one (Figure 4).

4.2. Model selection

We simulate a binary dynamic dataset with $Q = 4$ groups, transition matrix between states satisfies $\pi_{qq} = 0.91$ and $\pi_{ql} = 0.03$ for $q \neq l$. Bernoulli parameters are chosen as follows: we draw i.i.d. random variables $\{\epsilon_{ql}\}_{1 \leq q \leq l \leq 4} \in [-1, 1]$ and then choose

$$\begin{aligned} \forall q \in \mathcal{Q}, \quad & \beta_{qq} = 0.4 + \epsilon_{qq} 0.1 \\ \forall q \neq l \in \mathcal{Q}^2, \quad & \beta_{ql} = 0.1 + \epsilon_{ql} 0.1 \end{aligned}$$

We generate 100 datasets under this model and estimate the number of groups relying on ICL criterion. Results are presented in Figure 5. We observe that the correct number of groups is recovered in 88% of the cases (left panel). Moreover, the right panel shows that when ICL selects only 3 groups, ARI of the classification with 4 groups is rather low (less than 80%). This shows that in those cases, classification with 4 groups is not the correct one, so that `VEM` algorithm seems responsible for bad results (optimum has not been reached) more than the penalization term.

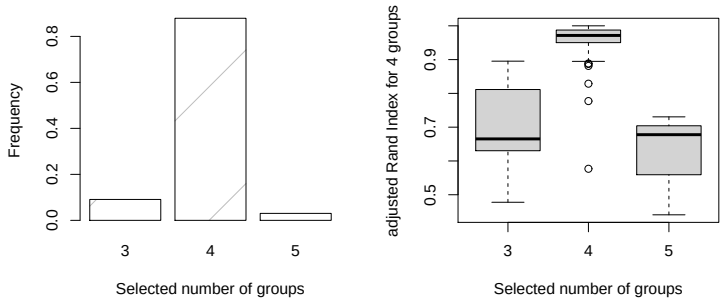

Figure 5. Estimation of the number of groups via ICL criterion. Left panel shows the frequency of the selected number of groups. Right panel shows ARI of the classification obtained with 4 groups depending on the selected number of groups.

5. Analyzing a real dataset

In this section, we illustrate our method through the analysis of a real data example.

The dataset consists in face-to-face encounters of high school students (measured through the use of wearable sensors) of a class from a French high school (see Fournet and Barrat, 2014, for a complete description of the experiment). In this class called 'PC' (as students focus on Physics and Chemistry), interactions were recorded during 4 days (Tuesday to Friday) in Dec. 2011. We kept only the 27 (out of 31) students that appear every day, i.e. that have at least one interaction with another student during each of the 4 days. Interaction times were aggregated by days to form a sequence of 4 different networks. These are undirected and weighted networks, the weight of an interaction between two individuals being the number of interactions between these 2 individuals divided by the number of time points for which at least two individuals interacted; thus a non negative real number that we call *interaction frequency*. After examination of the distribution of these weights, we choose to discretize these data into $M = 3$ bins (see Example 3) corresponding to *low, medium* and *high* interaction frequency. As already explained in Example 3, our model selection criterion is not fitted to this case. We thus choose to rely instead on the 'elbow' method, applied to the complete data log-likelihood. It consists in identifying a change of slope on the curve that represents this complete data log-likelihood for different values of Q. The method selects $Q = 4$ groups (see Figure 6) and we now present the results obtained with our model fitted with $Q = 4$ groups.

Figure 7 presents a summary of the estimators we obtain for interaction parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. In this figure, each of the 10 cells corresponds to a pair (q, l) with $1 \leq q \leq l \leq 4$. In each cell, there are $T = 4$ different colored barplots, each of them containing the proportions $\hat{\gamma}^t_{ql}(m)$ for $1 \leq m \leq 3$. Finally, the width of each barplot is proportional to the corresponding value of $\hat{\beta}^t_{ql}$. We recall that when considering the diagonal cells (q, q), parameters do not depend on t anymore. We observe that groups 2 and 3 are composed by students that are likely to interact together (i.e. $\hat{\beta}_{22}$ and $\hat{\beta}_{33}$ are close to 1). Furthermore, the frequency of their interactions inside their groups is higher than in the rest of the network ($\hat{\gamma}_{qq}(low) < \hat{\gamma}_{qq}(medium) < \hat{\gamma}_{qq}(high)$ for $q = 2, 3$). These two groups form two

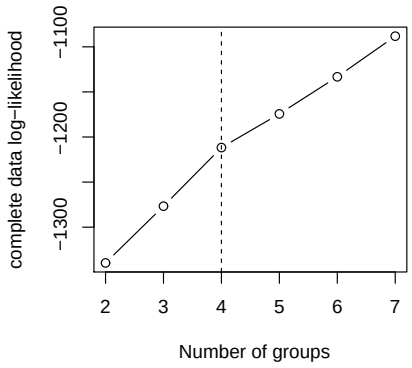

Figure 6. Complete data log-likelihood estimated for different numbers of groups on the dataset of interactions in the 'PC' class (Fournet and Barrat, 2014).

communities such as defined in Fortunato (2010). Moreover, we observe that both groups include a certain number of individuals (3 and 4 respectively) that permanently stay in the group over time (see Figure 8). These individuals may play the role of 'social attractors' or 'core leaders' around which the other students are likely to gravitate. Group 4 displays a similar pattern of community structure, with much less interaction (intermediate value of $\hat{\beta}_{44}$) but also a significant level of interaction with group 2. Interestingly, groups 2 and 4 also exchange students over time (see fluxes between groups in Figure 8) and this could reflect some cooperation or affinity between the students of these two groups. Group 1 is quite stable over time (7 permanent members, see Figure 8) and is characterized by a low rate of interactions inside and outside the group (Figure 7). It clearly gathers isolated students, but this does not mean that they do not interact with any student, they usually do so, but with a small number of partners. Therefore, we do not only decipher evolving communities (such as in Yang et al., 2011) but we also highlight the dynamics of aloneness inside this class.

We now investigate if gender differences may help in explaining or refining the interaction patterns that we reveal. We first note that group 3 is exclusively composed by male students: this observation along with the previous conclusions suggest that group 3 may be a closed/exclusive male-community. Meanwhile, some of these male students move to group 1 which is partly composed by a 'backbone' of female students that stay in group 1 (Figure 9). Moreover, we clearly observe that female students are likely to stay in their group (most of the moves between groups are realized by males, Figure 9) and that a majority of them are in low-interacting groups 1 and 4. But not any female student moves between these two groups, which supports a clear dichotomy pattern in the female organization with respect to male organization. In summary, we show evidence for some gender homophily (see Fournet and Barrat, 2014, for a precise definition), i.e. gender is a key factor for explaining the dynamics of the interactions between these young adults.

Lastly, we note that both information captured by our model (say $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) are often convergent/correlated in this case, but we note that studying this network with a binary model (i.e. not considering the interaction frequency) does not allow to capture interesting

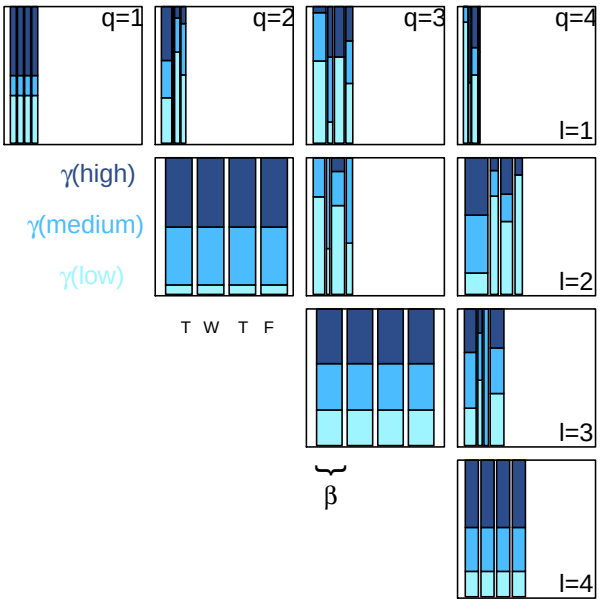

Figure 7. Summary of the interaction parameters $\hat{\beta}$ and $\hat{\gamma}$ estimated by our model with $Q = 4$ grous on the dataset of interactions in the 'PC' class (Fournet and Barrat, 2014). In each cell (q, l) with $1 \leq q \leq l \leq 4$, there are $T = 4$ barplots corresponding to the 4 measurements (Tuesday to Friday). Each barplot represents the distribution of the parameter γ_{ql}^t for the three categories of interaction frequency (low, $medium$ and $high$). The width of each barplot is proportional to the sparsity parameter β_{ql}^t.

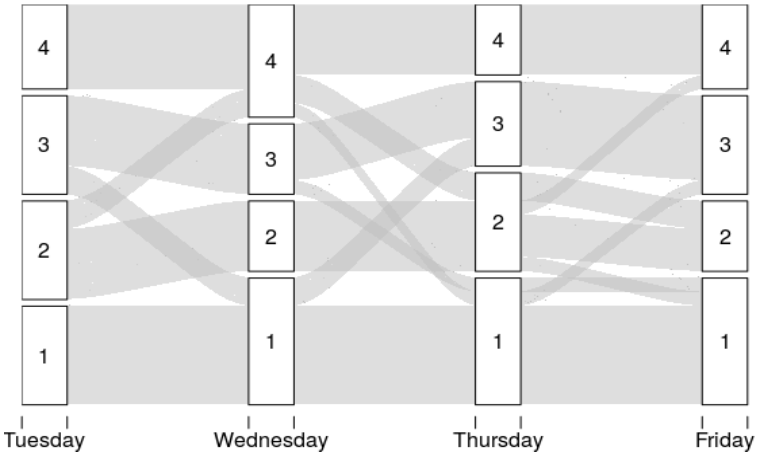

Figure 8. Alluvial plot showing the dynamics of the group membership estimated by our model on the dataset of interactions in the 'PC' class (Fournet and Barrat, 2014). Each line is a flux that represents the move of one or more students from a group to another group. The thickness of the lines is proportional to the number of students and the total height represents the 27 students.

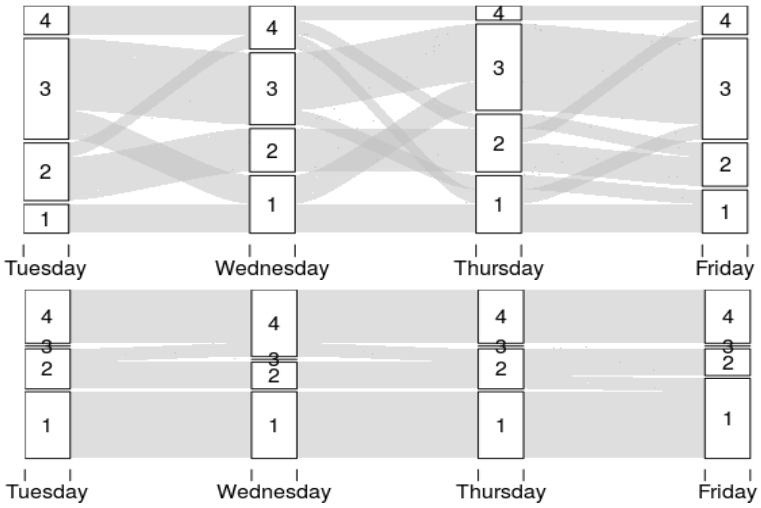

Figure 9. Same as in Figure 8 for the 15 male students (upper panel) and the 12 female students (lower panel).

structure (data not shown). Therefore, the presence/absence of an interaction as well as its frequency are important and require an explicit modelling such as in our approach.

6. Extensions

In the present work, we limited ourselves to the case where the list of nodes $\{1, \dots, N\}$ stays constant across time. However in real data applications it may happen that some actors enter or leave the study during the analysis. This may be handled in a simple way as follows. Let us consider $V = \{1, \dots, N\}$ as the total list of individuals and for each time step t, a subset V^t of V with cardinality N_t of actors are present. Data is formed by a series of adjacency matrices $\mathbf{Y} = (Y^t)_{1 \le t \le T}$ where each Y^t still has size $N \times N$. For all pair of present nodes $i, j \in V^t$, entry Y^t_{ij} characterizes the binary or weighted interaction between i, j while for any $i, j \in V$ such that $i \notin V^t$, entry Y^t_{ij} is set to 0. Now, we construct the latent process $\mathbf{Z} = (Z^t_i)_{1 \le t \le T, i \in V}$ on an extended set $\mathcal{Q}_a = \mathcal{Q} \cup \{a\}$ where the extra value a stands for *absent*. For each time step t and whenever $i \in V^t$, random variable Z^t_i is constrained to vary in $\mathcal{Q}$ while for any $i \notin V^t$ we fix $Z^t_i = a$. As previously, the random time series $(Z_i)_{i \in V}$ are supposed to be independent while for each individual $i \in V$, the sequence $Z_i = (Z^t_i)_{1 \le t \le T}$ forms an *inhomogeneous* Markov chain with values in $\mathcal{Q}_a$ and transitions $\boldsymbol{\pi}^t$ constrained by, for all $q, q' \in \mathcal{Q}$,

$$\begin{aligned}
\pi^t_{qa} &= \mathbb{P}(Z^t_i = a | Z^{t-1}_i = q) = 1\{i \notin V^t\}, \\
\pi^t_{aq} &= \mathbb{P}(Z^t_i = q | Z^{t-1}_i = a) = \alpha_q 1\{i \in V^t\}, \\
\pi^t_{qq'} &= \mathbb{P}(Z^t_i = q' | Z^{t-1}_i = q) = \pi_{qq'} 1\{i \in V^t\}.
\end{aligned}$$

Here, $\boldsymbol{\pi} = (\pi_{qq'})_{1 \le q, q' \le Q}$ stands as previously for a transition matrix on $\mathcal{Q}$ of an irreducible aperiodic stationary Markov chain with stationary distribution $\boldsymbol{\alpha}$. Note that the whole chain Z_i is not stationary anymore. The probability of any trajectory of the latent process

simply writes as

$$\mathbb{P}(\mathbf{Z}) = \prod_{i=1}^{N} \mathbb{P}(Z_i^1) \prod_{t=2}^{T} \mathbb{P}(Z_i^t | Z_i^{t-1}) = \prod_{q \in \mathcal{Q}} \alpha_q^{N_q} \times \prod_{q,q' \in \mathcal{Q}} \pi_{qq'}^{N_{qq'}},$$

where

$$N_q = \sum_{i \in V^1} 1\{Z_i^1 = q\} + \sum_{t=2}^{T} \sum_{i \in V^t, i \notin V^{t-1}} 1\{Z_i^t = q\},$$

$$\text{and} \quad N_{qq'} = \sum_{t=2}^{T} \sum_{i \in V^{t-1} \cap V^t} 1\{Z_i^{t-1} = q, Z_i^t = q'\}.$$

As such, a node that would not be present at each time point contributes to the likelihood only through the part of the trajectory where it is present. Moreover, given the latent groups $\mathbf{Z}$, for any $i, j \in V^t$, the conditional distribution of Y_{ij}^t is still given by (1) while whenever $i \notin V^t, j \in V$, we have Y_{ij}^t is deterministic and set to 0. Thus, a node absent at time t does not contribute to the likelihood of the observations. Generalization of our `VEM` algorithm easily follows.

Acknowledgments We would like to thank Tianbao Yang for making his code available from his web page. We sincerely acknowledge the SocioPatterns collaboration for providing the high school dataset (http://www.sociopatterns.org) and especially Alain Barrat for interesting discussions. This work was performed using the computing facilities of the CC LBBE/PRABI.

A. Optimization with respect to $\tau(i,q)$

In this section, we provide the exact fixed point equation satisfied by the values $\hat{\tau}(i,q)$ maximizing $J(\theta, \tau)$. We have

$$\hat{\tau}(i,q) \propto \alpha_q \prod_{j, j \neq i} \prod_{l=1}^{Q} [\phi_{ql}^1(Y_{ij}^1)]^{\hat{\tau}(j,l)} \prod_{t \geq 2} \prod_{q_2 \ldots q_t} \Big(\frac{\pi_{q_{t-1} q_t}}{\hat{\tau}(t,i,q_{t-1},q_t)} \Big)^{\hat{\tau}(2,i,q,q_2) \ldots \hat{\tau}(t,i,q_{t-1},q_t)}$$
$$\times \prod_{t \geq 2} \prod_{q_2 \ldots q_t, l} \phi_{q_t l}^t(Y_{ij}^t)^{\hat{\tau}_{\text{marg}}(t,j,l) \hat{\tau}(2,i,q,q_2) \ldots \hat{\tau}(t,i,q_{t-1},q_t)},$$

with the convention: whenever $t = 2$ then $q_{t-1} = q$. This equation is to be compared with our approximation given by (6).

References

Airoldi, E., D. Blei, S. Fienberg, and E. Xing (2008). Mixed-membership stochastic block-models. *Journal of Machine Learning Research 9*, 1981–2014.

Allman, E., C. Matias, and J. Rhodes (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist. 37*(6A), 3099–3132.

Allman, E., C. Matias, and J. Rhodes (2011). Parameters identifiability in a class of random graph mixture models. *Journal of Statistical Planning and Inference 141*, 1719–1736.

Ambroise, C. and C. Matias (2012). New consistent and asymptotically normal parameter estimates for random-graph mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 74*(1), 3–35.

Bickel, P., D. Choi, X. Chang, and H. Zhang (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist. 41*(4), 1922–1943.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Machine Intel. 22*(7), 719–725.

Celisse, A., J.-J. Daudin, and L. Pierre (2012). Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model. *Electron. J. Statist. 6*, 1847–1899.

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

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B 39*(1), 1–38.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports 486*, 75–174.

Fournet, J. and A. Barrat (2014, 09). Contact patterns among high school students. *PLoS ONE 9*(9), e107878.

Goldenberg, A., A. X. Zheng, S. E. Fienberg, and E. M. Airoldi (2010). A survey of statistical network models. *Found. Trends Mach. Learn. 2*(2), 129–233.

Gunawardana, A. and W. Byrne (2005). Convergence theorems for generalized alternating minimization procedures. *J. Mach. Learn. Res. 6*, 2049–2073.

Heaukulani, C. and Z. Ghahramani (2013). Dynamic probabilistic models for latent feature propagation in social networks. In S. Dasgupta and D. Mcallester (Eds.), *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, Volume 28, pp. 275–283. JMLR Workshop and Conference Proceedings.

Herlau, T., M. Mørup, and M. Schmidt (2013). Modeling temporal evolution and multiscale structure in networks. In S. Dasgupta and D. Mcallester (Eds.), *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, Volume 28, pp. 960–968. JMLR Workshop and Conference Proceedings.

Hoff, P., A. Raftery, and M. Handcock (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc. 97*(460), 1090–98.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*, 193–218.

Ishiguro, K., T. Iwata, N. Ueda, and J. B. Tenenbaum (2010). Dynamic infinite relational model for time-varying relational data analysis. In J. Lafferty, C. Williams, J. Shawe-taylor, R. Zemel, and A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23*, pp. 919–927.

Jernite, Y., P. Latouche, C. Bouveyron, P. Rivera, L. Jegou, and S. Lamassé (2014, 03). The random subgraph model for the analysis of an ecclesiastical network in merovingian gaul. *Ann. Appl. Stat. 8*(1), 377–405.

Jordan, M., Z. Ghahramani, T. Jaakkola, and L. Saul (1999). An introduction to variational methods for graphical models. *Machine Learning 37*(2), 183–233.

Karrer, B. and M. E. J. Newman (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E 83*, 016107.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer.

Latouche, P., E. Birmelé, and C. Ambroise (2012). Variational Bayesian inference and complexity control for stochastic block models. *Statistical Modelling 12*(1), 93–115.

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl. 40*(1), 127–143.

Liu, S., S. Wang, and R. Krishnan (2014). Persistent community detection in dynamic social networks. In V. Tseng, T. Ho, Z.-H. Zhou, A. Chen, and H.-Y. Kao (Eds.), *Advances in Knowledge Discovery and Data Mining*, Volume 8443 of *Lecture Notes in Computer Science*, pp. 78–89. Springer International Publishing.

Matias, C. and S. Robin (2014). Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proc. 47*, 55–74.

Sarkar, P. and A. W. Moore (2005, December). Dynamic social network analysis using latent space models. *SIGKDD Explor. Newsl. 7*(2), 31–40.

Snijders, T. A. (2011). Statistical models for social networks. *Annual Review of Sociology 37*, 129–151.

Teicher, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Statist 38*, 1300–1302.

Xing, E. P., W. Fu, and L. Song (2010). A state-space mixed membership blockmodel for dynamic network tomography. *Ann. Appl. Stat. 4*(2), 535–566.

Xu, A. and X. Zheng (2009). Dynamic social network analysis using latent space model and an integrated clustering algorithm. In *Dependable, Autonomic and Secure Computing, 2009. DASC '09. Eighth IEEE International Conference on*, pp. 620–625.

Xu, K. and A. Hero (2014, Aug). Dynamic stochastic blockmodels for time-evolving social networks. *Selected Topics in Signal Processing, IEEE Journal of 8*(4), 552–562.

Yang, T., Y. Chi, S. Zhu, Y. Gong, and R. Jin (2011). Detecting communities and their evolutions in dynamic social networks—a Bayesian approach. *Machine Learning 82*(2), 157–189.

Zanghi, H., C. Ambroise, and V. Miele (2008). Fast online graph clustering via Erdős Rényi mixture. *Pattern Recognition 41*(12), 3592–3599.

Zanghi, H., F. Picard, V. Miele, and C. Ambroise (2010). Strategies for online inference of model-based clustering in large and growing networks. *Ann. Appl. Stat. 4*(2), 687–714.

Zreik, R., P. Latouche, and C. Bouveyron (2015). The dynamic random subgraph model for the clustering of evolving networks. Technical report, HAL.