Estimation and clustering in a semiparametric Poisson process stochastic block model for longitudinal networks

Catherine Matias, Tabea Rebafka, Fanny Villers

Estimation and clustering in a semiparametric Poisson process stochastic block model for longitudinal networks

Catherine Matias, Tabea Rebafka and Fanny Villers

Sorbonne Universités, Université Pierre et Marie Curie, Université Paris Diderot, Centre National de la Recherche Scientifique, Laboratoire de Probabilités et Modèles Aléatoires, 4 place Jussieu, 75252 PARIS Cedex 05, FRANCE.
{catherine.matias,tabea.rebafka,fanny.villers}@upmc.fr

Abstract

In this work, we introduce a Poisson process stochastic block model for recurrent interaction events, where each individual belongs to a latent group and interactions between two individuals follow a conditional inhomogeneous Poisson process whose intensity is driven by the individuals' latent groups. The model is semiparametric as the intensities per group pair are modeled in a nonparametric way. First an identifiability result on the weights of the latent groups and the nonparametric intensities is established. Then we propose an estimation procedure, relying on a semiparametric version of a variational expectation-maximization algorithm. Two different versions of the method are proposed, using either histogram-type (with an adaptive choice of the partition size) or kernel intensity estimators. We also propose an integrated classification likelihood criterion to select the number of latent groups. Asymptotic consistency results are then explored, both for the estimators of the cumulative intensities per group pair and for the kernel procedures that estimate the intensities per group pair. Finally, we carry out synthetic experiments and analyse several real datasets to illustrate the strengths and weaknesses of our approach.

Keywords: Cox process, doubly stochastic Poisson process, dynamic interactions, expectation-maximization algorithm, integrated classification likelihood, longitudinal network, semiparametric model, stochastic block model, temporal networks, variational approximation.

1 Introduction

The past few years have seen a large increase in the interest for modeling dynamic interactions between individuals. While most of the recent approaches rely on discrete-time models, many real world interaction data contain continuous-time information on the interactions, e.g. email exchanges between employees in a company (Klimt and Yang, 2004) or between students and university staff (Kossinets and Watts, 2006), encounters between high school students (Fournet and Barrat, 2014) or animals (Jeanson, 2012) equipped with sensors. A frequently used discrete-time approach consists in aggregating the data on predefined time intervals to obtain a sequence of snapshots of interaction random graphs.

Besides the fact that this induces a loss of information, the choice of the time intervals has a direct impact on the results which is most often overlooked. Thus, developing models of interaction that exploit the continuous-time aspect of the data – either called *longitudinal networks, interaction event data or temporal networks* – is an important research issue.

Statistical methods for the analysis of longitudinal networks form a huge corpus, especially in social sciences and we do not pretend to provide an exhaustive bibliography on this topic. We refer to the very nice and recent review by Holme (2015) for a more complete view on temporal networks. A natural way of modeling temporal event data consists in the use of stochastic point processes defined on the positive real line. An important line of research has been the use of continuous-time Markov processes, with seminal works on dyad-independent models (Wasserman, 1980a,b) up to the development of so-called stochastic actor oriented models (e.g. Snijders and van Duijn, 1997; Snijders et al., 2010). In these works, observations consist in a series of time intervals of interactions and the models assume that each interaction lasts during the whole corresponding time interval, which is quite different from assuming that it occurred exactly at some time point. Here, we focus on this latter observation setup. Furthermore, we consider a model that allows for dependencies of the processes modeling the interactions of pairs of individuals.

The analysis of event data (also called time-to-event analysis, survival analysis, or lifetime analysis) is an old and important area in statistics (see for e.g. Andersen et al., 1993). Many authors have used multivariate point processes to model interaction events. More precisely, they consider a multivariate counting process $N(t) = (N_{i,j}(t))_{(i,j)}$ that counts the number of interactions of each pair (i, j) up to time t. In Butts (2008), counting processes have been introduced in the context of *action* data, which are a set of time-stamped directed interactions between individuals that moreover are marked by a label (representing a behavioral event). The author describes a very general setup where the realization of an event can be treated as independent given the events that occurred previously; each action may have its own hazard rate depending on the past of the process through the use of statistics that are based on the previous observations. The special case of piecewise constant hazard rates (possibly depending on covariates) is considered. The model may be viewed as a special instance of Cox's multiplicative hazard model with time-dependent covariates and constant baseline function. Maximum likelihood estimation based on numerical optimization is proposed. In the same vein, Vu et al. (2011) propose a general regression-based modeling of the intensity of (simple type) non recurrent interaction events. They consider two different frameworks: Cox's multiplicative and Aalen's additive hazard rates (see for instance Martinussen and Scheike, 2006). In the Cox setup, the main difference with the approach proposed by Butts is that the model includes a non constant baseline hazard and estimation relies on partial likelihood. Perry and Wolfe (2013) propose another variant of Cox's multiplicative intensity model for recurrent interaction events where the baseline function is specific to each individual. They establish consistency and asymptotic normality of a maximum partial likelihood estimator (under suitable regularity conditions). Note that in the above mentioned works, a set of statistics is chosen by the user as potential candidates that modulate the individuals' interactions. As in any regression framework, choosing these statistics might raise some issues, increasing their number potentially leads to a high-dimensional problem and interpretation of the results might be blurred by the correlation between these statistics.

In a slightly different setup, Malmgren et al. (2008, 2009) propose a cascading inhomogeneous Poisson process, where each individual is either in an *active* or in a *passive* phase. Phases are not observed and phase changes occur according to an inhomogeneous Poisson process. Here, data have the form $N(t) = (N_i(t))_i$ recording the activity of each individual with no distinction of whom he interacts to. The model can be formulated as a double-chain hidden Markov model, allowing the authors to use an efficient inference algorithm to estimate the model parameters from the observed data.

Clustering individuals based on interaction data represents a well-established technique for taking into account the intrinsic heterogeneity and summarizing information. In Navaroli et al. (2012), a model with latent groups of recipients, each of which is associated with a piecewise-constant Poisson rate over time is considered. Here again, observations consist in a number of interactions per individual and not per pair of individuals. In the context of dynamic random graphs (where a discrete-time sequence of graphs is observed), recent approaches propose to generalize the so-called stochastic block model (SBM) to a dynamic context (Yang et al., 2011; Xu and Hero, 2014; Matias and Miele, 2015). Stochastic block models posit that interaction between two individuals is conditionally independent of the interaction of any other pair, given latent discrete values on the interacting individuals (their latent group). Another attempt to use SBM in the context of interaction events appears in DuBois et al. (2013). The authors generalize the approach by Butts (2008) by adding discrete latent variables on the individuals. The conditional distribution of directed interactions, given the latent groups, depends on the past history of the process in a multiplicative form. Inference is carried out by a Markov chain Monte Carlo (MCMC) scheme in a Bayesian setup.

The approaches by Butts; Vu et al.; Perry and Wolfe and others rely on conditional Poisson processes (also known as doubly stochastic Poisson processes or Cox processes, the latter terminology should not be mistaken with Cox's multiplicative model of intensities). A particular instance of the conditional Poisson process is the Hawkes process, which consists in a collection of point processes with some background rate, and each event adds a nonnegative impulse to the intensity of all other processes. Cho et al. (2014) develop a model for spatial-temporal networks with missing information, relying on such self-exciting point processes for temporal dynamics, combined with a spatial Gaussian mixture for the spatial dynamics. They develop a variational expectation-maximization (`VEM`) algorithm to infer the unknown participants in an event given the location and the time of the event. Similarly, Linderman and Adams (2014) combine temporal Hawkes processes with latent distance models for implicit networks that cannot be observed directly.

In this work, a semiparametric stochastic block model for recurrent interaction events is considered. Each individual belongs to a latent group and conditional on these groups, two individuals interact according to an inhomogeneous Poisson process, whose intensity depends on the individuals' latent groups. We call our model a Poisson process stochastic block model (PPSBM). Contrarily to the works by Butts (2008); DuBois et al. (2013) and others, we do not rely on a parametric model that uses a set of predefined network statistics modulating the intensities. Instead, we propose nonparametric estimators of the intensities as well as a clustering of the individuals, using a semiparametric version of the `VEM` algorithm where the maximization step is replaced either by a histogram-type or by a

kernel estimation procedure. Semiparametric generalizations of the classical expectation-maximization (EM) algorithm (Dempster et al., 1977) have been proposed in many different contexts (see for e.g. Böhning (1995); Bordes et al. (2007); Robin et al. (2007) for semiparametric mixtures or Dannemann (2012) for a semiparametric hidden Markov model). However we are not aware of other attempts to incorporate a nonparametric estimate into a variational approximation of EM. Our nonparametric estimation of the intensities relies on the work of Reynaud-Bouret (2006) for the histogram-type approach and on Ramlau-Hansen (1983) for the kernel one.

The paper is organized as follows. In Section 2.1, the model and notations are presented, while the identifiability of the parameters is established (under natural conditions) in Section 2.2. Note that we provide results for the general case as well as for the affiliation model, where interaction intensities only depend on whether the two individuals are in the same group or not. Our estimation procedures are presented in Section 3. As already mentioned, they are based on a semiparametric variational expectation-maximization algorithm. Two different methods are given, relying either on a histogram-type or on a kernel estimator for the nonparametric part of the model. We also provide a model selection procedure for the partition used in the histogram-type approach. Furthermore, relying on the histogram-type approach, we propose an integrated classification likelihood (ICL) criterion that selects the number of groups adaptively (Section 3.6). Section 4 contains consistency results. Under a natural assumption on the convergence of the variational parameters (that estimate each individual's group), we first establish that Nelson-Aalen-type estimators that are based on our estimation procedure are consistent (as the total number of individuals increases) for estimating the cumulatives intensities per group pair. We also prove consistency of the corresponding kernel procedures that estimate the intensities per group pair. Synthetic experiments are carried out in Section 5, enlightening both the clustering capacities of our method as well as the performance of the nonparametric estimation of the different intensities. Finally in Section 6, the analysis of several real datasets illustrates the strengths and weaknesses of our approach.

2 A semiparametric Poisson process stochastic block model (PPSBM)

2.1 Model

We are interested in the pairwise interactions of n individuals during some time interval $[0, T]$. We suppose that every individual belongs to one out of Q groups, and the relation between two individuals, that is the way two individuals interact with another, is driven by their group membership. We choose to restrict our attention to undirected interactions with no self-interactions, but generalizations to the directed case, with or without self-interactions are straightforward. Denote by

$$\mathcal{R} = \{(i, j), 1 \leq i < j \leq n\},$$

the set of all possible pairs of individuals, which is also the set of all possible dyads in the graph. Its cardinality is $R = R_n = n(n-1)/2$. Observations consist in

$$\mathcal{O} = \{(t_m, (i_m, j_m)), m = 1, \dots, M\},$$

where $(t_m, (i_m, j_m)) \in [0, T] \times \mathcal{R}$ corresponds to the event that the individuals with indices i_m and j_m interact at time t_m and M is the total number of events in time interval $[0, T]$. We assume that $0 < t_1 < \cdots < t_M < T$, i.e. there is at most one event at a time. Now the distribution of these observations is modeled as follows. Let $Z_1, \dots, Z_n$ be i.i.d. (latent) random variables taking values in $\{1, \dots, Q\}$ with probabilities

$$\pi_q = \mathbb{P}(Z_1 = q), \quad q = 1, \dots, Q.$$

For the moment, we assume that Q is fixed and known and that $\pi_q \in (0, 1)$ for all $q \in \{1, \dots, Q\}$. When no confusion occurs, we also use the notation $Z_i = (Z^{i,1}, \dots, Z^{i,q})$ with $Z^{i,q} \in \{0, 1\}$ and then Z_i has multinomial distribution $\mathcal{M}(1, \pi)$ with $\pi = (\pi_1, \dots, \pi_Q)$.

Now, our Poisson process stochastic block model (PPSBM) is defined as follows. For every $(i, j) \in \mathcal{R}$, the interactions of individuals i and j, conditional on the latent groups Z_i, Z_j, are modeled by a conditional inhomogeneous Poisson process $N_{i,j}(\cdot)$ on $[0, T]$ with intensity depending only on the group values Z_i, Z_j. Thus, we suppose that the conditional intensities of processes $N_{i,j}(\cdot)$ are the same for any two dyads with same latent classes Z_i, Z_j. We consider nonnegative intensity functions $\alpha^{(q,l)}$ with $1 \leq q, l \leq Q$ such that the conditional intensity of process $N_{i,j}(\cdot)$ given that $Z^{i,q} Z^{j,l} = 1$ is $\alpha^{(q,l)}(\cdot)$ for any $(i, j) \in \mathcal{R}$. Note that $\alpha^{(q,l)} = \alpha^{(l,q)}$ and in the following we restrict our attention to the set of functions $\alpha = \{\alpha^{(q,l)}\}_{(q,l)\in\mathcal{Q}}$ where $\mathcal{Q} = \{(q, l); 1 \leq q \leq l \leq Q\}$. The corresponding cumulative intensities are denoted by

$$A^{(q,l)}(t) = \int_0^t \alpha^{(q,l)}(u)du, \quad \forall t \in [0, T].$$

The set of observations $\mathcal{O}$ forms a realization of the multivariate counting process $(N_{i,j}(\cdot))_{(i,j)\in\mathcal{R}}$ with conditional intensity process $(\alpha^{(Z_i,Z_j)}(\cdot))_{(i,j)\in\mathcal{R}}$. Note that $N_{i,j}$ is not a Poisson process, but a counting process with intensity $\sum_{q=1}^{Q}\sum_{l=1}^{Q} \pi_q \pi_l \alpha^{(q,l)}$. In the literature, conditional Poisson processes are also known as Cox processes or doubly stochastic Poisson processes. We denote $\theta = (\pi, \alpha) = (\pi, \{\alpha^{(q,l)}\}_{(q,l)\in\mathcal{Q}})$ the (infinite-dimensional) parameter of a PPSBM. The distribution of the multivariate counting process $(N_{i,j}(\cdot))_{(i,j)\in\mathcal{R}}$ under parameter value θ is denoted $\mathbb{P}_\theta$.

2.2 Identifiability

This section is concerned with the identifiability of parameter θ, up to label switching, from the distribution of the multivariate counting process $(N_{i,j}(\cdot))_{(i,j)\in\mathcal{R}}$ as defined below. Note that the functions $\alpha^{(q,l)}$ are intensities and thus are only identifiable almost everywhere (a.e.) on $[0, T]$. We denote $\mathfrak{S}_Q$ the set of permutations of $\{1, \dots, Q\}$.

Definition 1 (Identifiability up to label switching)**.** *The parameter $\theta = (\pi, \alpha)$ of a PPSBM is identifiable on $[0, T]$ up to label switching if*

$$\forall \theta, \tilde{\theta}, \quad \mathbb{P}_\theta = \mathbb{P}_{\tilde{\theta}} \implies \exists \sigma \in \mathfrak{S}_Q, \forall (q, l) \in \mathcal{Q}, \quad \begin{cases} \pi_q = \tilde{\pi}_{\sigma(q)}, \\ \alpha^{(q,l)} = \tilde{\alpha}^{(\sigma(q),\sigma(l))} \text{ a.e. on } [0, T]. \end{cases}$$

We now state an assumption that ensures identifiability up to label switching of the parameter in the general model. The particular case of the affiliation PPSBM is considered afterwards.

Assumption 1. *The set of intensities* $\{\alpha^{(q,l)}\}_{(q,l)\in\mathcal{Q}}$ *contains exactly* $Q(Q+1)/2$ *distinct functions.*

The intensities $\alpha^{(q,l)}$ may take identical values at some points or even on some subsets of $[0,T]$, but should not be equal almost everywhere.

Proposition 1. *Under Assumption 1, the parameter* $\theta = (\pi, \alpha)$ *is identifiable on* $[0,T]$*, up to label switching, from the PPSBM distribution of the multivariate counting process* $(N_{i,j}(\cdot))_{(i,j)\in\mathcal{R}}$ *on the same interval, as soon as* $n \geq 3$*.*

The proof is postponed to Appendix A. Note that the previous result does not cover the affiliation case, where only two intensities $\alpha^{\text{in}} \neq \alpha^{\text{out}}$ are considered and

$$\forall (q,l) \in \mathcal{Q}, \quad \alpha^{(q,l)} = \begin{cases} \alpha^{\text{in}} & \text{if } q = l, \\ \alpha^{\text{out}} & \text{if } q \neq l. \end{cases}$$

Proposition 2. *Assume that the intensities* α^{in} *and* α^{out} *are distinct functions on* $[0,T]$*. Then, both* α^{in} *and* α^{out} *are identifiable on* $[0,T]$ *from the affiliation PPSBM distribution of the multivariate counting process* $(N_{i,j}(\cdot))_{(i,j)\in\mathcal{R}}$ *on the same interval, as soon as* $n \geq 3$*. Moreover, as soon as* $n \geq \max\{Q,3\}$*, the proportions* $\{\pi_q; 1 \leq q \leq Q\}$ *are also identifiable, up to a permutation, from the same distribution.*

Again, the proof of this result is postponed to Appendix A.

2.3 Processes at stake

In this section, we introduce additional notations and especially some processes that will be used throughout the manuscript.

First, for any group $(q,l) \in \mathcal{Q}$, let us consider

$$Y^{(q,l)} = \begin{cases} \sum_{(i,j)\in\mathcal{R}} (Z^{i,q}Z^{j,l} + Z^{i,l}Z^{j,q}) & \text{if } q < l, \\ \sum_{(i,j)\in\mathcal{R}} Z^{i,q}Z^{j,q} & \text{if } q = l \end{cases} \tag{1}$$

the (unobserved) number of dyads $(i,j) \in \mathcal{R}$ with group membership (q,l). Note that the (unobserved) counting process

$$N_{\mathcal{Z}}^{(q,l)} = \begin{cases} \sum_{(i,j)\in\mathcal{R}} (Z^{i,q}Z^{j,l} + Z^{i,l}Z^{j,q}) N_{i,j} & \text{if } q < l, \\ \sum_{(i,j)\in\mathcal{R}} Z^{i,q}Z^{j,q} N_{i,j} & \text{if } q = l \end{cases} \tag{2}$$

has conditional intensity $Y^{(q,l)}\alpha^{(q,l)}$, which falls into the class of Aalen's multiplicative intensity models. This is a central property of this process, on which our work often relies. We also define $Z_m^{(q,l)} \in \{0,1\}$ as the (unobserved) binary indicator of observation (i_m, j_m) belonging to group (q,l)

$$Z_m^{(q,l)} = \begin{cases} Z^{i_m,q}Z^{j_m,l} + Z^{i_m,l}Z^{j_m,q} & \text{if } q < l, \\ Z^{i_m,q}Z^{j_m,q} & \text{if } q = l. \end{cases} \tag{3}$$

As those quantities are unobserved, our work is based on some proxies. We consider

$$\mathcal{T} = \left\{ \tau = (\tau^{i,q})_{1 \leq i \leq n, 1 \leq q \leq Q} : \forall i, q,\ \tau^{i,q} \in [0,1], \sum_{q=1}^{Q} \tau^{i,q} = 1 \right\}, \tag{4}$$

as a set of candidate proxies for the unobserved latent groups $Z^{i,q}$. Note that while the $Z^{i,q}$'s are indicators, their counterparts $\tau^{i,q}$ are weights, representing the probability that node i belongs to group q. For every $\tau \in \mathcal{T}$, we set

$$\bar{Y}^{(q,l)} = \begin{cases} \sum_{(i,j)\in\mathcal{R}} (\tau^{i,q}\tau^{j,l} + \tau^{i,l}\tau^{j,q}) & \text{if } q < l, \\ \sum_{(i,j)\in\mathcal{R}} \tau^{i,q}\tau^{j,q} & \text{if } q = l \end{cases} \tag{5}$$

and construct the weighted cumulative process $(N^{(q,l)})_{(q,l)\in\mathcal{Q}}$ defined by

$$N^{(q,l)} = \begin{cases} \sum_{(i,j)\in\mathcal{R}} (\tau^{i,q}\tau^{j,l} + \tau^{i,l}\tau^{j,q}) N_{i,j} & \text{if } q < l, \\ \sum_{(i,j)\in\mathcal{R}} \tau^{i,q}\tau^{j,q} N_{i,j} & \text{if } q = l. \end{cases} \tag{6}$$

If the $\tau^{i,q}$'s approximate the $Z^{i,q}$'s then the intensity of this process $N^{(q,l)}$ is approximately $\bar{Y}^{(q,l)}\alpha^{(q,l)}$. Finally we introduce a proxy for the indicator $Z_m^{(q,l)}$ that is defined as

$$\tau_m^{(q,l)} = \begin{cases} \tau^{i_m,q}\tau^{j_m,l} + \tau^{i_m,l}\tau^{j_m,q} & \text{if } q < l, \\ \tau^{i_m,q}\tau^{j_m,q} & \text{if } q = l. \end{cases} \tag{7}$$

3 Semiparametric estimation procedure

The complete-data likelihood (comprising observations $\mathcal{O}$ and latent variables $\mathcal{Z} = (Z_1, \ldots, Z_n)$) writes

$$\begin{aligned} \mathcal{L}(\mathcal{O}, \mathcal{Z}|\theta) &= \mathcal{L}(\mathcal{O}|\mathcal{Z}, \theta) \times \mathcal{L}(\mathcal{Z}|\theta) \\ &= \exp\left\{ - \sum_{(i,j)\in\mathcal{R}} A^{(Z_i,Z_j)}(T) \right\} \prod_{m=1}^{M} \alpha^{(Z_{i_m},Z_{j_m})}(t_m) \prod_{i=1}^{n}\prod_{q=1}^{Q} \pi_q^{Z^{i,q}}. \end{aligned} \tag{8}$$

The likelihood of the observed data $\mathcal{L}(\mathcal{O}|\theta)$ is obtained by summing the complete-data likelihood over the set of all possible configurations of the latent variables $\mathcal{Z}$, which is so huge that the likelihood of the observed data is intractable for direct maximization. Hence, we use an `EM` algorithm (Dempster et al., 1977), which is an iterative procedure that copes with latent variables. However, two different issues arise here. First, as already appears for SBM (Daudin et al., 2008), the `E`-step of the `EM` algorithm may not be performed in our context because it requires the computation of the conditional distribution of $\mathcal{Z}$ given the observations $\mathcal{O}$, which is not tractable. Therefore we rely on a variational approximation (Jordan et al., 1999) at the `E`-step of the algorithm. Second, part of our parameter is infinite dimensional so that the `M`-step is partly replaced by a nonparametric estimation procedure, giving rise to a semiparametric `EM` algorithm. As a consequence, we propose a variational approximation of a semiparametric `EM` algorithm, which is an iterative procedure that recursively applies the steps described below in detail. We refer for instance to Matias and Robin (2014) for a general description of the `VEM` algorithm and its links to `EM` in stochastic block models. Our complete algorithm is summarized in Section 3.5, Algorithm 1.

3.1 Variational E-step

We start by describing the variational approximation part of the algorithm. A classical E-step consists in computing the expectation of the complete log-likelihood, given the observations, at some current parameter value θ. Here, instead of computing the exact solution of a classical E-step, we search for an approximation within a restricted class of factorized distributions. Relying on the class of parameters $\mathcal{T}$ defined in (4), for every $\tau \in \mathcal{T}$, let the conditional distribution $\mathbb{P}_\tau(\cdot|\mathcal{O})$ of $\mathcal{Z}$ given $\mathcal{O}$ be defined by

$$\forall (q_1, \dots, q_n) \in \{1, \dots, Q\}^n, \quad \mathbb{P}_\tau\left(\mathcal{Z} = (q_1, \dots, q_n)|\mathcal{O}\right) = \prod_{i=1}^{n} \mathbb{P}_\tau(Z_i = q_i|\mathcal{O}) = \prod_{i=1}^{n} \tau^{i,q_i}. \quad (9)$$

We consider the quantity

$$Q_\tau(\theta) := \mathbb{E}_\tau\left[\log \mathcal{L}(\mathcal{O}, \mathcal{Z}|\theta)|\mathcal{O}\right],$$

as well as the entropy of distribution $\mathbb{P}_\tau(\cdot|\mathcal{O})$ given by

$$\mathcal{H}(\tau) := -\sum_{i=1}^{n}\sum_{q=1}^{Q} \tau^{i,q} \log \tau^{i,q},$$

with the convention that $0 * (-\infty) = 0$. The variational E-step consists in the maximization of

$$J(\theta, \tau) = Q_\tau(\theta) + \mathcal{H}(\tau),$$

with respect to (w.r.t) $\tau \in \mathcal{T}$.

Proposition 3. *The solution $\hat{\tau}$ to the maximization problem* $\mathrm{Argmax}_{\tau\in\mathcal{T}} J(\theta, \tau)$ *satisfies the following fixed-point equation*

$$\forall 1 \le i \le n, \forall 1 \le q \le Q, \quad \hat{\tau}^{i,q} \propto \pi_q \exp[D_{iq}(\hat{\tau}, \pi, \alpha)] \quad (10)$$

where $\propto$ means 'proportional to' and

$$D_{iq}(\tau, \pi, \alpha) = -\sum_{l=1}^{Q}\sum_{j\neq i} \tau^{j,l} A^{(q,l)}(T) + \sum_{l=1}^{Q}\sum_{m=1}^{M} (\tau^{j_m,l})^{\mathbb{1}\{i_m=i\}} (\tau^{i_m,l})^{\mathbb{1}\{j_m=i\}} \log\left(\alpha^{(q,l)}(t_m)\right),$$

with $\mathbb{1}\{A\}$ (or $\mathbb{1}_A$) the indicator function of set A.

From a practical point of view, the fixed point $\hat{\tau}$ is found by successively updating the variational parameters $\tau^{i,q}$ via Equation (10) until convergence.

Proof. According to (8), the complete-data log-likelihood writes

$$\log \mathcal{L}(\mathcal{O}, \mathcal{Z}|\theta) = -\sum_{(q,l)\in\mathcal{Q}} Y^{(q,l)} A^{(q,l)}(T) + \sum_{(q,l)\in\mathcal{Q}}\sum_{m=1}^{M} Z_m^{(q,l)} \log\left(\alpha^{(q,l)}(t_m)\right) + \sum_{i=1}^{n}\sum_{q=1}^{Q} Z^{i,q} \log \pi_q,$$

where $Y^{(q,l)}$ and $Z_m^{(q,l)}$ have been introduced in Equations (1) and (3), respectively. Now, note that $\mathbb{E}_\tau[Z^{i,q}|\mathcal{O}] = \mathbb{P}_\tau(Z^{i,q} = 1|\mathcal{O}) = \mathbb{P}_\tau(Z_i = q|\mathcal{O}) = \tau^{i,q}$. Moreover, by the factorization property (9), for every $i \neq j$ we have

$$\mathbb{E}_\tau[Z^{i,q}Z^{j,l}|\mathcal{O}] = \mathbb{E}_\tau[Z^{i,q}|\mathcal{O}]\mathbb{E}_\tau[Z^{j,l}|\mathcal{O}] = \tau^{i,q}\tau^{j,l}.$$

The quantity $\bar{Y}^{(q,l)}$ introduced in (5) is thus equal to $\mathbb{E}_\tau[Y^{(q,l)}|\mathcal{O}]$, namely the variational approximation of the mean number of dyads in group (q,l). Similarly, $\tau_m^{(q,l)}$ introduced in (7) equals $\mathbb{E}_\tau[Z_m^{(q,l)}|\mathcal{O}]$, the variational approximation of the probability that observation $(t_m,(i_m,j_m))$ corresponds to a dyad in group (q,l). It follows that

$$J(\theta,\tau) = -\sum_{(q,l)\in\mathcal{Q}} \bar{Y}^{(q,l)} A^{(q,l)}(T) + \sum_{(q,l)\in\mathcal{Q}} \sum_{m=1}^{M} \tau_m^{(q,l)} \log\left(\alpha^{(q,l)}(t_m)\right) + \sum_{i=1}^{n}\sum_{q=1}^{Q} \tau^{i,q} \log\frac{\pi_q}{\tau^{i,q}}. \tag{11}$$

The variational E-step consists in maximizing J with respect to the $\tau^{i,q}$'s which are constrained to satisfy $\sum_{q=1}^{Q}\tau^{i,q} = 1$ for all i. In other words, we maximize

$$M(\tau,\gamma) = J(\theta,\tau) + \sum_{i=1}^{n}\gamma_i\left(\sum_{q=1}^{Q}\tau^{i,q} - 1\right),$$

with Lagrange multipliers γ_i. The partial derivatives write

$$\begin{aligned}
\frac{\partial}{\partial\tau^{i,q}}M(\tau,\gamma) &= -\sum_{l=1}^{Q}\sum_{j\neq i}\tau^{j,l}A^{(q,l)}(T) + \sum_{l=1}^{Q}\sum_{m=1}^{M}(\tau^{j_m,l})^{\mathbb{1}\{i_m=i\}}(\tau^{i_m,l})^{\mathbb{1}\{j_m=i\}}\log\left(\alpha^{(q,l)}(t_m)\right)\\
&\quad + \log\pi_q + \gamma_i - \log\tau^{i,q} - 1,\\
\frac{\partial}{\partial\gamma_i}M(\tau,\gamma) &= \sum_{q=1}^{Q}\tau^{i,q} - 1.
\end{aligned}$$

The partial derivatives are null iff $\sum_{q=1}^{Q}\tau^{i,q} = 1$ and the $\tau^{i,q}$'s satisfy the fixed point equations (10), with $\exp(\gamma_i - 1)$ being the normalizing constant. □

3.2 Nonparametric M-step: general description

We now turn to the M-step of the algorithm. In a parametric context, the M-step consists in the maximization of

$$Q_\tau(\theta) = \mathbb{E}_\tau\left[\log\mathcal{L}(\mathcal{O},\mathcal{Z}|\theta)|\mathcal{O}\right]$$

with respect to $\theta = (\pi,\alpha)$, which is equivalent to minimizing $J(\theta,\tau)$ w.r.t. θ. Considering only the finite-dimensional part of the parameter, we easily obtain that the maximizer $\hat{\pi}$ of $Q_\tau(\pi,\alpha)$ w.r.t. π is given by

$$\hat{\pi}_q = \frac{\sum_{i=1}^{n}\tau^{i,q}}{\sum_{q=1}^{Q}\sum_{i=1}^{n}\tau^{i,q}} = \frac{1}{n}\sum_{i=1}^{n}\tau^{i,q}, \quad \forall q\in\{1,\dots,Q\}. \tag{12}$$

Now, as the parameter α is infinite-dimensional, we replace the maximization of $Q_\tau(\pi, \alpha)$ w.r.t. α by a nonparametric estimation step. Note that semiparametric versions of the `EM` algorithm have already been proposed in many different contexts (see for e.g. Böhning, 1995; Bordes et al., 2007; Robin et al., 2007; Dannemann, 2012). In the following, we consider two different approaches for updating α: a histogram-type and a kernel approach. In both cases, estimation would be straightforward using the process $(N_Z^{(q,l)})_{(q,l)\in\mathcal{Q}}$ defined by (2), which unfortunately is not observed. It is thus natural to consider its (current) variational approximation, namely the weighted cumulative process $(N^{(q,l)})_{(q,l)\in\mathcal{Q}}$ defined by (6).

3.3 Histogram-type `M`-step

In this part, each intensity $\alpha^{(q,l)}$ is estimated by a piecewise constant function and we propose a data-driven choice of the partition of the time interval $[0, T]$. In the following $(q, l) \in \mathcal{Q}$ is fixed and we start by considering a fixed partition $\mathcal{E} = (E_k)_{1\leq k\leq D_\mathcal{E}}$ of $[0, T]$ with partition size $D_\mathcal{E}$. Denote $S_\mathcal{E}$ the space of piecewise constant functions on $\mathcal{E}$. Note that the total number of dyads $R = n(n-1)/2$ is an upper bound for $\bar{Y}^{(q,l)}$ (the variational mean number of dyads in group (q, l)). Following Reynaud-Bouret (2006), we consider the projection estimator of $\alpha^{(q,l)}$ on $S_\mathcal{E}$ defined as

$$\hat{\alpha}_\mathcal{E}^{(q,l)} = \operatorname*{Argmin}_{f\in S_\mathcal{E}} \gamma_n^{(q,l)}(f),$$

where the least-squares contrast is defined (relatively to the counting process $N^{(q,l)}$ introduced in (6)) for all $f \in \mathbb{L}^2([0, T], dt)$ by

$$\gamma_n^{(q,l)}(f) = -\frac{2}{R}\int_0^T f(t)dN^{(q,l)}(t) + \frac{\bar{Y}^{(q,l)}}{R}\int_0^T f^2(t)dt.$$

The (variational) mean number of observations (i_m, j_m) with group membership (q, l) occurring in time interval E_k for $k \in \{1, \ldots, D_\mathcal{E}\}$ is given by

$$N^{(q,l)}(E_k) = \int_{E_k} dN^{(q,l)}(s) = \begin{cases} \sum_{m=1}^M \mathbb{1}_{E_k}(t_m)(\tau^{i_m,q}\tau^{j_m,l} + \tau^{i_m,l}\tau^{j_m,q}) & \text{if } q < l, \\ \sum_{m=1}^M \mathbb{1}_{E_k}(t_m)\tau^{i_m,q}\tau^{j_m,q} & \text{if } q = l. \end{cases} \quad (13)$$

Denote $|E_k|$ the length of interval E_k. Then the estimator $\hat{\alpha}_\mathcal{E}^{(q,l)}$ is given by

$$\hat{\alpha}_\mathcal{E}^{(q,l)}(\cdot) = \frac{1}{\bar{Y}^{(q,l)}}\sum_{k=1}^{D_\mathcal{E}} \frac{N^{(q,l)}(E_k)}{|E_k|}\mathbb{1}_{E_k}(\cdot). \quad (14)$$

We remark that with a fixed partition $\mathcal{E}$, the setup is purely parametric with a finite number of parameters $\alpha_k^{(q,l)} \geq 0$ that determine the piecewise constant function

$$\alpha_\mathcal{E}^{(q,l)}(\cdot) = \sum_{k=1}^{D_\mathcal{E}} \alpha_k^{(q,l)}\mathbb{1}_{E_k}(\cdot) \in S_\mathcal{E}.$$

This means that with this point of view, a classical `M`-step can be performed with some objective function $Q_\tau(\pi, \{\alpha_k^{(q,l)}\}_{(q,l)\in\mathcal{Q},1\leq k\leq D_\mathcal{E}})$ to be maximized w.r.t. π and $\{\alpha_k^{(q,l)}\}_{(q,l),k}$.

Interestingly, it turns out that the solution of this M-step is exactly the same as the projection estimators given by (14), that minimize the contrasts $\{\gamma_n^{(q,l)}\}_{(q,l)\in\mathcal{Q}}$ on $S_{\mathcal{E}}$. Note that as the estimators of $\alpha^{(q,l)}$ are computed separately, the approach allows to choose different partitions for different groups (q,l).

Now we turn to the choice of the partition and provide an adaptive model selection method, that is applied to every function $\alpha^{(q,l)}$ separately. Let $\mathcal{M}_n$ be a finite collection of partitions of $[0,T]$ considered for the estimation of $\alpha^{(q,l)}$ with fixed (q,l).

Adaptive estimation consists in choosing the best estimator among the collection of estimators $\{\hat{\alpha}_{\mathcal{E}}^{(q,l)}, \mathcal{E}\in\mathcal{M}_n\}$ with $\hat{\alpha}_{\mathcal{E}}^{(q,l)}$ defined by (14). The choice is based on a penalized least-squares criterion of the form

$$crit_n^{(q,l)}(\mathcal{E}) = \gamma_n^{(q,l)}(\hat{\alpha}_{\mathcal{E}}^{(q,l)}) + pen_n^{(q,l)}(\mathcal{E}),$$

for some penalty function $pen_n^{(q,l)} : \mathcal{M}_n \to \mathbb{R}^+$ that penalizes large partitions. Following Reynaud-Bouret (2006) we take for $\mathcal{M}_n$ either the collection of regular partitions $\mathcal{E}^r$ of $[0,T]$ with $D_{\mathcal{E}^r} = r$ intervals each of length T/r for $r\in\{1,\dots,r_{\max}\}$, or the collection of dyadic partitions $\mathcal{E}^d$ of $[0,T]$ with $D_{\mathcal{E}^d} = 2^d$ intervals of length $T/2^d$ for $d\in\{0,\dots,d_{\max}\}$ (where $r_{\max}$ and $d_{\max}$ are to be chosen). Furthermore, the penalty function is given by

$$pen_n^{(q,l)}(\mathcal{E}) = \frac{2D_{\mathcal{E}}}{R}C \qquad \text{with} \quad C = \frac{D_{\mathcal{E}^{\max}}}{T\bar{Y}^{(q,l)}} \sup_{1\le k\le D_{\mathcal{E}^{\max}}} N^{(q,l)}(E_k^{\mathcal{E}^{\max}}),$$

where $\mathcal{E}^{\max}$ denotes the finest partition in the collection $\mathcal{M}_n$, that is $D_{\mathcal{E}^{\max}} = r_{\max}$ in the regular case and $D_{\mathcal{E}^{\max}} = 2^{d_{\max}}$ in the dyadic case, and $E_k^{\mathcal{E}^{\max}}$ denotes the k-th interval of partition $\mathcal{E}^{\max}$.

Denote by $\hat{\mathcal{E}} = \hat{\mathcal{E}}^{(q,l)}$ the partition that minimizes $crit_n^{(q,l)}(\mathcal{E})$ over $\mathcal{M}_n$. Let $\hat{D}^{(q,l)} = D_{\hat{\mathcal{E}}^{(q,l)}}$ be the size of partition $\hat{\mathcal{E}}^{(q,l)}$. Then the adaptive estimator of intensity $\alpha^{(q,l)}$ is given by $\hat{\alpha}_{\hat{\mathcal{E}}}^{(q,l)}$ that writes

$$\forall t\in[0,T], \quad \hat{\alpha}_{\text{hist}}^{(q,l)}(t) = \hat{\alpha}_{\hat{\mathcal{E}}^{(q,l)}}^{(q,l)}(t) = \frac{1}{T\bar{Y}^{(q,l)}} \sum_{k=1}^{\hat{D}^{(q,l)}} \hat{D}^{(q,l)} N^{(q,l)}(E_k^{\hat{\mathcal{E}}}) \mathbb{1}_{E_k^{\hat{\mathcal{E}}}}(t). \tag{15}$$

Note that Reynaud-Bouret (2006) develops her approach in the Aalen multiplicative intensity model, which does not exactly correspond to our context. Moreover, our setup neither satisfies the assumptions of Theorem 1 in Reynaud-Bouret (2006), since the number of jumps of each process $N_{i,j}$ is not bounded by a known positive number, because here the $N_{i,j}$ are Poisson processes. Nevertheless, in our simulations, this procedure successfully estimates the intensities $\alpha^{(q,l)}$ (see Section 5). We refer to Baraud and Birgé (2009) for a theoretical study of an adaptive nonparametric estimation of the intensity of a Poisson process. Note also that Reynaud-Bouret (2006) studies other penalized least squares estimators (for e.g. relying on Fourier bases), which might be used here similarly. An alternative way for nonparametric intensity estimation relies on kernel estimators, that are explored in the following section.

3.4 Kernel estimation M-step

In this part kernel estimators of the intensities $\alpha^{(q,l)}$ are provided. A similar procedure has been proposed for a non variational version of the EM algorithm in Robin et al. (2007). Note

that in this approach, we do not maximize any criteria anymore. As already mentioned, if the $\tau^{i,q}$'s are good approximations of the latent $Z^{i,q}$'s, then the intensity of process $N^{(q,l)}$ defined by (6) is approximately $\bar{Y}^{(q,l)}\alpha^{(q,l)}$, where $\bar{Y}^{(q,l)}$ is defined in (5) as the variational mean number of dyads in group (q,l). Thus, a Nelson-Aalen-type estimator of the cumulative intensity $A^{(q,l)}(t)$ is given by

$$\frac{\bar{I}^{(q,l)}}{\bar{Y}^{(q,l)}} \int_0^t dN^{(q,l)}(u), \tag{16}$$

with indicator function $\bar{I}^{(q,l)} = \mathbb{1}\{\bar{Y}^{(q,l)} > 0\}$. Kernel methods are suited to estimate smooth functions. Following Ramlau-Hansen (1983) and considering a nonnegative kernel function K with support within $[-1,1]$ together with some bandwidth $b > 0$, the intensity $\alpha^{(q,l)}$ is estimated by

$$\begin{aligned}\hat{\alpha}^{(q,l)}_{\text{ker}}(t) &= \frac{\bar{I}^{(q,l)}}{b\bar{Y}^{(q,l)}} \int_0^T K\Big(\frac{t-u}{b}\Big) dN^{(q,l)}(u) \\ &= \frac{\bar{I}^{(q,l)}}{b\bar{Y}^{(q,l)}} \sum_{m=1}^{M} \tau_m^{(q,l)} K\Big(\frac{t-t_m}{b}\Big),\end{aligned} \tag{17}$$

where $\tau_m^{(q,l)}$ is defined in Equation (7). Note that the bandwidth b can be chosen adaptively from the data following the procedure proposed by Grégoire (1993). We choose not pursue this path here and in our simulations (Section 5) we rely on automatic choices provided by the software. Note also that kernel methods are not always suited to infer a function on a bounded interval as boundary effects may deteriorate their quality. It is out of the scope of this work to investigate refinements to correct for this issue.

3.5 Algorithm's full description

To recapitulate all steps, Algorithm 1 provides a full description of the procedure.

3.6 Model selection w.r.t. Q

In this section, a model selection criterion for choosing the best number of groups Q is proposed. We rely on the histogram-type `M`-step to construct this criterion. Indeed in this case, notwithstanding that the model is semiparametric and thus a nonparametric component is to be estimated, our estimator is finite dimensional. As such, the dimension of the selected adaptive estimator may be used in the penalization term. Anyway, we stress that once the number of groups is chosen, any of the two methods (histogram-type or kernel) may be used to estimate the intensities. We rely on an integrated classification likelihood (ICL) criterion introduced in the mixture context in (Biernacki et al., 2000) and adapted to the SBM in Daudin et al. (2008).

For any value $Q \geq 1$, let $\hat{\theta}_Q$ be the estimated parameter value with Q groups and $\hat{\mathcal{Z}}^Q$ the corresponding maximum a posteriori (MAP) classification at $\hat{\theta}_Q$ obtained by the `VEM` algorithm with the histogram-type `M`-step. For each value of Q, the parameter $\hat{\theta}(Q) = (\hat{\pi}(Q), \hat{\alpha}_{\text{hist}}(Q))$ has two components: the first one $\hat{\pi}(Q)$ is a vector of dimension $Q-1$, while the second has dimension $\sum_{(q,l)\in\mathcal{Q}} \hat{D}^{(q,l)}$, where $\hat{D}^{(q,l)}$ denotes the size of the partition

Algorithm 1: Semiparametric `VEM`–type algorithm

`//Initialization`
$s \leftarrow 0$
Initialize $\tau^{[0]}$
Initialize $J^{[0]} = -\infty$

while *convergence is not attained* **do**

`//M-step: Update group proportions`
Compute $\pi^{[s+1]}$ relying on Equation (12) with $\tau = \tau^{[s]}$

`//M-step: Update intensities`
Compute $\alpha^{[s+1]}$ relying either on Equation (15) (histogram method) or (17) (kernel method), with $\tau = \tau^{[s]}$

`//E-step: Update latent structure`
Compute $\tau^{[s+1]}$ relying on the fixed-point equation (10) using $(\pi, \alpha) = (\pi^{[s+1]}, \alpha^{[s+1]})$

`//Compute value of criteria` J `and test for convergence`
Compute $J^{[s+1]}$ through (11) with $\tau = \tau^{[s+1]}$ and $(\pi, \alpha) = (\pi^{[s+1]}, \alpha^{[s+1]})$
Test for convergence via $|J^{[s+1]} - J^{[s]}| \leq \epsilon$

$s \leftarrow s + 1$

used in the histogram estimator $\hat{\alpha}^{(q,l)}_{\text{hist}}(Q)$. In the adaptation of ICL to SBM these two components are penalized differently: the first one, that concerns the n individuals, is penalized by a $\log(n)/2$ term, while the second one concerning the dyads is penalized by a $\log(R_n)/2$ term. We refer to Daudin et al. (2008) for more details. In our case, the ICL criterion writes

$$ICL(Q) = \log \mathbb{P}_{\hat{\theta}_Q}(\mathcal{O}, \hat{\mathcal{Z}}^Q) - \frac{1}{2}(Q-1)\log(n) - \frac{1}{2}\log(R_n) \sum_{(q,l)\in\mathcal{Q}} \hat{D}^{(q,l)}. \tag{18}$$

Hence, after fixing an upper bound $Q_{\max}$ we select the number of groups

$$\hat{Q} = \underset{1\leq Q\leq Q_{\max}}{\text{Argmax}}\ ICL(Q).$$

The performance of the procedure is illustrated in Section 5.

4 Consistency results

In this section, we study the consistency (as n increases) of the Nelson-Aalen estimator of the integrated intensity $A^{(q,l)}(t) = \int_0^t \alpha^{(q,l)}(s)ds$ constructed with the values $\hat{\tau}^{i,q}$ obtained at the last step of the `VEM` algorithm. Provided that the variational parameters $\tau^{i,q}$ are close to the true membership indicators $Z^{i,q}$, we show that relying on the weighted processes $(N^{(q,l)})_{(q,l)\in\mathcal{Q}}$ defined in (6) is a good strategy. Moreover, we obtain that the kernel

estimators based on these variational parameters $\tau^{i,q}$ consistently estimate the intensities $\alpha^{q,l}$.

To ease the understanding of the study of the asymptotic behavior of the estimators, in this section we add subscripts n to all quantities previously defined that depend on the sample size R_n. Thus we let $\mathcal{Z}_n := (Z_1, \dots, Z_n)$ and $Y_n^{(q,l)} := Y^{(q,l)}$ for any $(q,l) \in \mathcal{Q}$. For $i \in \{1, \dots, n\}$ and $q \in \{1, \dots, Q\}$ denote by $\hat{\tau}_n^{i,q} \in [0,1]$ the variational parameters obtained at the last iteration of the `VEM` algorithm. Moreover, all previously defined quantities that are now considered at the specific values $\tau^{i,q} = \hat{\tau}_n^{i,q}$ are denoted with a hat. Thus, we consider the collection of weighted cumulative processes $(\hat{N}_n^{(q,l)})_{(q,l)\in\mathcal{Q}}$ defined by

$$\hat{N}_n^{(q,l)} = \begin{cases} \sum_{(i,j)\in\mathcal{R}_n} \left(\hat{\tau}_n^{i,q}\hat{\tau}_n^{j,l} + \hat{\tau}_n^{i,l}\hat{\tau}_n^{j,q}\right) N_{i,j} & \text{if } q < l, \\ \sum_{(i,j)\in\mathcal{R}_n} \hat{\tau}_n^{i,q}\hat{\tau}_n^{j,q} N_{i,j} & \text{if } q = l \end{cases}$$

and the value obtained at the last iteration of the `VEM` algorithm of the variational mean number of dyads (i,j) with group membership (q,l) given by

$$\hat{Y}_n^{(q,l)} = \begin{cases} \sum_{(i,j)\in\mathcal{R}_n} (\hat{\tau}_n^{i,q}\hat{\tau}_n^{j,l} + \hat{\tau}_n^{i,l}\hat{\tau}_n^{j,q}) & \text{if } q < l, \\ \sum_{(i,j)\in\mathcal{R}_n} \hat{\tau}_n^{i,q}\hat{\tau}_n^{j,q} & \text{if } q = l. \end{cases}$$

Furthermore, we define the Nelson-Aalen-type estimator for $A^{(q,l)}(t)$ by

$$\hat{A}_n^{(q,l)}(t) = \int_0^t \frac{\hat{I}_n^{(q,l)}}{\hat{Y}_n^{(q,l)}} d\hat{N}_n^{(q,l)}(s) = \frac{\hat{I}_n^{(q,l)}}{\hat{Y}_n^{(q,l)}} \hat{N}_n^{(q,l)}(t), \tag{19}$$

where $\hat{I}_n^{(q,l)} = \mathbb{1}\{\hat{Y}_n^{(q,l)} > 0\}$.

Assumption 2. *The variational estimators $\hat{\tau}_n = (\hat{\tau}_n^{i,q})_{iq}$ satisfy*

$$||\hat{\tau}_n - \mathcal{Z}_n||_\infty := \max_{i=1,\dots,n;q=1,\dots,Q} |\hat{\tau}_n^{i,q} - Z^{i,q}| = o_{\mathbb{P}_\theta}(1).$$

Note that establishing such a result on the behaviour of the variational parameters is beyond the scope of this paper. However, while the variational approximation comes with no theoretical guarantee of convergence in general (Gunawardana and Byrne, 2005), it is reasonable to think that it is consistent in our context. Indeed, convergence of the posterior distribution of group memberships given the observations to a factorized distribution for (classical) SBM has been established in Mariadassou and Matias (2015). We also refer to Celisse et al. (2012); Bickel et al. (2013) for related results on variational estimators in SBM.

4.1 Consistency of Nelson-Aalen-type estimators of the cumulated intensities

Theorem 1. *Under Assumption 2, the Nelson-Aalen-type estimator* (19) *is uniformly consistent, namely*

$$\forall (q,l) \in \mathcal{Q}, \quad \sup_{s\in[0,T]} |\hat{A}_n^{(q,l)}(s) - A^{(q,l)}(s)| \xrightarrow{\mathbb{P}_\theta} 0 \quad \text{as } n \to \infty.$$

We note that $N^{(q,l)}_{\mathcal{Z}_n}$ defined by (2) is an *oracle* version of our estimator $\hat{N}^{(q,l)}_n$ that would be obtained if the latent variables $\mathcal{Z}_n$ were known. Its conditional intensity has the multiplicative form $\alpha^{(q,l)}(\cdot)Y^{(q,l)}_n$. We introduce the corresponding Nelson-Aalen oracle for the cumulative intensity

$$A^{(q,l)}_{\mathcal{Z}_n}(t) = \frac{I^{(q,l)}_n}{Y^{(q,l)}_n} N^{(q,l)}_{\mathcal{Z}_n}(t), \tag{20}$$

where $I^{(q,l)}_n = \mathbb{1}\{Y^{(q,l)}_n > 0\}$. Some preliminary results are needed to prove the theorem.

Lemma 1. *As n tends to infinity, the quantities $Y^{(q,l)}_n$ defined by (1) satisfy*

$$Y^{(q,l)}_n = \begin{cases} n^2\pi_q\pi_l(1+o_{\mathbb{P}_\theta}(1)) & \text{if } q<l, \\ \frac{n^2}{2}\pi_q^2(1+o_{\mathbb{P}_\theta}(1)) & \text{if } q=l. \end{cases}$$

Proof of Lemma 1. Denote $Y^{(q)}_n = \sum_{i=1}^n Z^{i,q}$ the number of individuals in group q. The number $Y^{(q,l)}_n$ of dyads (i,j) with group membership (q,l) satisfies

$$Y^{(q,l)}_n = \begin{cases} Y^{(q)}_n Y^{(l)}_n & \text{if } q<l, \\ \frac{1}{2}Y^{(q)}_n(Y^{(q)}_n-1) & \text{if } q=l. \end{cases}$$

As $Z_i = (Z^{i,1},\dots,Z^{i,q})$ are i.i.d. with multinomial distribution $\mathcal{M}(1,\pi)$, the weak law of large numbers gives that $n^{-1}Y^{(q)}_n \xrightarrow{\mathbb{P}_\theta} \pi_q$ when $n\to\infty$, which implies Lemma 1. □

Lemma 2. *Under Assumption 2, as $n\to\infty$,*

$$\hat{Y}^{(q,l)}_n = \begin{cases} n^2\pi_q\pi_l(1+o_{\mathbb{P}_\theta}(1)) & \text{if } q<l, \\ \frac{n^2}{2}\pi_q^2(1+o_{\mathbb{P}_\theta}(1)) & \text{if } q=l. \end{cases}$$

Proof of Lemma 2. We prove the result for $1\le q<l\le Q$. One proceeds similarly for $q=l$. In order to control $\hat{Y}^{(q,l)}_n$ we introduce $Y^{(q,l)}_n$ so that

$$\hat{Y}^{(q,l)}_n = (\hat{Y}^{(q,l)}_n - Y^{(q,l)}_n) + Y^{(q,l)}_n.$$

Then Lemma 2 follows from Lemma 1, Assumption 2 and the following inequality

$$\begin{aligned} &|\hat{Y}^{(q,l)}_n - Y^{(q,l)}_n| \\ &\le \sum_{(i,j)\in\mathcal{R}_n} |\hat{\tau}^{i,q}_n\hat{\tau}^{j,l}_n + \hat{\tau}^{i,l}_n\hat{\tau}^{j,q}_n - Z^{i,q}Z^{j,l} - Z^{i,l}Z^{j,q}| \\ &\le \sum_{(i,j)\in\mathcal{R}_n} |(\hat{\tau}^{i,q}_n - Z^{i,q})\hat{\tau}^{j,l}_n + (\hat{\tau}^{i,l}_n - Z^{i,l})\hat{\tau}^{j,q}_n + (\hat{\tau}^{j,l}_n - Z^{j,l})Z^{i,q} + (\hat{\tau}^{j,q}_n - Z^{j,q})Z^{i,l}| \\ &\le 4\frac{n(n-1)}{2}||\hat{\tau}_n - \mathcal{Z}_n||_\infty. \end{aligned} \tag{21}$$

□

We are now ready to prove Theorem 1.

Proof of Theorem 1. Using the following decomposition

$$\sup_{s\in[0,T]} |\hat{A}_n^{(q,l)}(s) - A^{(q,l)}(s)| \leq \sup_{s\in[0,T]} |\hat{A}_n^{(q,l)}(s) - A_{\mathcal{Z}_n}^{(q,l)}(s)| + \sup_{s\in[0,T]} |A_{\mathcal{Z}_n}^{(q,l)}(s) - A^{(q,l)}(s)|, \quad (22)$$

the proof of the result splits into two parts.

First we work conditionally on the values of $\mathcal{Z}_n$. As already mentioned the multivariate oracle process $(N_{\mathcal{Z}_n}^{(q,l)})_{(q,l)\in\mathcal{Q}}$ with intensity $Y_n^{(q,l)}\alpha^{(q,l)}$ belongs to the multiplicative intensity model by Aalen. By Theorem IV.1.1 from Andersen et al. (1993) we obtain for any $(q,l) \in \mathcal{Q}$ that

$$\sup_{s\in[0,T]} |A_{\mathcal{Z}_n}^{(q,l)}(s) - A^{(q,l)}(s)| = o_{\mathbb{P}_\theta}(1), \quad \text{as } n \to \infty, \quad (23)$$

if the assumptions of this theorem are satisfied. Indeed, as by assumption $\pi_q > 0$ for all q, we have

$$\int_0^T \frac{I_n^{(q,l)}}{Y_n^{(q,l)}} \alpha^{(q,l)}(s)ds = \frac{I_n^{(q,l)}}{Y_n^{(q,l)}} A^{(q,l)}(T) \xrightarrow{\mathbb{P}} 0$$

$$\text{and} \quad \int_0^T (1 - I_n^{(q,l)})\alpha^{(q,l)}(s)ds = (1 - I_n^{(q,l)})A^{(q,l)}(T) \xrightarrow{\mathbb{P}} 0,$$

by Lemma 1 and the convergence of $I_n^{(q,l)} = \mathbb{1}\{Y_n^{(q,l)} > 0\} \xrightarrow{\mathbb{P}} 1$. Now, by the dominated convergence theorem, convergence (23) is also valid unconditional on $\mathcal{Z}_n$.

We now establish the second part of the proof, namely,

$$\sup_{s\in[0,T]} |\hat{A}_n^{(q,l)}(s) - A_{\mathcal{Z}_n}^{(q,l)}(s)| = o_{\mathbb{P}_\theta}(1). \quad (24)$$

We prove it for $1 \leq q < l \leq Q$ and one proceeds similarly for $q = l$. For $(i,j) \in \mathcal{R}_n$ we introduce the term

$$r_n^{i,j,q,l} := \frac{\hat{I}_n^{(q,l)}}{\hat{Y}_n^{(q,l)}} (\hat{\tau}_n^{i,q}\hat{\tau}_n^{j,l} + \hat{\tau}_n^{i,l}\hat{\tau}_n^{j,q}) - \frac{I_n^{(q,l)}}{Y_n^{(q,l)}} (Z^{i,q}Z^{j,l} + Z^{i,l}Z^{j,q}). \quad (25)$$

Then for any $s \in [0,T]$, we have

$$\hat{A}_n^{(q,l)}(s) - A_{\mathcal{Z}_n}^{(q,l)}(s) = \sum_{(i,j)\in\mathcal{R}_n} r_n^{i,j,q,l} N_{i,j}(s),$$

which leads to

$$\sup_{s\in[0,T]} |\hat{A}_n^{(q,l)}(s) - A_{\mathcal{Z}_n}^{(q,l)}(s)| \leq \max_{(i,j)\in\mathcal{R}_n} |r_n^{i,j,q,l}| \sup_{s\in[0,T]} \sum_{(i,j)\in\mathcal{R}_n} N_{i,j}(s). \quad (26)$$

The term $|r_n^{i,j,q,l}|$ may be bounded as follows

$$
\begin{aligned}
|r_n^{i,j,q,l}| &\leq \Big|(\hat{\tau}_n^{i,q} - Z^{i,q})\hat{\tau}_n^{j,l}\frac{\hat{I}_n^{(q,l)}}{\hat{Y}_n^{(q,l)}}\Big| + \Big|(\hat{\tau}_n^{j,q} - Z^{j,q})\hat{\tau}_n^{i,l}\frac{\hat{I}_n^{(q,l)}}{\hat{Y}_n^{(q,l)}}\Big| + \Big|Z^{i,q}(\hat{\tau}_n^{j,l} - Z^{j,l})\frac{\hat{I}_n^{(q,l)}}{\hat{Y}_n^{(q,l)}}\Big| \\
&+ \Big|Z^{j,q}(\hat{\tau}_n^{i,l} - Z^{i,l})\frac{\hat{I}_n^{(q,l)}}{\hat{Y}_n^{(q,l)}}\Big| + \Big|Z^{i,q}Z^{j,l}\Big(\frac{\hat{I}_n^{(q,l)}}{\hat{Y}_n^{(q,l)}} - \frac{I_n^{(q,l)}}{Y_n^{(q,l)}}\Big)\Big| + \Big|Z^{i,l}Z^{j,q}\Big(\frac{\hat{I}_n^{(q,l)}}{\hat{Y}_n^{(q,l)}} - \frac{I_n^{(q,l)}}{Y_n^{(q,l)}}\Big)\Big| \\
&\leq 4\frac{||\hat{\tau}_n - \mathcal{Z}_n||_\infty}{\hat{Y}_n^{(q,l)}} + 2\frac{|\hat{Y}_n^{(q,l)} - Y_n^{(q,l)}|}{\hat{Y}_n^{(q,l)}Y_n^{(q,l)}} \\
&\leq 4\frac{||\hat{\tau}_n - \mathcal{Z}_n||_\infty}{\hat{Y}_n^{(q,l)}} + 4n(n-1)\frac{||\hat{\tau}_n - \mathcal{Z}_n||_\infty}{\hat{Y}_n^{(q,l)}Y_n^{(q,l)}},
\end{aligned}
$$

where the last inequality comes from Inequality (21). Using Lemma 1 and Lemma 2, it follows that under Assumption 2, we have

$$
\max_{(i,j)\in\mathcal{R}_n} |r_n^{i,j,q,l}| \leq \frac{1}{n^2} o_{\mathbb{P}_\theta}(1). \tag{27}
$$

To deal with the term $\sup_{s\in[0,T]}\sum_{(i,j)\in\mathcal{R}_n} N_{i,j}(s)$ in (26), we rely on (2) and (20) and obtain

$$
\sup_{s\in[0,T]} \sum_{(i,j)\in\mathcal{R}_n} N_{i,j}(s) = \sum_{(i,j)\in\mathcal{R}_n} N_{i,j}(T) = \sum_{(q,l)\in\mathcal{Q}} N_{\mathcal{Z}_n}^{(q,l)}(T) = \sum_{(q,l)\in\mathcal{Q}} Y_n^{(q,l)} A_{\mathcal{Z}_n}^{(q,l)}(T),
$$

for n sufficiently large such that $I_n^{(q,l)} = 1$. Using (23) and Lemma 1, it follows that there exists some constant $K > 0$ (independent of n) such that

$$
\sup_{s\in[0,T]} \sum_{(i,j)\in\mathcal{R}_n} N_{i,j}(s) = Kn^2(1 + o_{\mathbb{P}_\theta}(1)). \tag{28}
$$

Combining (26) with (27) and (28) yields (24), which concludes the proof of Theorem 1. □

Let us briefly explain why second-order asymptotics cannot be obtained with this method. Indeed, it can be shown that the second term in the right-hand side of decomposition (22) converges at the rate $\sqrt{R_n} = O(n)$. However, it is not reasonable to assume that the variational parameters $\hat{\tau}_n$ satisfy $\|\hat{\tau}_n - \mathcal{Z}_n\|_\infty = o_{\mathbb{P}_\theta}(n^{-1})$ (which would induce that the first term in the right-hand side of (22) is $o_{\mathbb{P}_\theta}(n^{-1})$). Moreover, assuming that this term is $O_{\mathbb{P}_\theta}(n)$ would only yield that $\hat{A}_n^{(q,l)}$ converges at the parametric rate $\sqrt{R_n}$ and would not induce any characterization of this limiting distribution.

4.2 Consistency of kernel intensity estimators

In the same way, we can study the consistency of the kernel estimators built on the weighted cumulative process $(\hat{N}_n^{(q,l)})_{(q,l)\in\mathcal{Q}}$. We still consider the variational parameters $\hat{\tau}_n^{i,q} \in [0,1]$ obtained at the last iteration of the `VEM` algorithm and for each $(q,l) \in \mathcal{Q}$, the kernel estimator (defined in (17)) that relies on these values $\hat{\tau}_n^{i,q}$s, namely

$$
\hat{\alpha}_{\text{ker}}^{(q,l)}(t) = \frac{\hat{I}_n^{(q,l)}}{b_n \hat{Y}_n^{(q,l)}} \int_0^T K\Big(\frac{t-u}{b_n}\Big) d\hat{N}_n^{(q,l)}(u).
$$

Theorem 2. *Under Assumption 2, assuming that all intensities $\alpha^{(q,l)}$ are continuous at $t \in (0,T)$ and as soon as the bandwidth b_n satisfies $b_n \to 0$ and $b_n n^2 \to +\infty$, for any $(q,l) \in \mathcal{Q}$ the kernel estimator $\hat{\alpha}^{(q,l)}_{\text{ker}}(t)$ converges pointwisely in probability to $\alpha^{(q,l)}(t)$. Namely,*

$$\forall (q,l) \in \mathcal{Q}, \quad |\hat{\alpha}^{(q,l)}_{\text{ker}}(t) - \alpha^{(q,l)}(t)| \xrightarrow{\mathbb{P}_\theta} 0 \quad \text{as } n \to \infty.$$

Proof. Let $t \in (0,T)$ be such that all the intensities $\alpha^{(q,l)}$ are continuous at t and fix some particular $(q,l) \in \mathcal{Q}$. We introduce the *oracle* version of the kernel estimator, based on the (latent) process $N_{\mathcal{Z}_n}$, namely

$$\alpha^{(q,l)}_{\mathcal{Z}_n}(t) = \frac{I^{(q,l)}_n}{b_n Y^{(q,l)}_n} \int_0^T K\Big(\frac{t-u}{b_n}\Big) dN^{(q,l)}_{\mathcal{Z}_n}(u).$$

Relying on Theorem IV.2.1 in Andersen et al. (1993), conditional on $\mathcal{Z}_n$ and under the above assumptions, the pointwise convergence of $\alpha^{(q,l)}_{\mathcal{Z}_n}(t)$ to the true intensity $\alpha^{(q,l)}(t)$ in probability is obtained. Namely,

$$\forall \epsilon > 0, \quad \mathbb{P}_\theta(|\alpha^{(q,l)}_{\mathcal{Z}_n}(t) - \alpha^{(q,l)}(t)| > \epsilon | \mathcal{Z}_n) \to 0 \quad \text{as } n \to \infty.$$

By dominated convergence, this is also valid without conditioning on $\mathcal{Z}_n$. Now, it is sufficient to prove that $\hat{\alpha}^{(q,l)}_{\text{ker}}(t) - \alpha^{(q,l)}_{\mathcal{Z}_n}(t)$ converges in probability to zero. We easily obtain that

$$\hat{\alpha}^{(q,l)}_{\text{ker}}(t) - \alpha^{(q,l)}_{\mathcal{Z}_n}(t) = \frac{1}{b_n} \sum_{m=1}^{M} K\Big(\frac{t-t_m}{b_n}\Big) r_n^{i_m,j_m,q,l},$$

where the terms $r_n^{i,j,q,l}$ are defined by (25) for $q \neq l$ and similarly when $q = l$. Using a nonnegative kernel K, combined with Inequality (27), we get

$$\begin{aligned} |\hat{\alpha}^{(q,l)}_{\text{ker}}(t) - \alpha^{(q,l)}_{\mathcal{Z}_n}(t)| &\leq \max_{(i,j)\in\mathcal{R}_n} |r_n^{i,j,q,l}| \times \frac{1}{b_n} \sum_{m=1}^{M} K\Big(\frac{t-t_m}{b_n}\Big) \\ &\leq \frac{R_n}{n^2} o_{\mathbb{P}_\theta}(1) \times \frac{1}{b_n} \int_0^T K\Big(\frac{t-u}{b_n}\Big) \frac{dN_n(u)}{R_n}, \end{aligned}$$

where N_n is the cumulative process $N_n = \sum_{(i,j)\in\mathcal{R}_n} N_{i,j}$, whose intensity is given by

$$\sum_{q_1=1}^{Q} \cdots \sum_{q_n=1}^{Q} \pi_{q_1} \ldots \pi_{q_n} \sum_{(i,j)\in\mathcal{R}} \alpha^{(q_i,q_j)},$$

which is also continuous at t. Applying Theorem IV.2.1 in Andersen et al. (1993) to the kernel estimator

$$\frac{1}{b_n} \int_0^T K\Big(\frac{t-u}{b_n}\Big) \frac{dN_n(u)}{R_n},$$

we obtain its convergence in probability and then the fact that this term is bounded in probability. As $R_n = O(n^2)$, we thus obtain the result. □

5 Synthetic experiments

In this section, we investigate the numerical performance of our method first for clustering the n individuals and then for estimating the intensities of the $Q(Q+1)/2$ inhomogeneous Poisson processes. We also study the performance of the ICL criterion for recovering the true number of groups. Note that `R` programs are available on request to the second author.

To assess the clustering performance, we rely on the adjusted Rand index (ARI, Hubert and Arabie, 1985) that evaluates the agreement between the estimated and the true latent structure. For two classifications that are identical (up to label switching), this index equals 1, otherwise the ARI is smaller than 1 and negative values are possible. As for the estimation of the intensities, we use the $\mathbb{L}^2$-norm to measure the distance between the true and the estimated intensity. For each pair of groups $(q,l) \in \mathcal{Q}$, the quadratic risk is defined by

$$Risk(q,l) = ||\hat{\alpha}^{(q,l)} - \alpha^{(q,l)}||_2 = \Big(\int_0^T (\hat{\alpha}^{(q,l)}(t) - \alpha^{(q,l)}(t))^2 dt \Big)^{1/2}.$$

We consider various settings, grouped into two scenarios. The first scenario is intended to evaluate the classification performance and explores different levels of difficulty, by focusing on two shifted intensities with varying shifting parameter. The second scenario focuses on the recovery of the different intensities and it is also used to assess the performance of the ICL criterion, that selects the number of groups. There, intensities with different shapes and amplitudes are considered. More precisely, the two scenarios are set as follows.

1. We consider the affiliation model with $Q = 2$ latent groups, equal group proportions $\pi_q = 1/2$ for $q \in \{1,2\}$ and $n \in \{10, 30\}$ for the number of individuals. The intensities are sinusoids with different shifts (see Figure 1). More precisely, we take $T = 1$, $\alpha^{(q,q)}(t) = 10\sin(2\pi t) + 10$ and, for $q \neq l$, $\alpha^{(q,l)}(t) = 10\sin(2\pi(t+\phi)) + 10$ with shifting parameter $\phi \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$.

2. We choose $Q = 3$ latent groups with equal proportions $\pi_q = 1/3$ for $q \in \{1,2,3\}$ and $Q(Q+1)/2 = 6$ different intensity functions plotted in Figure 4. The number of individuals n varies in $\{20, 50\}$.

For every setting, 1000 simulations are carried out. For each simulation, a dataset from the corresponding PPSBM is generated, that is, latent groups are generated and then observations from $R_n = n(n-1)/2$ inhomogeneous Poisson processes on $[0,1]$ with intensity depending on the individuals' latent groups are simulated. Then, to estimate the latent structure and the intensities, the `VEM` algorithm is applied on the data. We rely on the true value of Q in the studies of clustering and estimation performances, and on several different values of Q in the case of model selection.

As the `VEM` algorithm requires initialization, we always run it several times with different initial values, and finally keep the result that optimizes the criterion J among the different runs. Note that the `VE`-step takes as arguments the current values of both τ and (π, α), while the `M`-step only relies on the current values of τ. It is hence natural to start the `VEM` algorithm with an `M`-step, to limit the number of parameters to initialize. We use two different methods to initialize τ, either by randomly chosen values or in a deterministic way by applying a `k-means` algorithm. For the latter, `k-means` is applied with Q groups on the rows of the $n \times n$ weighted adjacency matrix that contains the cumulated number

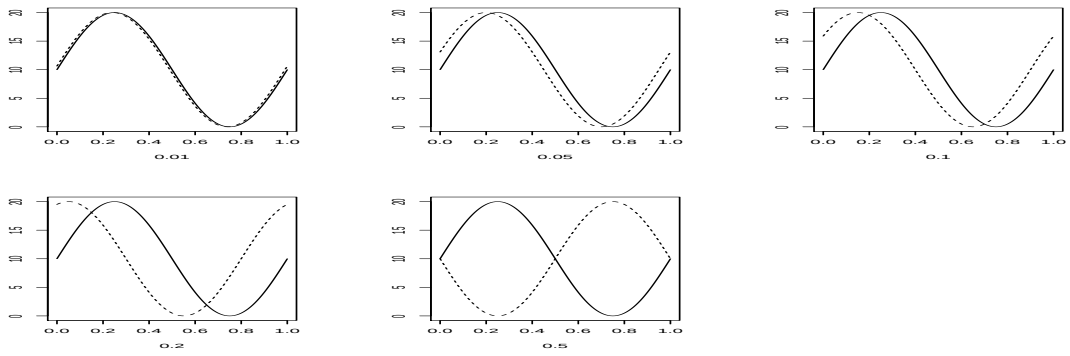

Figure 1: Intensities in scenario 1. Each picture represents the intra-group intensity $\alpha^{(q,q)}$ (bold line) and the inter-group intensity $\alpha^{(q,l)}$ for $q \neq l$ (dotted line) with different shifting parameter $\phi \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$.

of interactions between any two individuals (i.e. we forget about the times of interactions and consider an aggregated random graph).

Finally, the histogram-type estimator uses the penalty function described in Section 3.3 with a regular partition and $D_{\max} = 20$. As for the kernel estimator, we rely on the Epanechnikov kernel and the default bandwidth provided by the function `density` in the R software (R Development Core Team, 2008).

We report boxplots of the ARI obtained with the histogram and the kernel versions of our method in Figures 2 and 3 for the two scenarios, respectively.

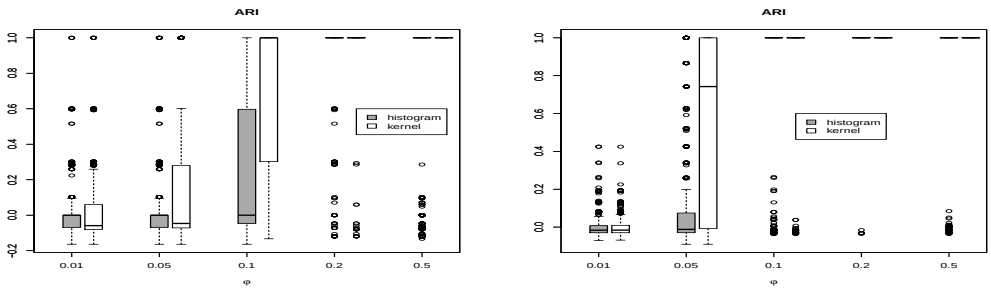

Figure 2: Boxplots of ARI in scenario 1 for the histogram (gray) and the kernel (white) estimators with $\phi \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$. Left panel $n = 10$, right panel $n = 30$.

For the first scenario, when considering small values of the shifting parameter ($\phi \in \{0.01, 0.05\}$), the intensities are so close that the classification is very difficult, especially when $n = 10$ is small. The classification improves when both the shift between the two intensities and the number of observations increase, with (almost) perfect classification when $\phi \geq 0.1$ ($n = 30$) or $\phi \geq 0.2$ ($n = 10$). We also observe that the kernel version of our method gives better classification results than the histogram method, which might be due to the choice of actually continuous intensities. For the second scenario, we can see that the classification is already very good with only $n = 20$ individuals and perfect with $n = 50$ scenarios. From the classification point of view, this scenario is somehow easy as

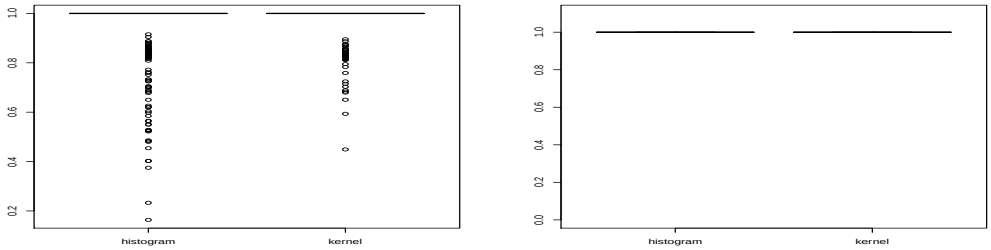

Figure 3: Boxplots of ARI in scenario 2 for the histogram (left) and the kernel (right) estimators. Left panel $n = 20$, right panel $n = 50$.

the intensities greatly differ from each other.

We now focus on the recovery of the different intensities and rely on the quadratic risk computed in scenario 2 where different shapes and amplitudes are explored. Table 1 gives for scenario 2 and for each pair of groups (q, l), the mean value of the risk $Risk(q, l)$ and its standard deviation estimated over the 1000 simulations, for the histogram and the kernel versions of our method. We also report in Table 1 the mean value (over those 1000 simulations) of the number of observations (i_m, j_m) with group membership (q, l), namely $\sum_{m=1}^{M} Z_m^{(q,l)}$. Moreover, Figure 4 and Figure 5 show for each pair of groups (q, l) the true and the estimated intensities for one simulation in scenario 2.

Table 1: Mean values for the number of points; mean values (with standard deviation) of the quadratic risks (averaged over 1000 simulations) for the histogram and the kernel estimators, for each group $(q, l) \in \mathcal{Q}$ and $n \in \{20, 50\}$ (scenario 2).

Index (q, l)	mean$(\sum_m Z_m^{(q,l)})$		Histogram $Risk(q, l)$		Kernel $Risk(q, l)$	
	$n = 20$	$n = 50$	$n = 20$	$n = 50$	$n = 20$	$n = 50$
$(1, 1)$	83.400	545.740	0.56 (0.62)	0.2 (0.19)	1.2 (0.64)	0.65 (0.12)
$(1, 2)$	146.810	949.480	1.23 (0.52)	1.03 (0.05)	2 (0.28)	1.61 (0.07)
$(1, 3)$	83.990	545.210	0.89 (0.53)	0.41 (0.07)	0.46 (0.36)	0.2 (0.06)
$(2, 2)$	32.830	210.480	1.15 (0.43)	0.73 (0.08)	0.75 (0.48)	0.36 (0.09)
$(2, 3)$	131.840	846.420	2.07 (0.52)	1.05 (0.14)	1.84 (0.28)	1.06 (0.11)
$(3, 3)$	47.100	295.310	1.5 (0.54)	0.76 (0.17)	0.87 (0.46)	0.44 (0.12)

When the true intensity is piecewise-constant, the histogram version of our method recovers the intensity and, as expected, outperforms the kernel estimator. Conversely, when the true intensity is smooth, the kernel estimator is better to recover the shape of the intensity. Note that for the intensity of groups $(2, 3)$ and when $n = 50$ is large, both histogram and kernel estimators perform well. Note also that a well-known drawback of the kernel estimator is that it suffers from boundary effects. This may be seen here for the intensities with groups $(1, 1)$, $(1, 2)$ and is less crucial for the other intensities that appear to be null at the interval boundaries.

Finally, we use scenario 2 to illustrate the performance of the ICL criterion. For each of the 1000 simulated datasets, the histogram-type method is run for every $q \in \{1, \ldots, 10\}$

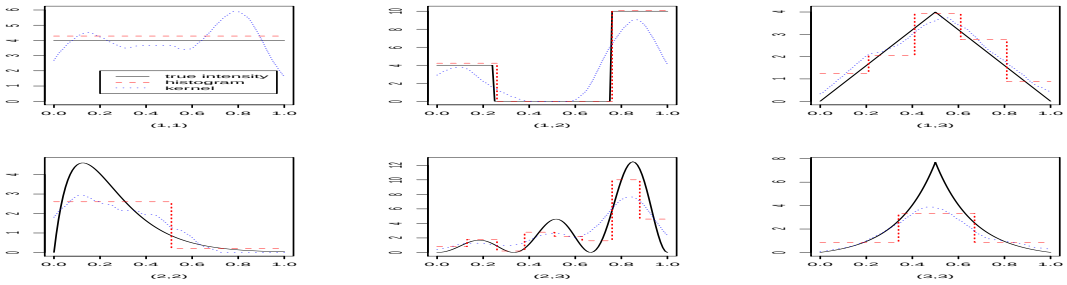

Figure 4: Scenario 2 with $n = 20$. True intensities (black continuous), histogram estimator (red dashed) and kernel estimator (blue dotted) for each pair of groups (q, l) with $1 \leq q \leq l \leq 3$.

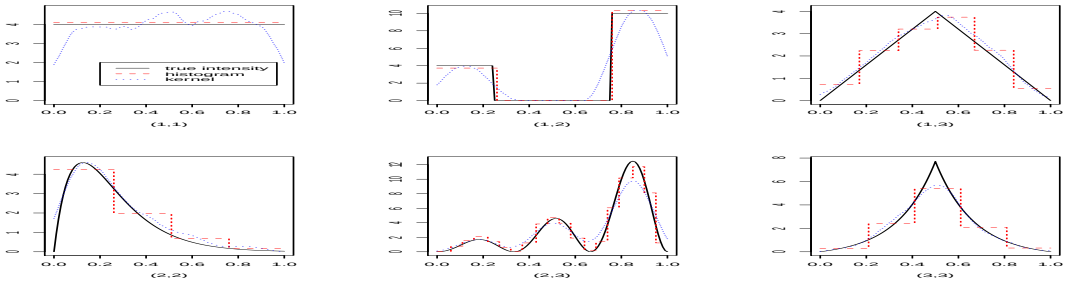

Figure 5: Scenario 2 with $n = 50$. True intensities (black continuous), histogram estimator (red dashed) and kernel estimator (blue dotted) for each pair of groups (q, l) with $1 \leq q \leq l \leq 3$.

to compute the values $\mathit{ICL}(q)$ given by (18) and finally evaluate the minimizer $\hat{Q}$ for this dataset. Results for $n = 20$ are reported in Figure 6. We observe that the correct number of groups is recovered in 95% of the cases (left panel). Moreover, the right panel shows that when ICL selects 2 or 4 groups, ARI of the classification with 3 groups is rather low. This shows that in those cases, classification with 3 groups is not the correct one, so that the VEM algorithm seems responsible for bad results more than the penalization term. For $n = 50$ our procedure selects the correct number of groupe for each simulated datasets.

6 Real datasets

6.1 London cycles dataset

The web site `https://api-portal.tfl.gov.uk/docs` provides the cycle hire usage data from the Santander stations (previously known as Barclays) of the city of London from 2012 to 2015. This dataset is also used in Guigourès et al. (To appear) with a different perspective. We choose to focus here on two consecutive days (randomly chosen), which are Wednesday, February 1, 2012 (dataset 1) and Thursday, February 2, 2012 (dataset 2). The data consist in pairs of stations associated to a single hiring/journey (departure station, ending station) and corresponding time stamp (hire time, with minute precision). Each

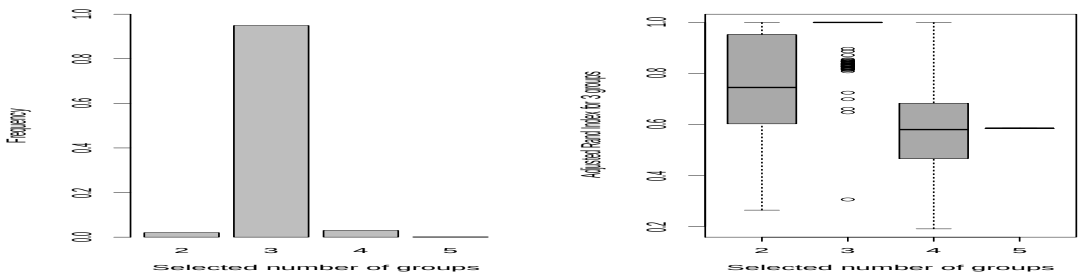

Figure 6: Scenario 3 with $n = 20$. Estimation of the number of groups via ICL criterion. Left panel shows the frequency of the selected number of groups. Right panel shows ARI of the classification with 3 groups depending on the selected number of groups.

dataset i contains n_i stations (the number of stations involved per day varies) and m_i hire events, with $n_1 = 415$, $n_2 = 417$, $m_1 = 17,631$ and $m_2 = 16,333$. Here we have a total time length $T = 86,400$ minutes and we first applied our procedure with histograms based on a dyadic partition with maximum size $D_{\max} = 2^5 = 32$.

An ICL criterion applied on these data shows a maximum value at $Q = 7$ clusters both for datasets 1 and 2 (data not shown). We plotted the geographic locations of the stations on a map of the city as well as the clusters obtained with our PPSBM (thanks to the OpenStreetMap project). Figure 7 shows the results for dataset 1. Similar results are obtained for dataset 2. We observe that our procedure mainly recovers geographic clusters, as stations are mainly linked through geographic proximity in the two datasets.

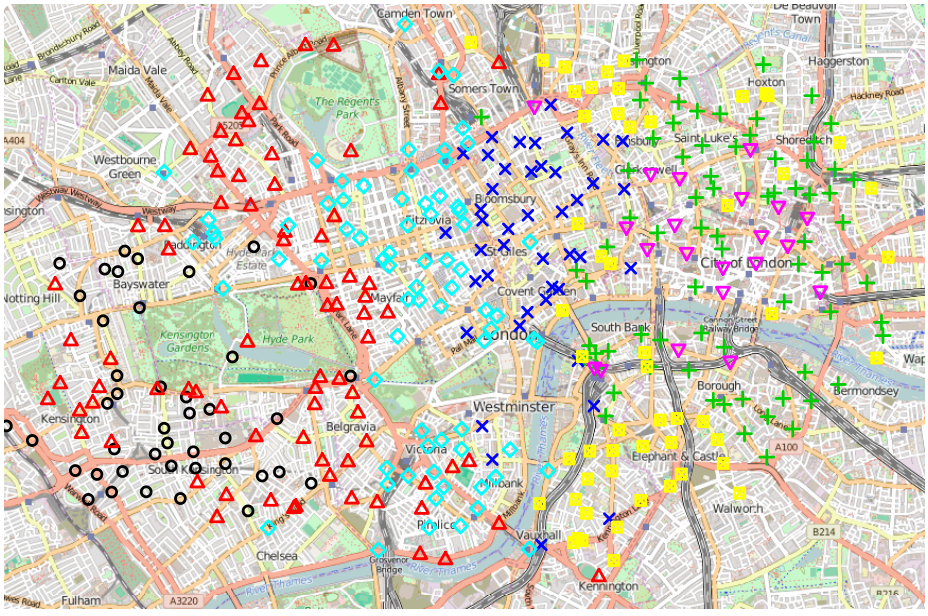

Figure 7: Geographic positions and PPSBM clusters of stations obtained for the London cycle hires on February 1, 2012 (dataset 1). There are 7 different clusters (represented by 7 different symbols).

By looking more closely at the results, it appears that the model selection steps acting on the partitions used in each histogram estimate always select $\hat{D}^{(q,l)} = 1$, resulting in constant intensity estimates. We interpret this result as the fact that in these datasets clusters are mainly driven by present or absent connections between the stations rather than the specific shape of intensities during the day. To confirm this assumption, we used a kernel estimator of the intensities per group pair relying on the clustering previously obtained.

When plotted each on their own scale, the intensities mainly exhibit 2 types of profiles: those with 2 modes around 8am and 7pm and those with an extra mode around noon. This is illustrated on Figure 8 that shows the estimated intensities for groups $(1, l)$ with $l \in \{1, \ldots, 5\}$. Intra-group intensities tend to exhibit the 3-modes shape: as the stations within a cluster tend to be geographically close, we observe activities during the lunch break. On the contrary, clusters that are geographically far from each other tend to exhibit a bimodal inter-group intensity (data not shown).

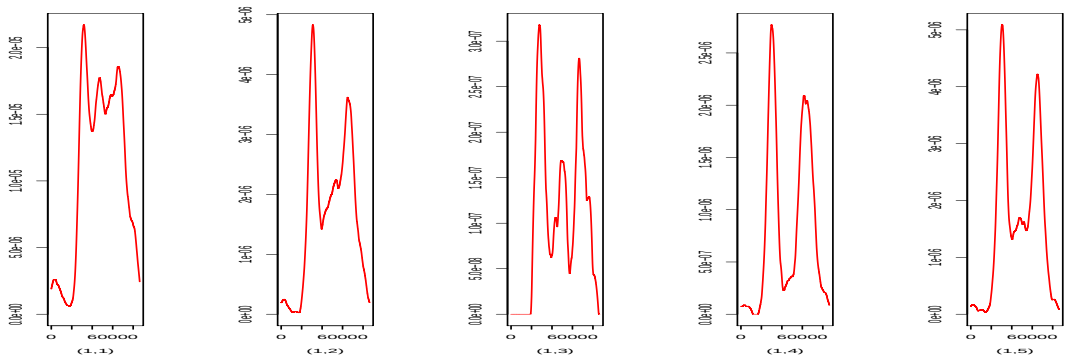

Figure 8: Kernel intensities estimates for the 7 clusters on dataset 1. The plots show estimated intensities for groups $(1, l)$ with $l \in \{1, \ldots, 5\}$.

Now, looking at the y-scales of the different intensities, we note that they are quite different. When plotting the intensities on the same y-scale it appears that out of the 28 plots, 15 of them are almost null intensities. One group pair (namely $(6, 6)$) shows very high intensity values, that are 3 times larger than the values of the intensity that exhibits the second largest values (group pair $(3, 6)$). These are plotted on Figure 9, together with two other group pairs: one exhibiting an almost null intensity ($(1, 3)$ already plotted on Figure 8 in its own scale) and one with small but non null intensity (namely group pair $(5, 6)$). This confirms that here the clusters are mainly driven by present or absent connections between the stations rather than specific intensity shapes. We mention that group 6 is plotted as a pink down-pointing triangle ($\triangledown$) in Figure 7 and corresponds to a cluster with small geographic expansion.

6.2 Primary school temporal network dataset

We analysed the dataset presented in Stehlé et al. (2011) and available at the website `www.sociopatterns.org/datasets/primary-school-temporal-network-data/`. To understand contacts between children at school and quantify the transmission opportunities of

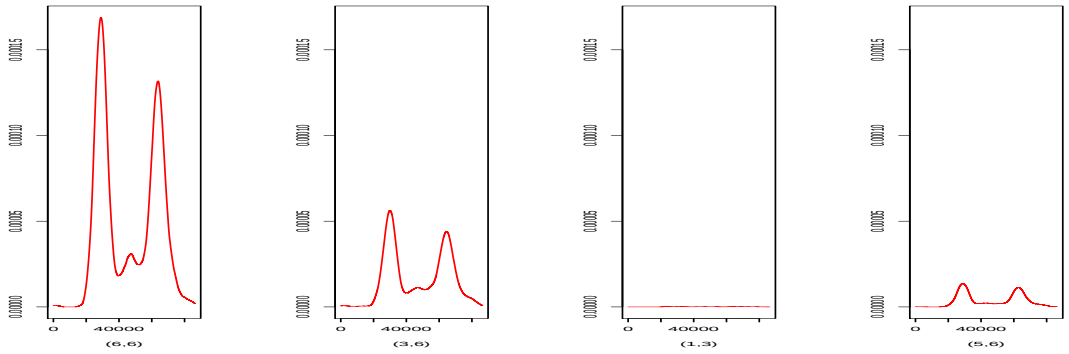

Figure 9: Kernel intensity estimates for the 7 clusters on dataset 1, plotted on the same y-scale. The plots show the group pairs $(6,6)$ (largest intensity values), $(3,6)$ (second largest), $(1,3)$ (among the 15 plots with intensity almost null) and $(5,6)$ (small intensity values, but non null).

respiratory infections, data on face-to-face interactions were collected in a French primary school (children from 6 to 12 years old) during two days in October 2009. The school is composed by 5 grades, each of them comprising two classes, for a total of 10 classes (denoted by $1A, 1B, \ldots, 5A, 5B$). Each class has an assigned teacher and an assigned room. The school day runs from approximately 8.30am to 4.30pm, with a lunch break from 12pm to 2pm and two breaks of 20-25 min around 10.30am and 3.30pm. Lunches are served in a common canteen, and a shared playground is located outside the main building. As the playground and the canteen do not have enough capacity to host all pupils at the same time, only two or three classes have breaks at the same time, and lunches are taken in two consecutive turns. The dataset contains $125,773$ face to face contacts between $n = 242$ individuals (232 children and 10 teachers) observed during a total time length $T = 116,900$ seconds (from 8.45am on the first day to around 5.05pm on the second day). We refer to (Stehlé et al., 2011) for more details on the dataset.

We applied our procedure with histograms based on a dyadic partition with maximum size $D_{max} = 2^{10} = 1024$, and with the number of groups Q varying in $\{1, \ldots, 30\}$. For Q varying between 2 and 11, Figure 10 shows the repartition of the n individuals (232 children coming from 10 classes and the 10 teachers appear with different colors) into those Q groups. When Q is small ($Q \leq 6$), our procedure gathers different classes and their corresponding teachers. When Q is larger, our procedure makes a sharper clustering according to the behavior of the children. For example for $Q = 11$, the procedure separates children from the same class: either to isolate a few of them in a group (3 children of class $1B$ are put together in one group), or to put together children of different classes (one group is made of children of classes $1A, 3A$ and $4B$). But overall, the classes of children are mainly conserved inside a same group. Note that teachers never form a particular group apart, but are gathered with their assigned class, suggesting (Stehlé et al. (2011) made the same observation) that contacts between teachers are sparse and that in this dataset clustering is mainly driven by communities (*i.e.* groups of highly connected individuals, with few inter-groups interactions).

Our model selection criterion for choosing the best number of groups Q does not provide

a reasonably small number of clusters that could be used for interpretation on this dataset. Indeed, the ICL criterion increases on the whole range of values $Q \in \{1, \ldots, 30\}$. It has been observed by other authors that this may happen for large datasets (see Guigourès et al., To appear, and the references therein). Thus, building on the previous remarks concerning the repartition of the children classes across the groups, we choose to analyse below the data with both $Q = 6$ and $Q = 11$ groups.

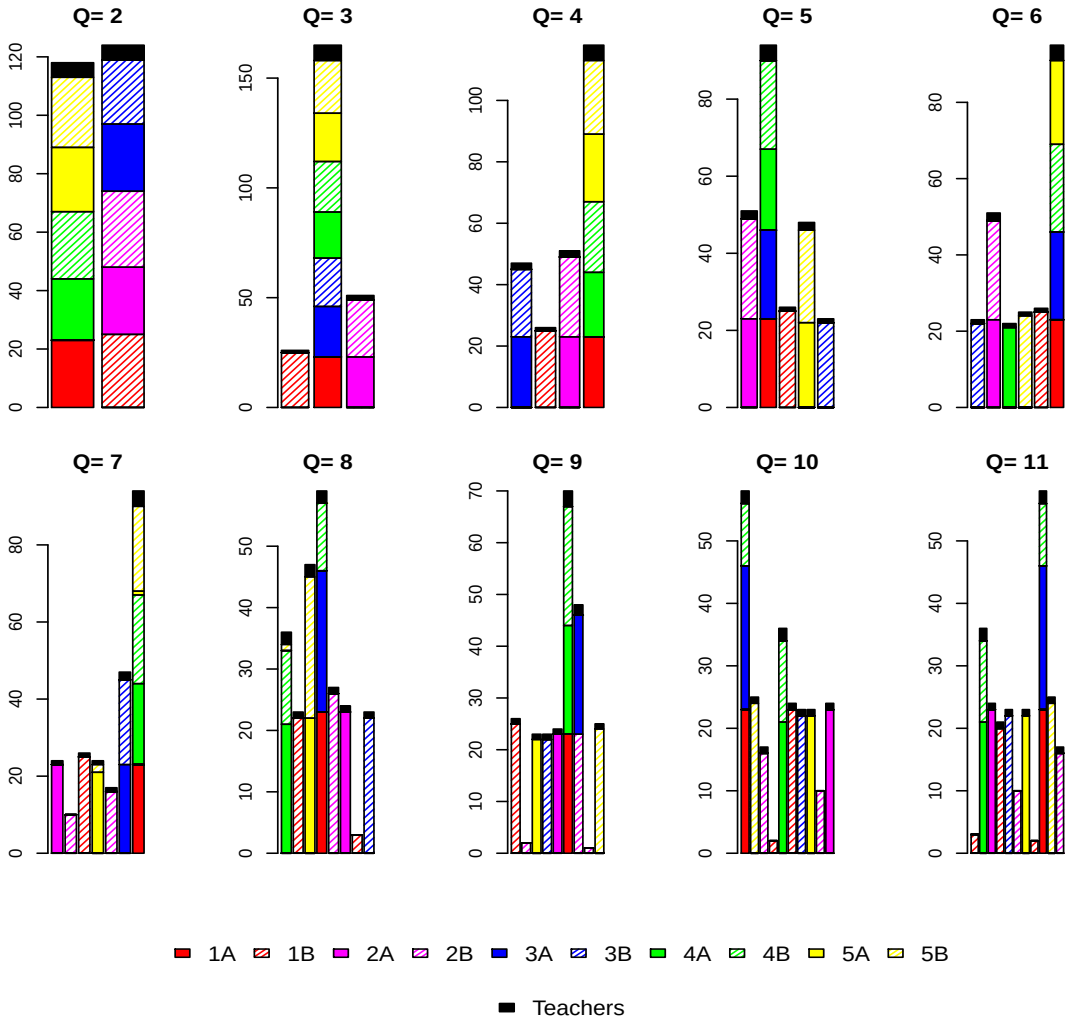

Figure 10: Partition of the $n = 232$ individuals into Q groups for $Q \in \{2, \ldots, 11\}$. Each of the five colors (red, purple, blue, green, yellow) represents a grade while black is for teachers. Plain and hatching distinguish the two classes in the same grade. For each picture, there are Q vertical bars corresponding to the Q different groups.

We thus consider the intensities estimated by our algorithm. Note that the contacts are measured during two days, with no encounters between the end of afternoon on the first day and the early morning on the second day. However, we did not input this information in our algorithm and the procedure automatically selects the time partition adapted to the data.

For $Q = 6$, the estimated intensities for groups (q, l) with $q \neq l$ can be considered as null. As groups mainly correspond to classes, this highlights that most contacts involve children of the same class and that the dataset is structured into communities. Figures 11 and 12 show the estimated intra-group intensities with $Q = 6$ and $Q = 11$ groups, respectively. In the case $Q = 11$, note that groups $3, 5, 7$ and 10 correspond exactly to classes $2A, 3B, 5A$ and $5B$, respectively (with their corresponding teacher). In those intra-group intensities, we recover peaks of interactions during the two breaks of 20-25 minutes around 10.30am and 3.30pm. During the lunch breaks, it appears that interactions between children may vary from the first to the second day, and are less important than during the breaks where they play together.

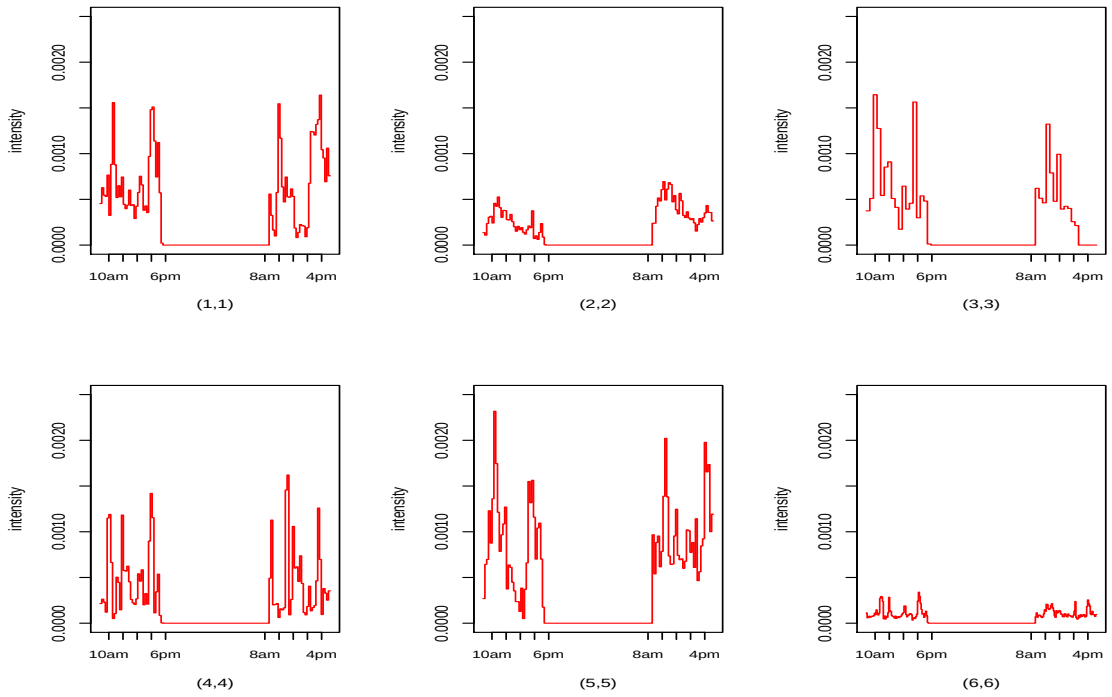

Figure 11: Estimated intra-group intensities for $Q = 6$ groups (plotted on the same y-scale).

We now consider the inter-group connections obtained with $Q = 11$ groups. Here again, most of the estimated intensities for groups (q, l) with $q \neq l$ can be considered as null, except for some that we discuss now. First, we recall that our procedure splits some children of the same class into separate groups. Indeed, groups $1, 4, 8$ turn out to form one class, while groups $6, 11$ form another class (each one with its teacher). That is, the inter-group interactions $(1, 4), (1, 8), (4, 8)$ and $(6, 11)$ (from our clustering point of view) correspond in fact to intra-class interactions. The corresponding estimated intensities are shown in Figure 13. Groups 6 and 11 are respectively constituted of 10 and 16 children of class $2B$ (26 children in this class). During lunch time, children of group 6 do not interact with other children (especially on the second day), whereas children of group 11 interact a lot during lunch. It seems that our procedure has recognized two subgroups in class $2B$: children

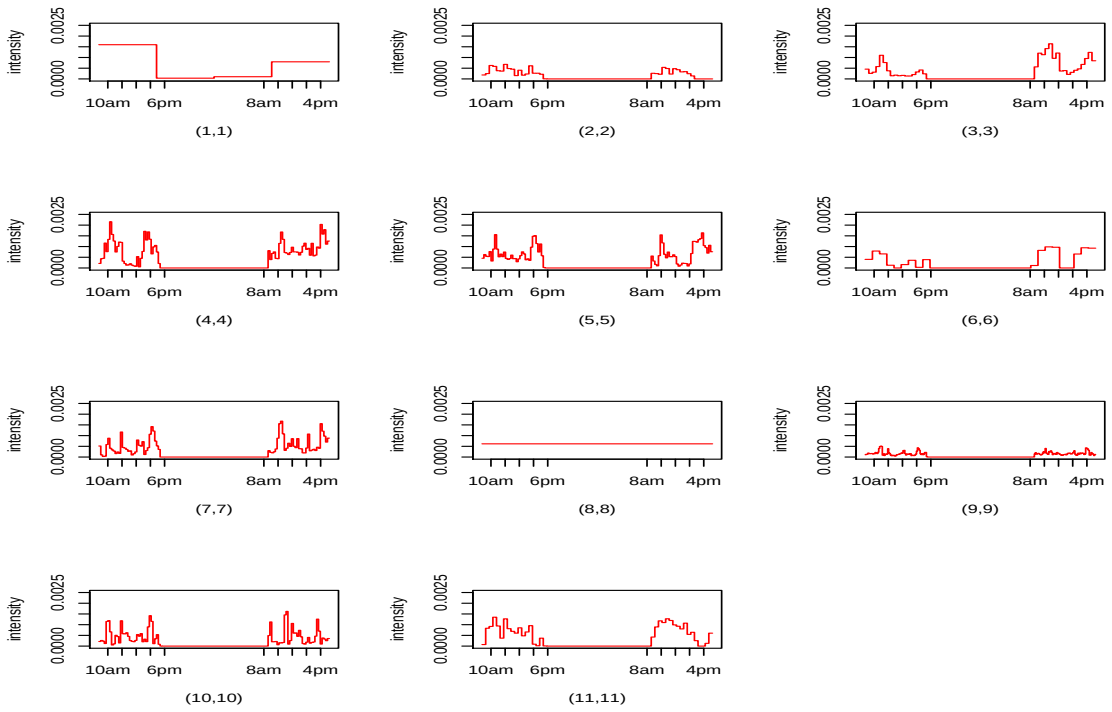

Figure 12: Estimated intra-group intensities for $Q = 11$ groups (plotted on the same y-scale).

having lunch at school and those going home for lunch. Groups $1, 4$ and 8 are respectively constituted of $3, 20$ and 2 children of class $1B$ (25 children in this class). The estimated intensities suggest a particular behaviour of some children of class $1B$: group 8 consists of two children without any contact between them, but with strong interaction with the children of group 1. Indeed, the intensity between groups $(1, 8)$ is four times higher than the other intensities.

Second, from Figure 14 we observe that intensities between groups made of children of the same grade are significant, suggesting that children mostly interact with children of the same age. Figure 14 shows the estimated interactions between classes $5A$ (group 7) and $5B$ (group 10) and between the classes $2A$ (group 3) and $2B$ (group 11). Those interactions especially occur during the lunch break. As lunch is served in a common canteen with two or three classes at a time, it is likely that classes are grouped by grade to get their lunch. Note however that the estimated intensities between classes of the same grade are lower than the intra-classes intensities.

As a conclusion, let us note we recover many results of Stehlé et al. (2011). In particular, we observe that contacts occur mostly within each class, and that more contacts are observed between children of the same grade than with other grades. We also recover periods of breaks during which interactions are more important. We also see that when Q is relatively large, PPSBM allows for detecting subgroups in the classes with specific behavior of some individuals (leaving school for lunch) and interesting behavior between

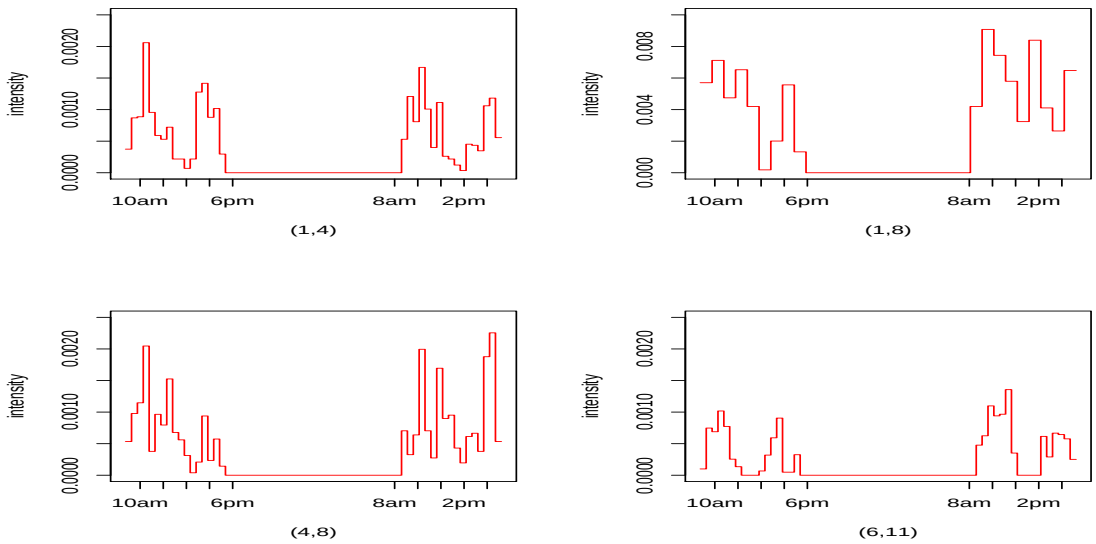

Figure 13: Inter-group interactions that correspond to intra-class interactions with $Q = 11$ groups.

classes (interaction with classes of the same grade). We think that this is mainly due to the fact that PPSBM takes into account the information provided by the timestamps of the events. Without this temporal information, it is hard to imagine to obtain similar results. In particular, aggregating data on a day scale (to construct for discrete time networks) would not provide such a refined analysis.

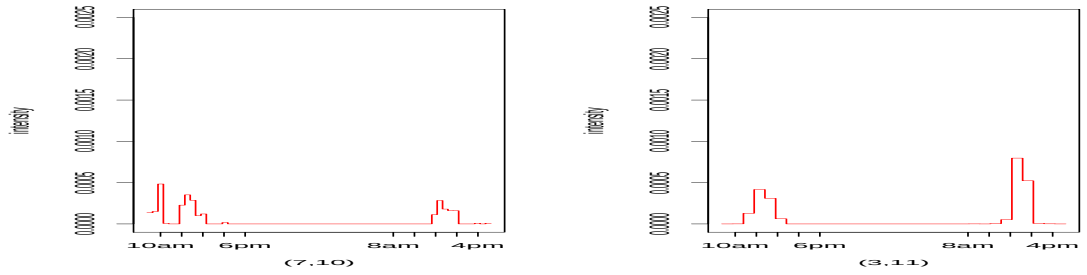

Figure 14: Estimated inter-group intensities between classes $5A$ (group 7) and $5B$ (group 10) and between classes $2A$ (group 3) and $2B$ (group 11) for $Q = 11$ groups.

6.3 Enron dataset

The Enron dataset is composed of email exchanges between $n = 147$ persons working at Enron, mostly in the senior management, covering the period of the affair that led to the bankruptcy of the company in 2001 (Klimt and Yang, 2004). Our dataset (retrieved from `http://www.cs.cmu.edu/~enron/`) contains $23,456$ emails exchanged among these

147 persons between November 14, 1998 and July 17, 2002, for which the sender, the recipient and the time when the email was sent are known. Note that we analyse these data with our undirected model.

Here again, our ICL criterion did not provide a reasonably small enough number of clusters that could be used for interpretation. We thus choose to apply our procedure on the data with $Q = 3$ groups and dyadic partitions for the histogram-type estimators of the intensities using $D_{\max} = 2^{10} = 1024$. Figure 15 shows the estimated intensities. The second group is a rather silent group with very little activity. The first group is more active with substantial intra-group communication. The third group is characterized by a very high level of intra-group communication and some interaction with the first group. Figure 16 illustrates how communication evolves in the three groups over time. In this picture, we choose to partition the whole time interval into 4 regular subintervals (with approximate length of 10 months each) and represent the mean intensities (intra- and inter-groups) over these subintervals. We see that intra- and inter-group communications are not stationary through time.

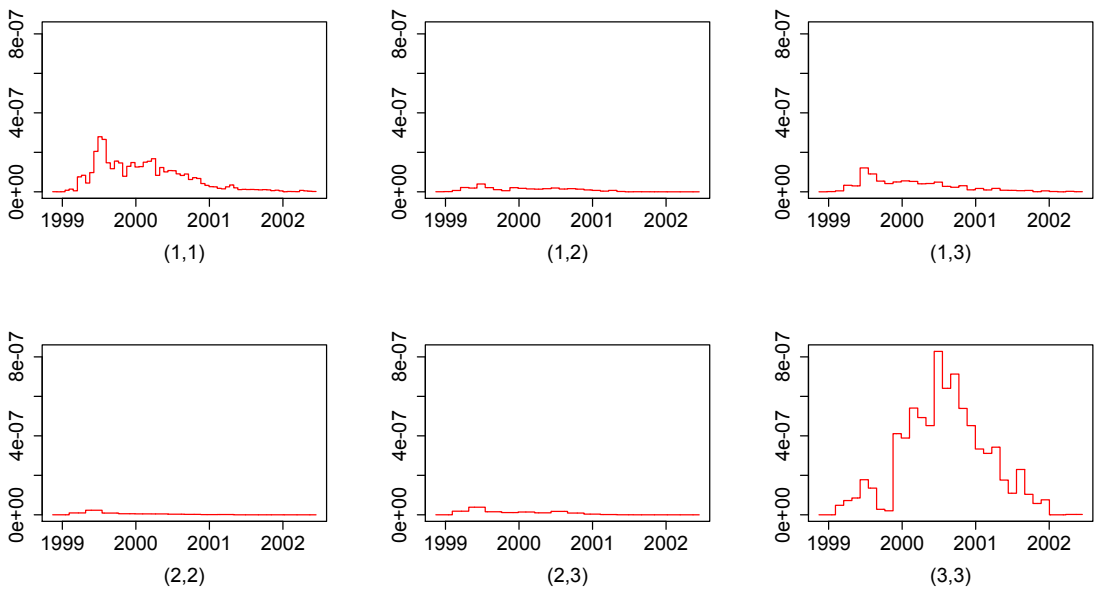

Figure 15: Estimated intensities for the Enron dataset using $Q = 3$ groups and a dyadic partition with $D_{\max} = 1024$.

Now we would like to compare these results with those obtained using a classical SBM. Note that taking $D_{\max} = 1$ in our approach amounts to forget the timestamps of the emails, as the algorithm then only considers email counts over the whole observation period. In other words, using $D_{\max} = 1$ boils down to rely on a classical SBM with Poisson emission distribution and mean parameter $A^{(q,l)}(T)$ (see for instance Mariadassou et al., 2010). Thus, we compared the classifications obtained by the two procedures (PPSBM and classical SBM) for different values of Q. Figure 17 shows the ARI values between those two classifications. For $Q = 2$ the classification of the n individuals is very similar with both models. However,

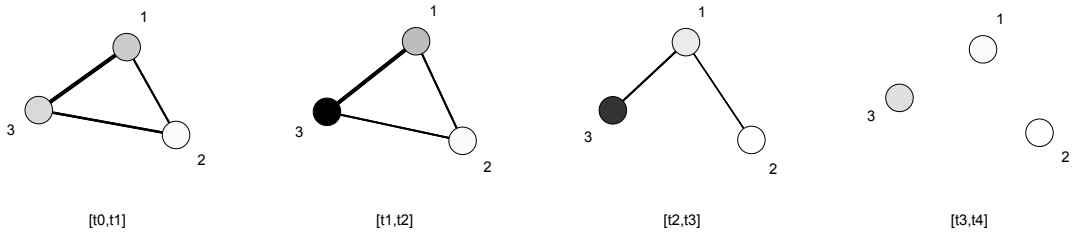

Figure 16: Temporal evolution of inter- and intra-communication activity of the $Q = 3$ groups for the Enron dataset, for a partition of the observation time into 4 regular intervals. Line widths (inter groups) and different shades of gray of the nodes (intra groups) indicate the mean intensities $\int_{t_{k-1}}^{t_k} \hat{\alpha}^{(q,l)}(s)ds$ over the considered subinterval, where darker gray shades represent higher intensities.

when increasing Q, the two classifications tend to differ more and more, without ever becoming completely discordant (the smallest ARI value, around 0.3 is still not negative). When $Q = 3$, a closer look on the classifications obtained by PPSBM and classical SBM shows that differences mainly concern group 1 and 3. That is, it depends much on the model, whether an individual is associated with group 1 or 3. Indeed, both groups are characterized by a high level of intra-communication and a smaller amount of communication with the other group. Intuitively, it may be difficult to distinguish group 1 from group 3 when looking only at the total count data as in classical SBM. However, the temporal distribution of the intra-group communication is different for both groups: in particular, the intra-group communication of group 3 achieves its maximum during the second half of year 2000, while the intensity of the intra-group communication of group 1 declines during this same period (Figure 15). This means that taking into account the time information of the events may be very useful to improve the classification of the individuals compared to classical SBM.

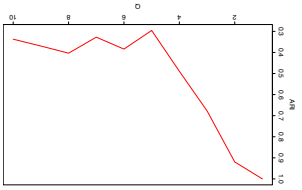

Figure 17: ARI for Enron dataset between our classification and a classification obtained by a simple SBM as a function of the number of groups Q.

Acknowledgments. We would like to thank Agathe Guilloux for pointing out valuable references, Nathalie Eisenbaum for her help on doubly stochastic counting processes and Pierre Latouche for sharing information on datasets.

References

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

P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993.

Y. Baraud and L. Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Related Fields*, 143(1-2):239–284, 2009.

P. Bickel, D. Choi, X. Chang, and H. Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.*, 41(4):1922–1943, 08 2013.

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

D. Böhning. A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference*, 47(1–2):5 – 28, 1995.

L. Bordes, D. Chauveau, and P. Vandekerkhove. A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics & Data Analysis*, 51(11):5429 – 5443, 2007.

C. T. Butts. A relational event framework for social action. *Sociological Methodology*, 38 (1):155–200, 2008.

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

Y.-S. Cho, A. Galstyan, P. J. Brantingham, and G. Tita. Latent self-exciting point process model for spatial-temporal networks. *Discrete and Continuous Dynamical Systems - Series B*, 19(5):1335–1354, 2014.

D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2003. Elementary theory and methods.

J. Dannemann. Semiparametric Hidden Markov models. *J. Comput. Graph. Statist.*, 21 (3):677–692, 2012.

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

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

C. DuBois, C. T. Butts, and P. Smyth. Stochastic blockmodeling of relational event dynamics. In *AISTATS*, volume 31, pages 238–246. JMLR Workshop and Conference Proceedings, 2013.

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

V. Gemmetto, A. Barrat, and C. Cattuto. Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC infectious diseases*, 14(1):695, 2014.

G. Grégoire. Least squares cross-validation for counting process intensities. *Scandinavian Journal of Statistics*, 20(4):pp. 343–360, 1993.

R. Guigourès, M. Boullé, and F. Rossi. Discovering patterns in time-varying graphs: a triclustering approach. *Adv Data Anal Classif*, To appear.

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

P. Holme. Modern temporal network theory: a colloquium. *Eur. Phys. J. B*, 88(9):234, 2015.

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

R. Jeanson. Long-term dynamics in proximity networks in ants. *Animal Behaviour*, 83(4): 915 – 923, 2012.

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

B. Klimt and Y. Yang. The Enron Corpus: A new dataset for email classification research. In J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, editors, *Machine Learning: ECML 2004*, volume 3201 of *Lecture Notes in Computer Science*, pages 217–226. Springer Berlin Heidelberg, 2004.

G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

S. Linderman and R. Adams. Discovering latent network structure in point process data. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, volume 32, page 1413–1421, 2014.

R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. N. Amaral. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105(47):18153–18158, 2008.

R. D. Malmgren, J. M. Hofman, L. A. Amaral, and D. J. Watts. Characterizing individual communication patterns. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 607–616. ACM, 2009.

M. Mariadassou and C. Matias. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21(1):537–573, 2015.

M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: a variational approach. *Ann. Appl. Stat.*, 4(2):715–42, 2010.

T. Martinussen and T. H. Scheike. *Dynamic regression models for survival data*. Statistics for Biology and Health. Springer, New York, 2006.

C. Matias and V. Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. Technical report, ArXiv:1506.07464, 2015.

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

N. Navaroli, C. DuBois, and P. Smyth. Statistical models for exploring individual email communication behavior. In S. Hoi and W. Buntine, editors, *Proceedings of the 4th Asian conference on machine learning (ACML 2012)*, volume 25, pages 317–332. JMLR Workshop and Conference Proceedings, 2012.

P. O. Perry and P. J. Wolfe. Point process modelling for directed interaction networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(5):821–849, 2013.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`.

H. Ramlau-Hansen. Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, 11(2):pp. 453–466, 1983.

P. Reynaud-Bouret. Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli*, 12(4):633–661, 2006.

S. Robin, A. Bar-Hen, J.-J. Daudin, and L. Pierre. A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics & Data Analysis*, 51(12):5483 – 5493, 2007.

T. Snijders and M. van Duijn. Simulation for statistical inference in dynamic network models. In R. Conte, R. Hegselmann, and P. Terna, editors, *Simulating Social Phenomena*, volume 456 of *Lecture Notes in Economics and Mathematical Systems*, pages 493–512. Springer Berlin Heidelberg, 1997.

T. A. B. Snijders, J. Koskinen, and M. Schweinberger. Maximum likelihood estimation for social network dynamics. *Ann. Appl. Stat.*, 4(2):567–588, 06 2010.

J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, and et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8): e23176, 2011.

D. Q. Vu, D. Hunter, P. Smyth, and A. U. Asuncion. Continuous-time regression models for longitudinal networks. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2492–2500. Curran Associates, Inc., 2011.

S. Wasserman. A stochastic model for directed graphs with transition rates determined by reciprocity. *Sociological Methodology*, 11:pp. 392–412, 1980a.

S. Wasserman. Analyzing social networks as stochastic processes. *Journal of the American Statistical Association*, 75(370):280–294, 1980b.

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

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

A Identifiability proofs

Proof of Proposition 1. We start by considering the distribution of one marginal process $N_{i,j}$, which is a Cox process directed by the random measure $A_{i,j}$ such that

$$A_{i,j} \sim \sum_{q=1}^{Q} \sum_{l=1}^{Q} \pi_q \pi_l \delta_{A^{(q,l)}}.$$

(Here, for any $(q,l) \in \mathcal{Q}$, we use the notation $A^{(q,l)}$ for the measure on $[0,T]$ defined by $A^{(q,l)}(I) = \int_I \alpha^{(q,l)}(u)du$ for all measurable $I \subset [0,T]$. We also recall that δ_u is the Dirac mass at point u). It is known that the mapping of probability laws of random measures into laws of Cox processes directed by them is a bijection (see for example Proposition 6.2.II in Daley and Vere-Jones, 2003). In other words, here the distribution of $N_{i,j}$ uniquely determines the finite measure (on the set of measures on $[0,T]$) $\sum_{q=1}^{Q} \sum_{l=1}^{Q} \pi_q \pi_l \delta_{A^{(q,l)}}$. Then, under Assumption 1 that the intensities $\alpha^{(q,l)}$ are distinct, the corresponding measures $A^{(q,l)}$ are all different and we may recover from the distribution of our counting process $N_{i,j}$ the set of values $\{(\pi_q^2, A^{(q,q)}); 1 \leq q \leq Q\} \cup \{(2\pi_q\pi_l, A^{(q,l)}); 1 \leq q < l \leq Q\}$ or equivalently the set $\{(\pi_q^2, \alpha^{(q,q)}); 1 \leq q \leq Q\} \cup \{(2\pi_q\pi_l, \alpha^{(q,l)}); 1 \leq q < l \leq Q\}$. In particular, we recover the functions $\alpha^{(q,l)}$ almost everywhere on $[0,T]$, up to a permutation of these $Q(Q+1)/2$ values. However, to recover those values up to a permutation in $\mathfrak{S}_Q$, it is necessary to consider higher-order marginals.

We now fix three distinct indices $1 \leq i,j,k \leq n$ and consider the trivariate counting process $(N_{i,j}, N_{i,k}, N_{j,k})$. In the same way, these are Cox processes directed by the triplet of random measures $(A_{i,j}, A_{i,k}, A_{j,k})$ such that

$$(A_{i,j}, A_{i,k}, A_{j,k}) \sim \sum_{1 \leq q,l,m \leq Q} \pi_q \pi_l \pi_m \delta_{(A^{(q,l)}, A^{(q,m)}, A^{(l,m)})}.$$

We write this distribution in such a way that distinct components appear only once

$$\begin{aligned}
&\sum_{q=1}^{Q} \pi_q^3 \delta_{(A^{(q,q)}, A^{(q,q)}, A^{(q,q)})} \\
&+ \sum_{1 \leq q \neq l \leq Q} \pi_q^2 \pi_l \left[\delta_{(A^{(q,q)}, A^{(q,l)}, A^{(q,l)})} + \delta_{(A^{(q,l)}, A^{(q,q)}, A^{(q,l)})} + \delta_{(A^{(q,l)}, A^{(q,l)}, A^{(q,q)})} \right] \\
&+ \sum_{\substack{q,l,m \\ |\{q,l,m\}|=3}} \pi_q \pi_l \pi_m \delta_{(A^{(q,l)}, A^{(q,m)}, A^{(l,m)})}.
\end{aligned} \tag{29}$$

Using the same reasoning, we identify the triplets of values $\{(A^{(q,l)}, A^{(q,m)}, A^{(l,m)}); 1 \leq q, l, m \leq Q\}$ up to a permutation on the triplets (q, l, m). Among these, the only values with three identical components are $\{(A^{(q,q)}; A^{(q,q)}; A^{(q,q)}); 1 \leq q \leq Q\}$ and thus the measures $\{A^{(q,q)}\}_q$ are identifiable, up to a permutation in $\mathfrak{S}_Q$. Going back to (29) and looking for the Dirac terms at points that have two identical components (of the form $(A^{(q,q)}, A^{(q,l)}, A^{(q,l)})$ and two other with permuted components), we can now identify the set of measures

$$\{(A^{(q,q)}, \{A^{(q,l)}\}_{l \neq q}); 1 \leq q \leq Q\}.$$

This is equivalent to saying that we identify the measures $\{A^{(q,l)}; (q,l) \in \mathcal{Q}\}$ up to a permutation in $\mathfrak{S}_Q$. Obviously, this also identifies the corresponding intensities $\{\alpha^{(q,l)}; (q,l) \in \mathcal{Q}\}$ almost everywhere on $[0,T]$, up to a permutation in $\mathfrak{S}_Q$. To finish the proof, we need to identify the proportions π_q. Note that as we identified the components $\{A^{(q,q)}; 1 \leq q \leq Q\}$, we recover from (29) the set of values $\{\pi_q^3; 1 \leq q \leq Q\}$ up to the same permutation as on the $A^{(q,q)}$'s. This concludes the proof. □

Proof of Proposition 2. We follow some of the arguments already appearing in the proof of Proposition 1. Let A^{in} (resp. A^{out}) denote the measure whose intensity is α^{in} (resp. α^{out}) The univariate process $N_{i,j}$ is a Cox process directed by the random measure $A_{i,j}$ that is now distributed as

$$A_{i,j} \sim (\sum_{q=1}^{Q} \pi_q^2)\delta_{A^{\text{in}}} + (\sum_{1 \leq q \neq l \leq Q} \pi_q \pi_l)\delta_{A^{\text{out}}}.$$

Thus the measures A^{in} and A^{out} are identifiable from the distribution of $N_{i,j}$, but only up to a permutation. Once again, we rather consider the trivariate Cox processes $(N_{i,j}, N_{i,k}, N_{j,k})$ directed by the random measures $(A_{i,j}, A_{i,k}, A_{j,k})$ whose distribution in the affiliation case has now five atoms

$$\Big(\sum_{q=1}^{Q} \pi_q^3\Big)\delta_{(A^{\text{in}}, A^{\text{in}}, A^{\text{in}})} + \Big(\sum_{q \neq l} \pi_q^2 \pi_l\Big)\delta_{(A^{\text{in}}, A^{\text{out}}, A^{\text{out}})} + \Big(\sum_{q \neq l} \pi_q^2 \pi_l\Big)\delta_{(A^{\text{out}}, A^{\text{in}}, A^{\text{out}})} + \Big(\sum_{q \neq l} \pi_q^2 \pi_l\Big)\delta_{(A^{\text{out}}, A^{\text{out}}, A^{\text{in}})} + \Big(\sum_{\substack{q,l,m \\ |\{q,l,m\}|=3}} \pi_q \pi_l \pi_m\Big)\delta_{(A^{\text{out}}, A^{\text{out}}, A^{\text{out}})}.$$

As previously, these five components are identifiable, up to a permutation on $\mathfrak{S}_5$. Now it is easy to identify the three components for which two marginals have same parameters and the third one has a different parameter. Thus, we recover exactly the measures A^{in} and A^{out}. This also identifies the corresponding intensities α^{in} and α^{out} almost everywhere on $[0,T]$.

Now, the identification of the proportions $\{\pi_q\}_q$ follows an argument already used in the proof of Theorem 13 in Allman et al. (2011) that we recall here for completeness. From the trivariate distribution of $(N_{i,j}, N_{i,k}, N_{j,k})$ and the already recovered values A^{in} and A^{out}, we identify the proportion $\sum_q \pi_q^3$. Similarly, for any $n \geq 1$, by considering the multivariate distribution of $(N_{i,j})_{(i,j) \in \mathcal{R}}$, we can identify the Dirac mass at point $(A^{\text{in}}, \ldots, A^{\text{in}})$ and thus its weight, which is equal to $\sum_q \pi_q^n$. By the Newton identities, the values $\{\sum_q \pi_q^n; 1 \leq n \leq Q\}$ determine the values of elementary symmetric polynomials $\{\sigma_n(\pi_1, \ldots, \pi_Q); 1 \leq n \leq Q\}$.

These, in turn, are (up to sign) the coefficients of the monic polynomial whose roots (with multiplicities) are precisely $\{\pi_q; 1 \leq q \leq Q\}$. Thus, the proportion parameters are recovered up to a permutation. □