A semiparametric extension of the stochastic block model for longitudinal networks

Catherine Matias, Tabea Rebafka, Fanny Villers

A semiparametric extension of the 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

To model recurrent interaction events in continuous time, we propose an extension of the stochastic block model 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 shown to be identifiable and an estimation procedure is proposed based on a semiparametric variational expectation-maximization algorithm. Two versions of the method are developed, using either a nonparametric histogram approach (with an adaptive choice of the partition size) or kernel intensity estimators. The number of latent groups can be selected by an integrated classification likelihood criterion. Finally, we demonstrate the performance of our procedure on synthetic experiments and the analysis of several real datasets illustrates the utility of our approach.

Keywords: dynamic interactions; expectation-maximization algorithm; integrated classification likelihood; link streams; longitudinal network; semiparametric model; stochastic block model; variational approximation.

1 Introduction

The past few years have seen a large increase in the interest for modeling dynamic interactions between individuals. While many real world data contain continuous-time information on the interactions, as e.g. email exchanges between employees in a company (Klimt & Yang, 2004) or face-to-face contact between individuals measured through sensors (Stehlé et al., 2011), most models are discrete in time. Commonly, data are aggregated on predefined time intervals to obtain a sequence of snapshots of interaction random graphs. Besides the loss of information induced by data aggregation, the specific 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, link streams* 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 is based on stochastic point processes. An important line of research involves continuous-time Markov processes with seminal works on dyad-independent models (Wasserman, 1980b,a) up to the development of so-called stochastic actor oriented models (e.g. Snijders & van Duijn, 1997; Snijders et al., 2010). In these works observations consist in a series of time intervals of interaction and interactions are assumed to last during the whole corresponding time interval. Here, we focus on a rather different setup where each interaction is identified with a time point. Furthermore, we consider a model that allows for dependencies of the processes modeling the interactions of pairs of individuals.

The analysis of event data is an old and important area in statistics (see e.g. Andersen et al., 1993). Generally a multivariate counting process $N(t) = (N_{i,j}(t))_{(i,j)}$ is considered, that counts the number of interactions of each pair (i, j) of individuals 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, in addition, are marked by a label (representing a behavioral event). The model may be viewed as an instance of Cox's multiplicative hazard model with time-dependent covariates and constant baseline function. In the same vein, Vu et al. (2011) propose a general regression-based modeling of the intensity of non recurrent interaction events. They consider two different frameworks: Cox's multiplicative and Aalen's additive hazard rates (see e. g. Martinussen & Scheike, 2006). Perry & Wolfe (2013) propose another variant of Cox's multiplicative intensity model for recurrent interaction events where the baseline function is specific to each individual. In the above mentioned works a set of statistics is chosen by the user as potential candidates that modulate the interactions. As in any regression framework, the choice of 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.

The approaches by Butts, Vu et al., Perry & Wolfe and others are based on conditional Poisson processes characterized by random intensities, also known as doubly stochastic Poisson processes or Cox processes. A particular instance of the conditional Poisson process is the Hawkes process, which is a collection of point processes with some background rate, where 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, based on such self-exciting point processes for temporal dynamics combined with a Gaussian mixture for the spatial dynamics. Similarly, Linderman & Adams (2014) combine temporal Hawkes processes with latent distance models for implicit networks that cannot be observed directly.

Clustering individuals based on interaction data represents a well-established technique for taking into account the intrinsic heterogeneity and summarizing information. 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 to a dynamic context (Yang et al., 2011; Xu & Hero, 2014; Matias & Miele, to appear; Corneli et al., 2016). Stochastic block models posit that each individual belongs to a latent group and interactions between two individuals are conditionally independent of the interactions of any other pair, given the latent groups of the interacting individuals. Another attempt to use stochastic block models in the context of interaction events appears in DuBois et al.

(2013) generalizing the approach of Butts (2008) by adding discrete latent variables on the individuals.

In this work a semiparametric stochastic block model for recurrent interaction events in continuous time is introduced, to which we refer as the Poisson process stochastic block model. This is a stochastic block model where interactions are modeled by conditional inhomogeneous Poisson processes, whose intensities only depend on the latent groups of the interacting individuals. In contrast to many other works, we do not rely on a parametric model where intensities are modulated by predefined network statistics, but intensities are modeled and estimated in a nonparametric way. The model is shown to be identifiable. Our estimation and clustering approach is a semiparametric version of the variational expectation-maximization algorithm, where the maximization step is replaced by nonparametric estimators of the intensities. Semiparametric generalizations of the classical expectation-maximization (`EM`) algorithm have been proposed in many different contexts (see 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 nonparametric estimates in a variational approximation of `EM`. Two versions are developed for the nonparametric part of the model: a histogram approach based on the work of Reynaud-Bouret (2006) and a kernel estimator based on Ramlau-Hansen (1983). For the histogram approach, an integrated classification likelihood criterion is proposed to select the number of latent groups adaptively. Synthetic experiments enlighten both the clustering capacities of our method as well as the performance of the nonparametric estimation of the different intensities. Moreover, the analysis of several real datasets illustrates the strengths and weaknesses of our approach. The Supplementary Material, whose references appear as S.xx, provides the proofs of all theoretical results, technical details on the algorithm and more detailed results of the analysis of the real data examples.

2 A semiparametric Poisson process stochastic block model

2.1 Model

We are interested in the pairwise interactions of n individuals during some time interval $[0, T]$. For notational convenience, we choose to restrict our attention to directed interactions without self-interactions. The undirected case as well as self-interactions are treated similarly and simulations as well as a real data example from Section 5 use the undirected setup. The set of all possible pairs of individuals, which is also the set of all possible dyads in the graph, is denoted by

$$\mathcal{R} = \{(i, j), i, j = 1, \dots, n, i \neq j\}.$$

The cardinality of $\mathcal{R}$ is $r = n(n-1)$. The observations $\mathcal{O}$ are the interactions occurring in time interval $[0, T]$, that is

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

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. The number of events in time interval $[0, T]$ is M. We assume that $0 < t_1 < \cdots < t_M < T$, i.e. there is at most one event at a time.

To model the distribution of these observations, every individual is assumed to belong to one out of Q groups, and the relation between two individuals, that is the way they interact with another, is driven by their group membership. More precisely, let $Z_1, \ldots, Z_n$ be independent and identically distributed (latent) random variables taking values in $\{1, \ldots, Q\}$ with non zero probabilities

$$\pi_q = \Pr(Z_1 = q) \quad (q = 1, \ldots, Q).$$

For the moment, Q is considered to be fixed and known. When no confusion occurs, we also use the notation $Z_i = (Z^{i,1}, \ldots, Z^{i,Q})$ with $Z^{i,q} \in \{0, 1\}$ such that Z_i has multinomial distribution $\mathcal{M}(1, \pi)$ with $\pi = (\pi_1, \ldots, \pi_Q)$.

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 and Z_j, are modeled by a conditional inhomogeneous Poisson process $N_{i,j}(\cdot)$ on $[0, T]$ with intensity depending only on the latent groups Z_i and Z_j. We consider nonnegative intensity functions $\alpha^{(q,l)}$ with $q, l = 1, \ldots, Q$ such that the conditional intensity of process $N_{i,j}(\cdot)$, given that $Z_i = q$ and $Z_j = l$, is $\alpha^{(q,l)}(\cdot)$ for any $(i, j) \in \mathcal{R}$. The corresponding cumulative intensities are denoted by

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

The set of observations $\mathcal{O}$ is a realization of the multivariate counting process $(N_{i,j}(\cdot))_{(i,j)\in\mathcal{R}}$ with conditional intensity process $(\alpha^{(Z_i,Z_j)}(\cdot))_{(i,j)\in\mathcal{R}}$. The process $N_{i,j}$ is not a Poisson process, but a counting process with intensity $\sum_{q=1}^{Q}\sum_{l=1}^{Q} \pi_q \pi_l \alpha^{(q,l)}$. We denote $\theta = (\pi, \alpha)$ the infinite-dimensional parameter of a Poisson process stochastic block model. The distribution of the multivariate counting process $(N_{i,j}(\cdot))_{(i,j)\in\mathcal{R}}$ under parameter value θ is denoted $\Pr_\theta$.

2.2 Identifiability

Concerning the identifiability of parameter θ from the distribution of the multivariate counting process $(N_{i,j}(\cdot))_{(i,j)\in\mathcal{R}}$, it is clear that at best the Poisson process stochastic block model is identifiable up to label switching, as defined below. Furthermore, as the functions $\alpha^{(q,l)}$ are intensities, they are only identifiable almost everywhere on $[0, T]$. We denote $\mathfrak{S}_Q$ the set of permutations of $\{1, \ldots, Q\}$.

Definition 1 (Identifiability up to label switching)**.** *The parameter $\theta = (\pi, \alpha)$ of a Poisson process stochastic block model is identifiable on $[0, T]$ up to label switching if for all θ and $\tilde{\theta}$ such that $\mathbb{P}_\theta = \mathbb{P}_{\tilde{\theta}}$, there exists a permutation $\sigma \in \mathfrak{S}_Q$ such that*

$$\pi_q = \tilde{\pi}_{\sigma(q)}, \quad \alpha^{(q,l)} = \tilde{\alpha}^{(\sigma(q),\sigma(l))} \text{ almost everywhere on } [0, T], \qquad (q, l = 1, \ldots, Q).$$

The following assumption ensures identifiability up to label switching in a very general setting.

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

The intensities $\alpha^{(q,l)}$ may take identical values at some points or on 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 Poisson process stochastic block model distribution of the multivariate counting process $(N_{i,j}(\cdot))_{(i,j)\in\mathcal{R}}$ on the same interval, as soon as $n \geq 3$.*

This result strongly relies on the only available identifiability result for weighted stochastic block models, namely Theorem 12 in Allman et al. (2011), which can be applied under Assumption 1. One may wonder whether the necessary condition that any two rows (or any two columns) of the parameter matrix α are distinct is, in fact, a sufficient condition for identifiability. However, to our knowledge such a result has never been established even in the simple binary case. In the binary stochastic block model, the results in Allman et al. (2009, 2011) establish *generic* identifiability, which means identifiability except on a subset of parameters with Lebesgue measure zero, without specifying the exceptional subset. For the directed and binary stochastic block model, Celisse et al. (2012) establish identifiability under the assumption that the product vector $\alpha\pi$ (or $\pi^\intercal\alpha$) has distinct coordinates. This condition is slightly stronger than the one previously mentioned. Another partial identifiability result appears in Bickel et al. (2011) for some block models. These last two approaches are specifically adapted to the discrete setup (maybe even to the binary one) and cannot be generalized to the continuous case.

Proposition 1 does not cover the affiliation case, where only two intensities α^{in} and α^{out} are considered such that for all (q,l)

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

Proposition 2. *If 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 Poisson process stochastic block model 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, for any $n \geq \max\{Q,3\}$, the proportions $\pi_1,\ldots,\pi_Q$ are also identifiable, up to a permutation, from the same distribution.*

2.3 Description of relevant processes

In this section we introduce processes and notation that will be used throughout the manuscript. First, for any q,l, we consider the (unobserved) number of dyads $(i,j) \in \mathcal{R}$ with latent groups (q,l)

$$Y^{(q,l)} = \sum_{(i,j)\in\mathcal{R}} Z^{i,q} Z^{j,l}. \tag{1}$$

The (unobserved) counting process

$$N_{\mathcal{Z}}^{(q,l)} = \sum_{(i,j)\in\mathcal{R}} Z^{i,q} Z^{j,l} \tag{2}$$

has conditional intensity $Y^{(q,l)}\alpha^{(q,l)}$ and falls in the class of Aalen's multiplicative intensity models. This is a central property 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 groups (q,l) by

$$Z_m^{(q,l)} = Z^{i_m,q} Z^{j_m,l}. \tag{3}$$

As these quantities are unobserved, our work relies on proxies. We consider the set $\mathcal{T}$ of candidate proxies for the unobserved latent groups $Z^{i,q}$ given by

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

While the latent variables $Z^{i,q}$ are indicators, their counterparts $\tau^{i,q}$ are weights representing the probability that node i belongs to group q. Now, for every $\tau \in \mathcal{T}$, replacing all latent variables $Z^{i,q}$ in (1)–(3) by $\tau^{i,q}$, we define $\bar{Y}^{(q,l)}$, $N^{(q,l)}$ and $\tau_m^{(q,l)}$ which are estimators of $Y^{(q,l)}$, $N_{\mathcal{Z}}^{(q,l)}$ and $Z_m^{(q,l)}$, respectively.

3 Semiparametric estimation procedure

3.1 A variational semiparametric expectation-maximization algorithm

The complete-data likelihood of observation $\mathcal{O}$ and latent variables $\mathcal{Z} = (Z_1, \dots, Z_n)$ is

$$\begin{aligned} \mathcal{L}(\mathcal{O}, \mathcal{Z}|\theta) &= \mathcal{L}(\mathcal{O}|\mathcal{Z}, \theta) \times \mathcal{L}(\mathcal{Z}|\theta) \\ &= \exp\left\{ - \sum_{(i,j)\in\mathcal{R}} A^{(Z_i,Z_j)}(T) \right\} \prod_{m=1}^{M} \alpha^{(Z_{i_m},Z_{j_m})}(t_m) \prod_{i=1}^{n} \prod_{q=1}^{Q} \pi_q^{Z^{i,q}}. \end{aligned} \tag{5}$$

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 $\mathcal{Z}$. This set is so huge that the likelihood of the observed data is intractable for direct maximization. Hence, an expectation-maximization (`EM`) algorithm (Dempster et al., 1977) is used, which is an iterative procedure especially adapted to cope with latent variables. The `EM` algorithm consists of an `E`-step and an `M`-step that are iterated until convergence. In our model two different issues arise. First, as already observed for the standard stochastic block model (Daudin et al., 2008), the `E`-step requires the computation of the conditional distribution of $\mathcal{Z}$ given the observations $\mathcal{O}$, which is not tractable. Therefore, we use a variational approximation (Jordan et al., 1999) of the latent variables conditional distribution to perform the `E`-step. We refer, for instance, to Matias & Robin (2014) for a general description of the variational `EM` algorithm and its links to `EM` in stochastic block models. 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. Our complete algorithm is summarized in Section 3.6, Algorithm 1.

3.2 Variational `E`-step

The standard `E`-step consists in computing the expectation of the complete log-likelihood given the observations at some current parameter value θ. This requires the knowledge

of the conditional latent variables distribution $\Pr_\theta(\mathcal{Z}|\mathcal{O})$, which is not tractable, mainly because the latent variables Z_i are not conditionally independent. The idea is to perform a variational approximation of the conditional latent variables distribution $\Pr_\theta(\cdot|\mathcal{O})$ by a simpler distribution. More precisely, using the class of parameters $\mathcal{T}$ defined in (4), we consider for every $\tau \in \mathcal{T}$ the conditional factorized distribution $\Pr_\tau(\cdot|\mathcal{O})$ of $\mathcal{Z}$ given $\mathcal{O}$ defined by

$$\Pr_\tau\left(\mathcal{Z} = (q_1, \ldots, q_n)|\mathcal{O}\right) = \prod_{i=1}^{n} \Pr_\tau(Z_i = q_i|\mathcal{O}) = \prod_{i=1}^{n} \tau^{i,q_i}, \qquad (q_1, \ldots, q_n) \in \{1, \ldots, Q\}^n, \tag{6}$$

with corresponding expectation $E_\tau(\cdot|\mathcal{O})$. Then, we search for the parameter $\hat{\tau} \in \mathcal{T}$ that yields the best approximation $\Pr_\tau(\cdot|\mathcal{O})$ of $\Pr_\theta(\cdot|\mathcal{O})$. More precisely,

$$\hat{\tau} = \underset{\tau \in \mathcal{T}}{\text{Argmin}}\, KL\left(\Pr_\tau(\cdot|\mathcal{O})||\Pr_\theta(\cdot|\mathcal{O})\right), \tag{7}$$

where $KL(\cdot||\cdot)$ denotes the Kullback-Leibler divergence. The variational `E`-step is completed by the computation of the current expected complete data log-likelihood

$$Q_\tau(\theta) = E_\tau\left(\log \mathcal{L}(\mathcal{O}, \mathcal{Z}|\theta)|\mathcal{O}\right).$$

From a practical point of view, it can be shown that the solution $\hat{\tau}$ of (7) is also the solution of a fixed point equation, which in practice is found by successively updating the variational parameters $\tau^{i,q}$ via the following Equation (8) until convergence.

Proposition 3. *The solution $\hat{\tau}$ to the minimization problem given in* (7) *satisfies the following fixed-point equation*

$$\hat{\tau}^{i,q} \propto \pi_q \exp[D_{iq}(\hat{\tau}, \alpha)], \qquad (i = 1, \ldots, n,\ q = 1, \ldots, Q), \tag{8}$$

where $\propto$ means 'proportional to' and

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

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

3.3 Nonparametric `M`-step: general description

In a parametric context, the `M`-step consists in the maximization of $Q_\tau(\theta)$ with respect to $\theta = (\pi, \alpha)$. Considering only the finite-dimensional part π of the parameter, we easily obtain that the maximizer $\hat{\pi}$ of $Q_\tau(\pi, \alpha)$ with respect to π is

$$\hat{\pi}_q = \frac{\sum_{i=1}^{n} \tau^{i,q}}{\sum_{q=1}^{Q} \sum_{i=1}^{n} \tau^{i,q}} = \frac{1}{n} \sum_{i=1}^{n} \tau^{i,q}. \tag{9}$$

Concerning the infinite-dimensional parameter α, we replace the maximization of $Q_\tau(\pi, \alpha)$ with respect to α by a nonparametric estimation step. In the following, we develop two different approaches for updating α: a histogram and a kernel method. In both cases estimation would be straightforward using the process $(N_{\mathcal{Z}}^{(q,l)})_{q,l}$ defined by (2), which unfortunately is not observed. It is thus natural to use its (current) variational approximation, namely the weighted cumulative process $(N^{(q,l)})_{q,l}$ defined in Section 2.3.

3.4 Histogram-based M-step

In this part the intensities $\alpha^{(q,l)}$ are estimated by piecewise constant functions and we propose a data-driven choice of the partition of the time interval $[0, T]$. The procedure is based on a least-squares penalized criterion following the work of Reynaud-Bouret (2006). The detailed construction is provided in the Supplementary Material.

For $D = 1, \ldots, D_{\max}$, where $D_{\max}$ is to be chosen, we denote by $\mathcal{E}^D$ the regular partition of $[0, T]$ into D intervals with length T/D, namely

$$\mathcal{E}^D = \left\{ E_{k,D} = \Big[(k-1)\frac{T}{D}; k\frac{T}{D}\Big); k = 1, \ldots, D \right\}.$$

One may also use regular dyadic partitions, where $D = 2^d$, defining nested models that have some practical advantages concerning the implementation of the algorithm. In the following let (q, l) be fixed. For $k = 1, \ldots, D$ the estimated mean number of observed interactions between individuals (i_m, j_m) with latent groups (q, l) occurring in time interval $E_{k,D}$ is

$$N^{(q,l)}(k, D) = \int_{E_{k,D}} dN^{(q,l)}(s) = \sum_{m=1}^{M} \tau^{i_m,q} \tau^{j_m,l} \mathbb{1}_{E_{k,D}}(t_m). \tag{10}$$

For any fixed value of D, a projection estimator $\hat{\alpha}_D^{(q,l)}$ on the space of piecewise constant functions on $\mathcal{E}^D$ is given by

$$\hat{\alpha}_D^{(q,l)}(\cdot) = \frac{1}{\bar{Y}^{(q,l)}} \sum_{k=1}^{D} \frac{N^{(q,l)}(k, D)}{|E_{k,D}|} \mathbb{1}_{E_{k,D}}(\cdot),$$

where $|E_{k,D}| = T/D$ is the length of interval $E_{k,D}$. Now, adaptive estimation consists in choosing the best estimator among the collection of estimators $\{\hat{\alpha}_D^{(q,l)}, D = 1, \ldots, D_{\max}\}$. So we introduce an estimator of the partition through $\hat{D}$ that minimizes a penalized least-squares criterion, that simplifies to

$$\hat{D} = \hat{D}^{(q,l)} = \underset{D=1,\ldots,D_{\max}}{\text{Argmin}} \Big\{ -D \sum_{k=1}^{D} N^{(q,l)}(k, D)^2 + 2DD_{\max} \sup_{k=1,\ldots,D_{\max}} N^{(q,l)}(k, D_{\max}) \Big\}.$$

The selected partition size $\hat{D}$ depends on the groups (q, l) and may be different for different values of (q, l). Finally, the adaptive estimator of intensity $\alpha^{(q,l)}$ is

$$\hat{\alpha}_{\text{hist}}^{(q,l)}(t) = \hat{\alpha}_{\hat{D}}^{(q,l)}(t) = \frac{\hat{D}}{T\bar{Y}^{(q,l)}} \sum_{k=1}^{\hat{D}} N^{(q,l)}(k, \hat{D}) \mathbb{1}_{E_{k,\hat{D}}}(t). \tag{11}$$

Reynaud-Bouret (2006) develops her approach in the Aalen multiplicative intensity model, which is slightly different from our context. Moreover, our setup does not satisfy the assumptions of Theorem 1 in Reynaud-Bouret (2006), since the number of jumps of the processes $N_{i,j}$ is not bounded by a known positive number, because here the $N_{i,j}$ are counting processes. Nevertheless, in our simulations this procedure successfully estimates the intensities $\alpha^{(q,l)}$ (see Section 4). We refer to Baraud & Birgé (2009) for a theoretical study of an adaptive nonparametric estimation of the intensity of a Poisson process. Reynaud-Bouret (2006) also studies other penalized least squares estimators (for e.g. using Fourier bases), which might be used here similarly. An alternative way for nonparametric intensity estimation is based on kernel estimators, that are explored in the following section.

3.5 Kernel-based `M`-step

Kernel methods are suited to estimate smooth functions. 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). If the variational parameters $\tau^{i,q}$ are good approximations of the latent variables $Z^{i,q}$, then the intensity of process $N^{(q,l)}$ defined in Section 2.3 is approximately $\bar{Y}^{(q,l)}\alpha^{(q,l)}$, where $\bar{Y}^{(q,l)}$ is the variational mean number of dyads with latent groups (q,l). 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

$$\hat{\alpha}^{(q,l)}_{\text{ker}}(t) = \frac{1}{b\bar{Y}^{(q,l)}} \int_0^T K\Big(\frac{t-u}{b}\Big) dN^{(q,l)}(u) = \frac{1}{b\bar{Y}^{(q,l)}} \sum_{m=1}^{M} \tau_m^{(q,l)} K\Big(\frac{t-t_m}{b}\Big), \qquad (12)$$

if $\bar{Y}^{(q,l)} > 0$ and $\hat{\alpha}^{(q,l)}_{\text{ker}}(t) = 0$ otherwise, where $\tau_m^{(q,l)}$ is defined in Section 2.3. The bandwidth b can be chosen adaptively from the data following the procedure proposed by Grégoire (1993). Kernel methods are not always suited to infer a function on a bounded interval as boundary effects may deteriorate their quality. However, it is out of the scope of this work to investigate refinements of this kind.

3.6 Algorithm's full description

During the implementation of the algorithm, two issues arise: convergence and initialization. As our algorithm is an iterative procedure, one has to test for convergence. A stopping criterion can be defined based on the current expected complete data log-likelihood $Q_{\tau^{[s]}}(\theta^{[s]})$. Concerning initialization the algorithm be may run several times with different starting values. One can choose them randomly or by some k-means method. See the Supplementary Material for details. Algorithm 1 provides a full description of the procedure.

3.7 Model selection with respect to Q

To choose the best number of groups Q, we propose an integrated classification likelihood criterion that performs data-driven model selection. Roughly, this criterion is based on the complete data log-likelihood penalized by the number of parameters. It has been introduced in the mixture context in (Biernacki et al., 2000) and adapted to the stochastic block model in Daudin et al. (2008). The issue here is that our model contains a nonparametric part, so

Algorithm 1: Semiparametric variational expectation-maximization algorithm

$s \leftarrow 0$
Initialize $\tau^{[0]}$
while *convergence is not attained* **do**
  Update $\pi^{[s+1]}$ via Equation (9) with $\tau = \tau^{[s]}$
  Update $\alpha^{[s+1]}$ via either Equation (11) (histogram method) or (12) (kernel method) with $\tau = \tau^{[s]}$
  Update $\tau^{[s+1]}$ via the fixed-point equation (8) using $(\pi, \alpha) = (\pi^{[s+1]}, \alpha^{[s+1]})$
  Evaluate the stopping criterion
  $s \leftarrow s + 1$
Output $(\pi^{[s]}, \alpha^{[s]})$

that the parameter is infinite dimensional. However, in the case of histogram estimators, once the partition is selected, there is only a finite number of parameters to estimate, which can be used to build our integrated classification likelihood criterion.

More precisely, for any Q let $\hat{\theta}_Q$ be the estimated parameter with Q groups and $\hat{\mathcal{Z}}^Q$ the corresponding maximum a posteriori classification at $\hat{\theta}_Q$ obtained by our variational `EM` algorithm. 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 $\sum_{q,l} \hat{D}^{(q,l)}(Q)$, where $\hat{D}^{(q,l)}(Q)$ denotes the size of the partition used in the histogram estimator $\hat{\alpha}_{\text{hist}}^{(q,l)}(Q)$. In the adaptation of the integrated classification likelihood criterion to the stochastic block model these components are treated 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 $\log(r)/2$ term. We refer to Daudin et al. (2008) for more details. In our case, the integrated classification likelihood criterion is

$$\textsc{Icl}(Q) = \log \mathbb{P}_{\hat{\theta}_Q}(\mathcal{O}, \hat{\mathcal{Z}}^Q) - \frac{1}{2}(Q-1)\log(n) - \frac{1}{2}\log(r)\sum_{q=1}^{Q}\sum_{l=1}^{Q}\hat{D}^{(q,l)}(Q). \tag{13}$$

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

$$\hat{Q} = \underset{Q=1,\dots,Q_{\max}}{\text{Argmax}}\ \textsc{Icl}(Q). \tag{14}$$

4 Synthetic experiments

In this section we investigate the numerical performance of our method for clustering individuals and estimating the intensities of the inhomogeneous Poisson processes. We also study the performance of the integrated classification likelihood criterion for recovering the true number of latent groups. The following two scenarios are used in our simulations in the undirected setup where $\alpha^{(q,l)} = \alpha^{(l,q)}$ for any q, l.

1. We consider the affiliation model with $Q = 2$ latent groups and equal group probabilities $\pi_q = 1/2$. To evaluate the classification performance, the intensities are sinusoids

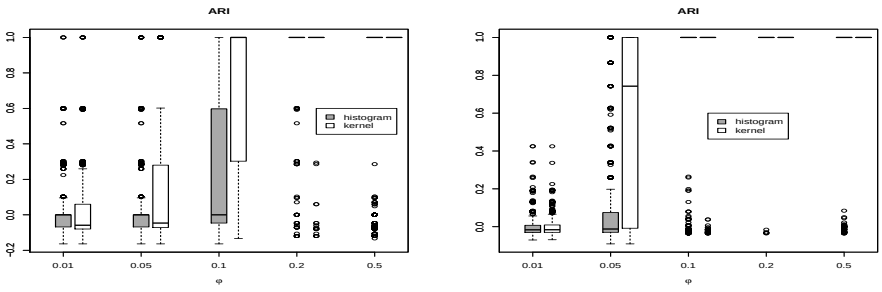

Figure 1: Boxplots of the adjusted rand index in Scenario 1 for histogram (gray) and kernel (white) estimators with $\varphi \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$. Left: $n = 10$, right: $n = 30$.

with varying shifting parameter φ. Clustering is supposed to be more difficult for small values of φ. The intensities are shown in Figure S.1 in the Supplementary Material. The number of individuals n varies in $\{10, 30\}$.

2. To evaluate the intensity estimators, we consider a Poisson process stochastic block model with $Q = 3$ groups with equal probabilities $\pi_q = 1/3$. The six intensity functions have rather different shapes and amplitudes (see Figure 2). The number of individuals n varies in $\{20, 50\}$.

For every setting, 1000 datasets are simulated under the corresponding Poisson process stochastic block model and the variational `EM` algorithm is applied. The histogram estimator is applied with a regular partition and $D_{\max} = 20$, while the kernel estimator uses the Epanechnikov kernel.

To assess the clustering performance, we use the adjusted rand index (Hubert & 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 adjusted rand index is smaller than 1 and negative values are possible. Figure 1 shows the boxplots of the adjusted rand index obtained with the histogram and the kernel versions of our method in Scenario 1. For small values of the shifting parameter ($\varphi \in \{0.01, 0.05\}$), the intensities are so close that the classification is very difficult, especially when $n = 10$ is small. The classification improves when the shift between the intensities and/or the number of observations increase, achieving (almost) perfect classification for large values of φ and/or n. 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.

Concerning the recovery of the intensities in Scenario 2, the quadratic risk is used to measure the distance between the true intensity $\alpha^{(q,l)}$ and its estimate $\hat{\alpha}^{(q,l)}$ defined by

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

Table 1 gives the mean value of the risk $\textsc{Risk}(q, l)$ and its standard deviation estimated over 1000 repetitions for both the histogram and the kernel version of our method. Table 1 also reports the mean number of observations (i_m, j_m) with latent groups (q, l), namely

Table 1: Mean number of events with latent groups (q, l) and mean quadratic risk (with standard deviation) for the histogram and the kernel estimators of the intensities in Scenario 2 averaged over 1000 repetitions. All values associated with the risk are multiplied by 100.

	Nb of events with group (q,l)		Histogram Risk(q,l)				Kernel Risk(q,l)			
Index (q,l)	$n=20$	$n=50$	$n=20$		$n=50$		$n=20$		$n=50$	
$(1,1)$	83	546	56	(62)	20	(19)	120	(64)	65	(12)
$(1,2)$	147	949	123	(52)	103	(5)	200	(28)	161	(7)
$(1,3)$	84	545	89	(53)	41	(7)	46	(36)	20	(6)
$(2,2)$	33	210	115	(43)	73	(8)	75	(48)	36	(9)
$(2,3)$	132	846	207	(52)	105	(14)	184	(28)	106	(11)
$(3,3)$	47	295	150	(54)	76	(17)	87	(46)	44	(12)

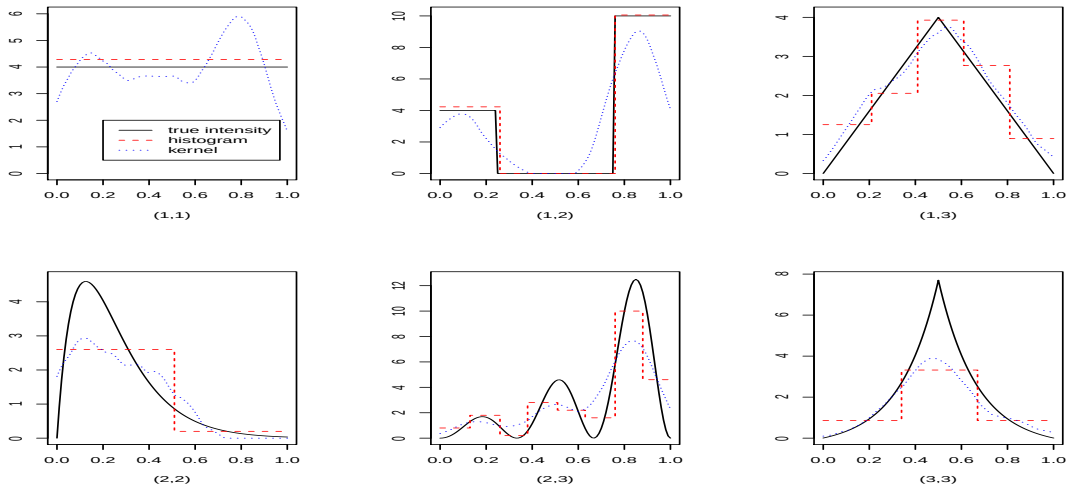

Figure 2: Scenario 2 with $n = 20$. True intensities (black continuous), histogram estimator (red dashed) and kernel estimator (blue dotted) for each pair of groups (q, l).

$\sum_{m=1}^{M} Z_m^{(q,l)}$. Moreover, Figure 2 shows for each pair of groups (q, l) the true and the estimated intensities for one dataset with $n = 20$.

As expected, when the true intensity is piecewise-constant, the histogram version of our method outperforms the kernel estimator. Conversely, when the true intensity is smooth, the kernel estimator is more appropriate to recover the shape of the intensity. In some cases, as e.g. for the intensity with latent groups $(2, 3)$, the estimators achieve comparable results. A well-known drawback of the kernel estimator is that it suffers from boundary effects. This is observed here for the intensities with groups $(1, 1)$ and $(1, 2)$, but 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 integrated classification likelihood criterion to select the number Q of latent groups from the data. For each of the 1000 simulated datasets, the maximizer $\hat{Q}$ of the integrated classification likelihood criterion defined in (14) with $Q_{\max} = 10$ is computed. Results are reported in Figure S.4 in the Supplementary Material. For $n = 20$ the correct number of groups is recovered in

95% of the cases. Moreover, when the criterion does not select the correct number Q, the adjusted rand index of the classification with three groups is rather low. This indicates that in those cases the classification obtained with three groups is not the correct one, so that rather the variational EM algorithm is to blame for bad results than the integrated classification likelihood criterion. For $n = 50$ our procedure selects the correct number of groups for each simulated datasets.

5 Real datasets

5.1 London cycles dataset

Here, we use the cycle hire usage data from the bike sharing system of the city of London from 2012 to 2015 (Transport for London, 2016). These data are also analyzed in Guigourès et al. (2015) with a different perspective. We focus on two randomly chosen weekdays, which are February 1st, 2012 (day 1) and February 2nd, 2012 (day 2). Data consist in pairs of stations associated with a single hiring/journey (departure station, ending station) and corresponding time stamp (hire time with second precision). The datasets have been pre-processed to remove journeys that either correspond to loops, last less than 1 minute or more than 3 hours or do not have an ending station (lost or stolen bikes). The datasets contain $n_1 = 415$ and $n_2 = 417$ stations on day 1 and day 2 with $M_1 = 17,631$ and $M_2 = 16,333$ hire events respectively. With more than 170,000 oriented pairs of stations the number of processes $N_{i,j}$ is huge, but only a very small fraction – around 7% – of these point processes are non null (i.e. contain at least one hiring event between these stations). This is to be expected as bike sharing systems are mostly used for short trips and stations far one from another are unlikely to be connected. As data correspond to origin/destination flows, it is natural to work with a directed setup and we applied the histograms version of our algorithm on a dyadic partition with maximum size $D_{\max} = 32$.

The integrated classification likelihood criterion achieves its maximum with $\hat{Q} = 6$ latent groups for day 2 and $\hat{Q} = 8$ on day 1. In order to compare results across the two datasets and keep interpretation simple, we focus on the classification obtained with $Q = 6$ clusters. Geographic locations of the bike stations and the clusters are represented on a city map (thanks to the OpenStreetMap project), see Figure S.5 in the Supplementary Material for day 2. Clusters for day 1 are very similar, so that in the following we only concentrate on day 2. We observe that our procedure globally recovers geographic clusters, as stations are expected to be mainly linked through geographic proximity in the datasets. A closer look at the clusters then reveals more information.

There is one cluster containing only four bike stations (cluster number 5, appearing as light blue diamonds ◇ in Figure S.5), while all other clusters contain between 38 and 125 stations. This small cluster contains one bike station at Kings Cross railway station and three stations next to Waterloo railway station. Indeed, two of these bike stations are among those with the highest activities (for both departures and arrivals) in comparison to all other stations, while this is not the case of the other two stations in the cluster. Thus it is very unlikely that a snapshot approach, where events are aggregated over a predetermined time window, would have yield the same cluster. Consequently, the explanation for this clustering is the similarity of the temporal profiles of these four stations. Indeed, Figure 3 shows that these four stations are 'outgoing' stations in the morning with much more

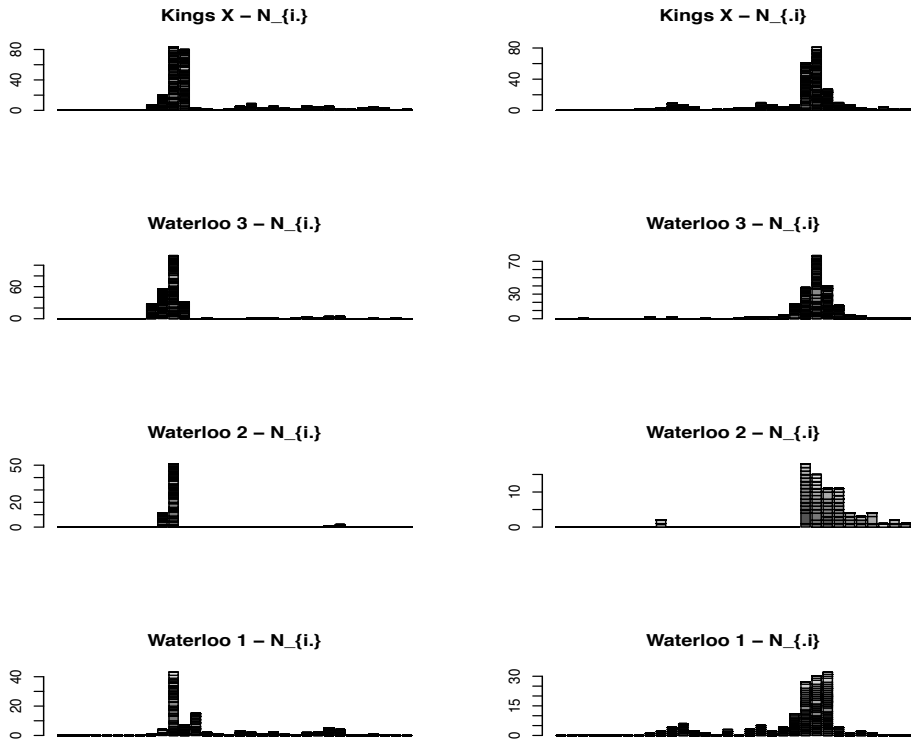

Figure 3: London bike sharing system: Barplots of outgoing ($N_{i\cdot}(\cdot)$) and incoming ($N_{\cdot i}(\cdot)$) processes from the 4 stations i in the smallest cluster: representation of volumes of connections to all other stations during day 2.

departures than arrivals around 8 a.m. and 'incoming' stations at the end of the day, with more arrivals than departures between 5p.m and 7p.m. Looking at the temporal profiles of stations close to the two other main railway stations in London (Victoria and Liverpool Street stations), this pattern is not observed and the stations are clustered differently. Thus, this small cluster is characterized by stations used by people living in the suburbs and working in the city center. This result highlights the specificity of our model that is able to find clusters from similar temporal profiles in sharp contrast with aggregated data approaches.

We then used a kernel estimator of the intensities per (directed) groups pairs $1 \leq q, l \leq Q$ (see Figure S.6 in the Supplementary Material). Cluster number 5 (which is the small cluster mentioned above) has high (directed) intensities of connections with cluster number 3 (shown in green plus sign $+$ in Figure S.5). This last cluster groups the stations belonging to the business city center of London. We observe a large intensity of connections from cluster 5 (King's Cross and Waterloo railway stations) to cluster 3 (City business center) in the morning and in the other direction (from cluster 3 to cluster 5) at the end of the day. Moreover, cluster 5 also appears to have (a smaller amount of) connections with clusters 1,4 and itself. Connections with cluster 1 (black circles $\circ$ in Figure S.5) and cluster 4 (blue cross $\times$ in Figure S.5) are similar (from cluster 5 in the morning and back to cluster 5 in the evening) to those with cluster 3 but at a smaller scale. Those results support and extend

the previous description of the role of cluster 5.

To conclude this section we mention that Randriamanamihaga et al. (2014) use a completely different approach, relying on Poisson mixture models on the same Origin/Destination flows. This approach does not take into account the network structure of the data (where e.g. two flows from the same station are related). As a consequence, clusters are obtained on pairs of stations from which interpretation is completely different and in a way less natural.

5.2 Primary school temporal network dataset

To understand contacts between children at school and to quantify the transmission opportunities of respiratory infections, data on face-to-face interactions in a French primary school were collected. The dataset is presented in detail in Stehlé et al. (2011) and available online (SocioPatterns, 2015). Children are aged from 6 to 12 years and the school is composed of five 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 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. Lunch is 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 a time, only two or three classes have breaks together, and lunch is served in two turns. The dataset contains $125,773$ face to face contacts among $n = 242$ individuals (232 children and 10 teachers) observed during two days.

We applied our procedure in the undirected setup with histograms based on a dyadic partition with maximum size $D_{\max} = 1024$. For $Q = 2, \ldots, 11$, Figure 4 shows the clustering of the n individuals into Q groups, where children from different classes are represented in different colors. When Q is small ($Q \leq 6$), our procedure gathers all pupils from one class and their corresponding teacher in the same cluster. For larger values of Q, 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 $4B$). Teachers never form a particular group apart, but they are in the cluster of their assigned class, suggesting that contacts among 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).

The 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 of the data. It has already been observed by other authors that this may happen for large datasets (see Guigourès et al., 2015, and the references therein). Thus, we choose to further analyse the data for $Q = 11$ groups.

We observe that the intensities representing the most activity are the intra-group intensities. As clusters mainly correspond to classes, this highlights that most contacts involve children of the same class and that the dataset is structured into communities. Moreover, peaks of interactions are observed during the two breaks around 10.30am and 3.30pm. At lunch time interactions between children vary from the first to the second day and are less

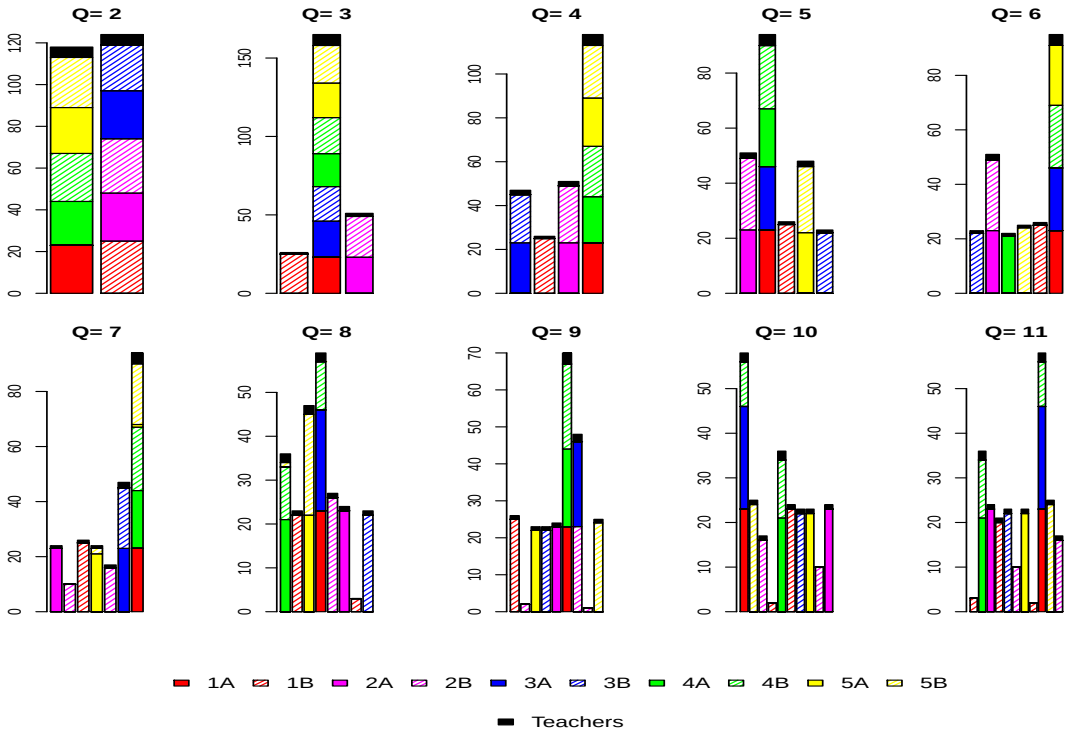

Figure 4: Primary school: clustering of the 232 individuals (represented by different colours) into $Q = 2, \ldots, 11$ groups. For each picture the vertical bars represent the Q clusters. Colours indicate the grades and the teachers, plain and hatching distinguish the two classes in the same grade.

important than during the breaks where they play together.

Concerning inter-group connections, 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, as our procedure splits some children of the same class into separate groups, the inter-group interactions associated with these clusters correspond in fact to intra-class interactions. Second, intensities between groups made of children of the same grade are significant, suggesting that children mostly interact with children of the same age. Third, the children of class $2B$ are partitioned into two clusters, and the intensity of one of these clusters drops to zero during lunch time, whereas children of the other group interact a lot during lunch. It seems that our procedure has recognized two subgroups in class $2B$: children having lunch at school and those going home for lunch. Fourth, class $1B$ is split into three groups with 20, 3 and 2 pupils, respectively. The estimated intensities suggest a particular behaviour of some of the children: there is no contact between the two children in the smallest cluster, but they have very strong interaction with the three pupils in the other cluster.

As a conclusion, we recover many results of Stehlé et al. (2011). In particular, we detect subgroups of pupils with a specific behavior at some period of the day (leaving school for lunch). We think that this is mainly due to the fact that the Poisson process stochastic block model takes into account the information provided by the timestamps of the events.

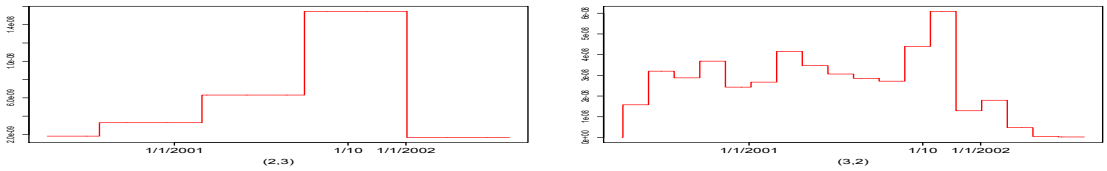

Figure 5: Enron: Asymmetric intra-group intensities.

Without this temporal information, it is hard to imagine to obtain similar results. In particular, aggregating data on a day scale to construct discrete time networks would not provide such a refined analysis.

5.3 Enron dataset

We analyse the email exchanges between 147 persons working at Enron, mostly in the senior management, covering the period of the affair that led to the bankruptcy of the company (Klimt & Yang, 2004). The dataset (CALO Project, 2015) contains 22,531 emails exchanged among these 147 persons between June 14, 2000 and June 13, 2002, for which the sender, the recipient and the time when the email was sent are known. As data are obviously directed, we use the directed version of our method to analyse them.

Here again, the integrated classification likelihood criterion does not provide a reasonably small enough number of clusters that could be used for interpretation. We choose to analyse the data with $Q = 4$ clusters. There is one group which is rather silent with very little activity and intensities close to zero. The other groups are characterized by substantial intra-group communication. It is instructive to compare the temporal profiles of the estimated intensities. First, we observe that communication is not symmetric. For example, people in cluster 3 regularly send emails to cluster 2 over the whole observation period, but the latter only respond rather late during the second half of 2001, see Figure 5. Second, the intra-group communication differs a lot from one group to the other, see Figure 6. Cluster 3 has a rather constant communication intensity over the whole period, while the intra-group intensity of cluster 1 is increasing with a peak at the end of 2002, which seems to be the consequence of the beginning of the investigations. In contrast, cluster 4 has an important intra-group activity until June 2001, then the intensity drops down and achieves another peak just before the inquiry, when it is known that number of individuals acted to hide sensitive information on the scandal. Thus, one may suspect people from cluster 4 to hold relevant information for the investigators.

Finally, we compare these results with those obtained using a classical stochastic block model. Indeed, 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 a classical stochastic block model with Poisson emission distribution and mean parameter $A^{(q,l)}(T)$ (see for instance Mariadassou et al., 2010). We compare the classifications obtained by the two procedures for $Q = 4$ clusters. The associated adjusted rand index is 0.74 indicating that part of the individuals are clustered in the same way and some are treated very differently in the two models. Indeed, both methods find a large common cluster with people that do not communicate a lot, while most hesitation is about persons in cluster 3 in the Poisson process stochastic

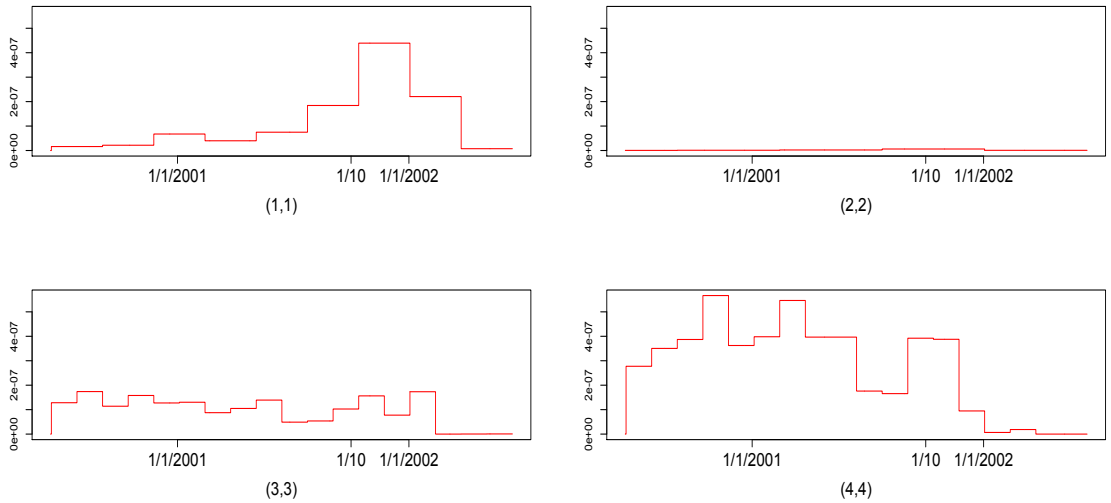

Figure 6: Enron: Inter-group intensities.

block model, which are mainly split into two groups in the classical stochastic block model and partly mixed with people from cluster 1. In fact, cluster 1 and 3 in the Poisson process stochastic block model have significant intra-group communication but very different temporal profiles as mentioned above. Intuitively, it may be difficult to distinguish these groups when we only look at the total count data as in classical stochastic block model. 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 stochastic block model.

Acknowledgment

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

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

Allman, E., Matias, C. & Rhodes, J. (2011). Parameters identifiability in a class of random graph mixture models. *J. Stat. Plan. Inference* **141**, 1719–1736.

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

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

Bickel, P. J., Chen, A. & Levina, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39**, 2280–2301.

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

Böhning, D. (1995). A review of reliable maximum likelihood algorithms for semiparametric mixture models. *J. Stat. Plan. Inference* **47**, 5 – 28.

Bordes, L., Chauveau, D. & Vandekerkhove, P. (2007). A stochastic EM algorithm for a semiparametric mixture model. *Comput. Stat. Data Anal.* **51**, 5429 – 5443.

Butts, C. T. (2008). A relational event framework for social action. *Sociol. Methodol.* **38**, 155–200.

CALO Project (2015). http://www.cs.cmu.edu/ enron/.

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

Cho, Y.-S., Galstyan, A., Brantingham, P. J. & Tita, G. (2014). Latent self-exciting point process model for spatial-temporal networks. *Discrete Continuous Dyn. Syst. Ser. B* **19**, 1335–1354.

Corneli, M., Latouche, P. & Rossi, F. (2016). Exact ICL maximization in a non-stationary temporal extension of the stochastic block model for dynamic networks. *Neurocomputing* **192**, 81 – 91.

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

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

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

DuBois, C., Butts, C. T. & Smyth, P. (2013). Stochastic blockmodeling of relational event dynamics. In *AISTATS*, vol. 31. JMLR Workshop and Conference Proceedings.

Grégoire, G. (1993). Least squares cross-validation for counting process intensities. *Scand. J. Statist.* **20**, pp. 343–360.

Guigourès, R., Boullé, M. & Rossi, F. (2015). Discovering patterns in time-varying graphs: a triclustering approach. *Adv. Data Anal. Classif.* , 1–28.

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

Hubert, L. & Arabie, P. (1985). Comparing partitions. *J. Classif.* **2**, 193–218.

Jordan, M., Ghahramani, Z., Jaakkola, T. & Saul, L. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37**, 183–233.

Klimt, B. & Yang, Y. (2004). The Enron Corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004*, J.-F. Boulicaut, F. Esposito, F. Giannotti & D. Pedreschi, eds., vol. 3201 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg.

Linderman, S. & Adams, R. (2014). Discovering latent network structure in point process data. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, E. P. Xing & T. Jebara, eds., vol. 32.

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

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

Matias, C. & Miele, V. (to appear). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* Available at ArXiv:1506.07464.

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

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

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11**, pp. 453–466.

Randriamanamihaga, A. N., Côme, E., Oukhellou, L. & Govaert, G. (2014). Clustering the Vélib' dynamic origin/destination flows using a family of Poisson mixture models. *Neurocomputing* **141**, 124 – 138.

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

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

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

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

SocioPatterns (2015). http://www.sociopatterns.org/.

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

Transport for London (2016). Cycle hire usage data 2012 - 2015. http://cycling.data.tfl.gov.uk/.

Vu, D. Q., Hunter, D., Smyth, P. & Asuncion, A. U. (2011). Continuous-time regression models for longitudinal networks. In *Adv Neural Inf Process Syst 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira & K. Weinberger, eds. Curran Associates, Inc., pp. 2492–2500.

Wasserman, S. (1980a). Analyzing social networks as stochastic processes. *JASA* **75**, 280–294.

Wasserman, S. (1980b). A stochastic model for directed graphs with transition rates determined by reciprocity. *Sociol. Methodol.* **11**, pp. 392–412.

Xu, K. & Hero, A. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE J. Sel. Topics Signal Process.* **8**, 552–562.

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

Supplementary material for: A semiparametric extension of the stochastic block model for longitudinal networks

All the references are from the main manuscript, except for those appearing as S-xx that are within this file.

S.1 Identifiability proofs

Proof of Proposition 1. For notational convenience, the proofs are presented in the undirected setup, where the set of intensities is $\alpha = \{\alpha^{(q,l)}\}_{1\leq q\leq l\leq Q}$. The directed case can be treated in the same way. 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 $q \leq l$, 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

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

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 $\{(A^{(q,q)}; A^{(q,q)}; A^{(q,q)}); q = 1, \ldots, Q\}$ and thus the measures $\{A^{(q,q)}\}_q$ are identifiable, up to a permutation in $\mathfrak{S}_Q$. Going back to (S.1) 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

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

This is equivalent to saying that we identify the measures $\{A^{(q,l)}; q \leq l\}$ up to a permutation in $\mathfrak{S}_Q$. Obviously, this also identifies the corresponding intensities $\{\alpha^{(q,l)}; q \leq l\}$ 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 $\{A^{(q,q)}; q = 1, \ldots, Q\}$, we recover from (S.1) the set of values $\{\pi_q^3; q = 1, \ldots, Q\}$ 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 $\{\sum_q \pi_q^n; n = 1, \ldots, Q\}$ determine the values of elementary symmetric polynomials $\{\sigma_n(\pi_1, \ldots, \pi_Q); n =$

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

S.2 Technical details on the estimation procedure

S.2.1 Variational E-step

Proof of Proposition 3. For the Kullback-Leibler divergence we compute

$$KL\left(\Pr_\tau(\cdot|\mathcal{O})||\Pr_\theta(\cdot|\mathcal{O})\right) = E_\tau\left(\log\frac{\Pr_\tau(\mathcal{Z}|\mathcal{O})}{\Pr_\theta(\mathcal{Z}|\mathcal{O})}\middle|\mathcal{Z}\right) = E_\tau\left(\log\frac{\Pr_\tau(\mathcal{Z}|\mathcal{O})\Pr_\theta(\mathcal{O})}{\mathcal{L}(\mathcal{O},\mathcal{Z}|\theta)}\middle|\mathcal{Z}\right)$$
$$= \sum_{i=1}^{n} E_\tau\left(\log\tau^{i,Z_i}\middle|\mathcal{O}\right) + \log\Pr_\theta(\mathcal{O}) - E_\tau\left(\log\mathcal{L}(\mathcal{O},\mathcal{Z}|\theta)|\mathcal{Z}\right).$$

According to (5), the complete-data log-likelihood $\log\mathcal{L}(\mathcal{O},\mathcal{Z}|\theta)$ is

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

where $Y^{(q,l)}$ and $Z_m^{(q,l)}$ have been introduced in Equations (1) and (3), respectively. Now, note that $E_\tau[Z^{i,q}|\mathcal{O}] = \Pr_\tau(Z^{i,q} = 1|\mathcal{O}) = \Pr_\tau(Z_i = q|\mathcal{O}) = \tau^{i,q}$. Moreover, by the factorization property (6), for every $i \neq j$ we have

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

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

$$\hat{\tau} = \operatorname*{Argmin}_{\tau\in\mathcal{T}} KL\left(\Pr_\tau(\cdot|\mathcal{O})||\Pr_\theta(\cdot|\mathcal{O})\right) = \operatorname*{Argmax}_{\tau\in\mathcal{T}} J(\theta,\tau),$$

where $J(\theta,\tau)$ is

$$-\sum_{q=1}^{Q}\sum_{l=1}^{Q}\bar{Y}^{(q,l)}A^{(q,l)}(T) + \sum_{q=1}^{Q}\sum_{l=1}^{Q}\sum_{m=1}^{M}\tau_m^{(q,l)}\log\left(\alpha^{(q,l)}(t_m)\right) + \sum_{i=1}^{n}\sum_{q=1}^{Q}\tau^{i,q}\log\frac{\pi_q}{\tau^{i,q}}. \quad \text{(S.2)}$$

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 are

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

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 (8), with $\exp(\gamma_i - 1)$ being the normalizing constant. □

S.2.2 Histogram-based 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) 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 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 $N^{(q,l)}$) 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 $k = 1,\ldots,D_\mathcal{E}$ is

$$N^{(q,l)}(E_k) = \int_{E_k} dN^{(q,l)}(s) = \sum_{m=1}^{M} \tau^{i_m,q}\tau^{j_m,l}\mathbb{1}_{E_k}(t_m). \tag{S.3}$$

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

$$\hat{\alpha}_\mathcal{E}^{(q,l)}(\cdot) = \frac{1}{\bar{Y}^{(q,l)}}\sum_{k=1}^{D_\mathcal{E}} \frac{N^{(q,l)}(E_k)}{|E_k|}\mathbb{1}_{E_k}(\cdot). \tag{S.4}$$

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 $Q_\tau(\pi, \{\alpha_k^{(q,l)}\}_{(q,l),1\leq k\leq D_\mathcal{E}})$ to be maximized with respect to 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 (S.4), that minimize the contrasts $\{\gamma_n^{(q,l)}\}_{q,l}$ 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 (S.4). 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 $\mathcal{E}^d_{\text{reg}}$ of $[0,T]$ with $D_{\mathcal{E}^d_{\text{reg}}} = d$ intervals each of length T/d for $d\in\{1,\dots,D_{\max}\}$, or the collection of dyadic partitions $\mathcal{E}^d_{\text{dy}}$ of $[0,T]$ with $D_{\mathcal{E}^d_{\text{dy}}} = 2^d$ intervals of length $T/2^d$ for $d\in\{0,\dots,d_{\max}\}$ (where $D_{\max}$ and $d_{\max}$ are to be chosen). Furthermore, the penalty function is given by

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

where $\mathcal{E}^{\max}$ denotes the finest partition in the collection $\mathcal{M}_n$, that is $D_{\mathcal{E}^{\max}} = D_{\max}$ 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

$$\forall t\in[0,T], \quad \hat\alpha_{\text{hist}}^{(q,l)}(t) = \hat\alpha_{\hat\mathcal{E}^{(q,l)}}^{(q,l)}(t) = \frac{1}{T\bar Y^{(q,l)}} \sum_{k=1}^{\hat D^{(q,l)}} \hat D^{(q,l)} N^{(q,l)}(E_k^{\hat\mathcal{E}}) \mathbb{1}_{E_k^{\hat\mathcal{E}}}(t). \tag{S.5}$$

S.2.3 Details on the Algorithm

A natural stopping criterion of the variational EM algorithm is based on function J defined in (S.2). Indeed, $J(\theta,\tau) = Q_\tau(\theta) + \mathcal{H}(\tau)$, where $\mathcal{H}(\tau)$ denotes the entropy of the distribution τ defined as $\mathcal{H}(\tau) = -\sum_{i=1}^n\sum_{q=1}^Q \tau^{i,q}\log\tau^{i,q}$. As our estimation procedure aims at maximizing J, the algorithm may be stopped at iteration $s+1$ if the increase of J is less than a given threshold $\varepsilon>0$, that is when

$$\left|\frac{J(\theta^{[s+1]},\tau^{[s+1]}) - J(\theta^{[s]},\tau^{[s]})}{J(\theta^{[s]},\tau^{[s]})}\right| < \varepsilon.$$

We use several initializations of the algorithm, relying on different aggregated datasets (on the whole time interval or on sub-intervals) and applying a k-means algorithm on the rows of the adjacency matrix of these aggregated datasets.

S.3 Additional figures

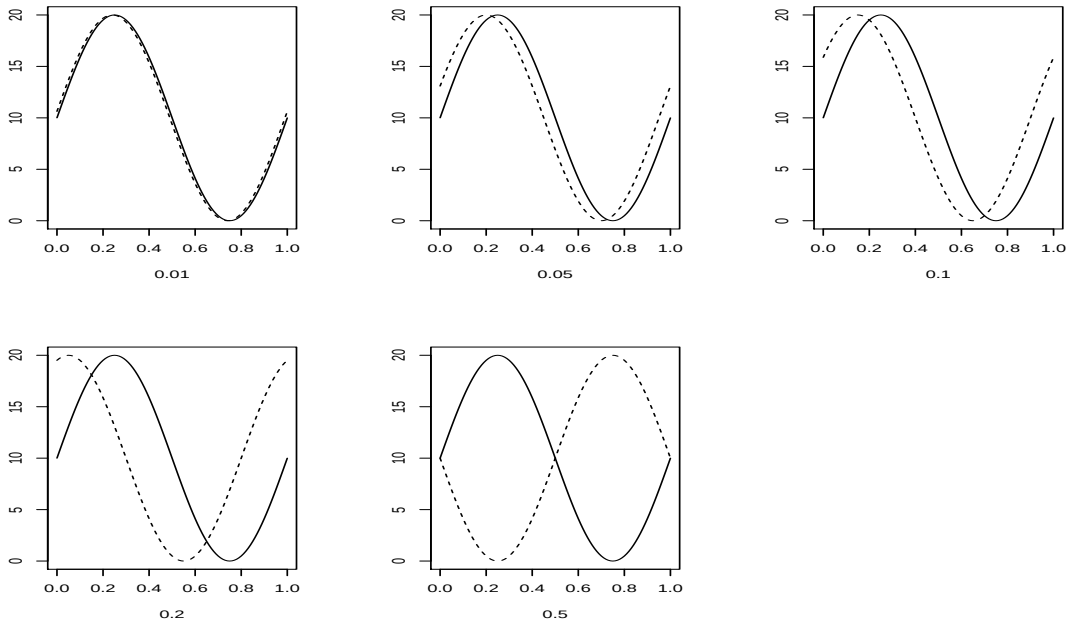

Figure S.1: Intensities in Scenario 1. Each picture represents the intra-group intensity $\alpha^{(q,q)}$ (bold line) and the inter-group intensity $\alpha^{(q,l)}$ for $q \neq l$ (dotted line) with different shifting parameter $\varphi \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$.

References

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

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.

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

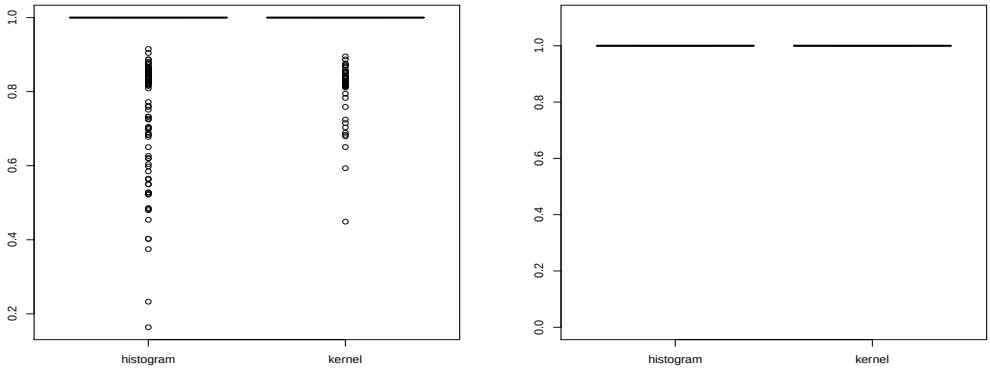

Figure S.2: Boxplots of the adjusted rand index in Scenario 2 for the histogram (left) and the kernel (right) estimators. Left panel $n = 20$, right panel $n = 50$.

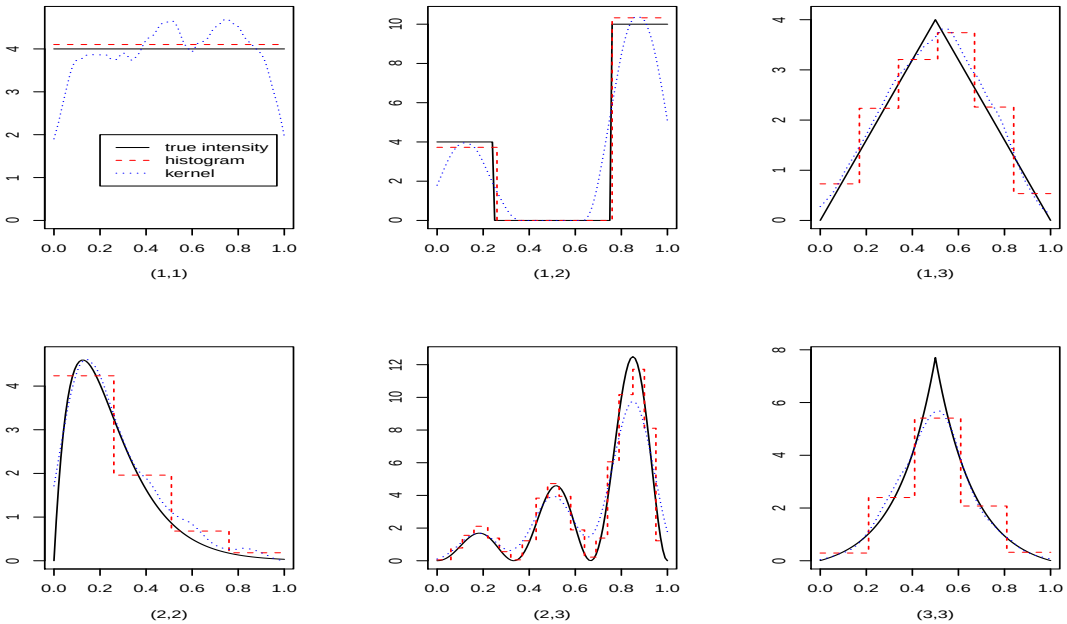

Figure S.3: Intensities in Scenario 2 with $n = 50$: True intensities (black continuous), histogram (red dashed) and kernel estimates (blue dotted) for each pair of groups (q, l) with $1 \leq q \leq l \leq 3$.

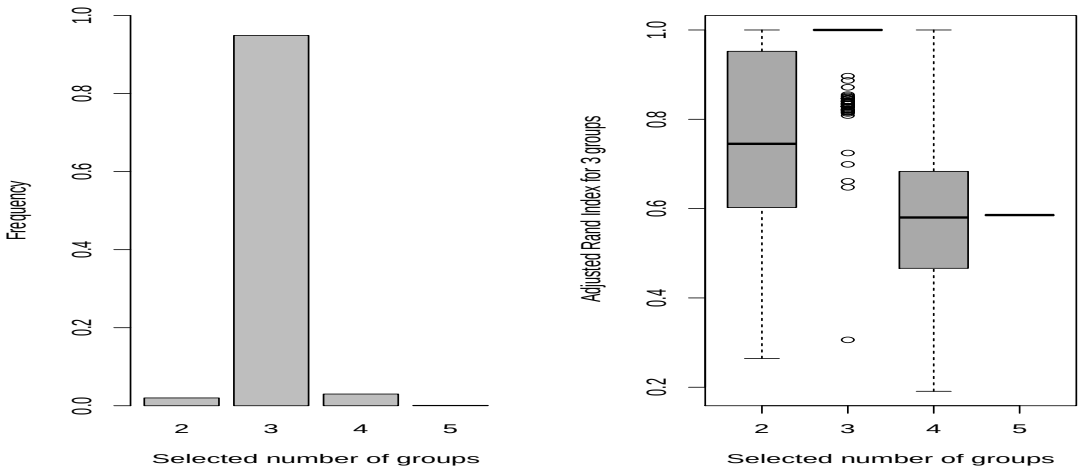

Figure S.4: Selection of the number of latent groups via the integrated classification likelihood criterion in Scenario 2 with $n = 20$. Left panel: frequencies of selected number of groups. Right panel: adjusted rand index of the classification into three groups as a function of the number of selected groups.

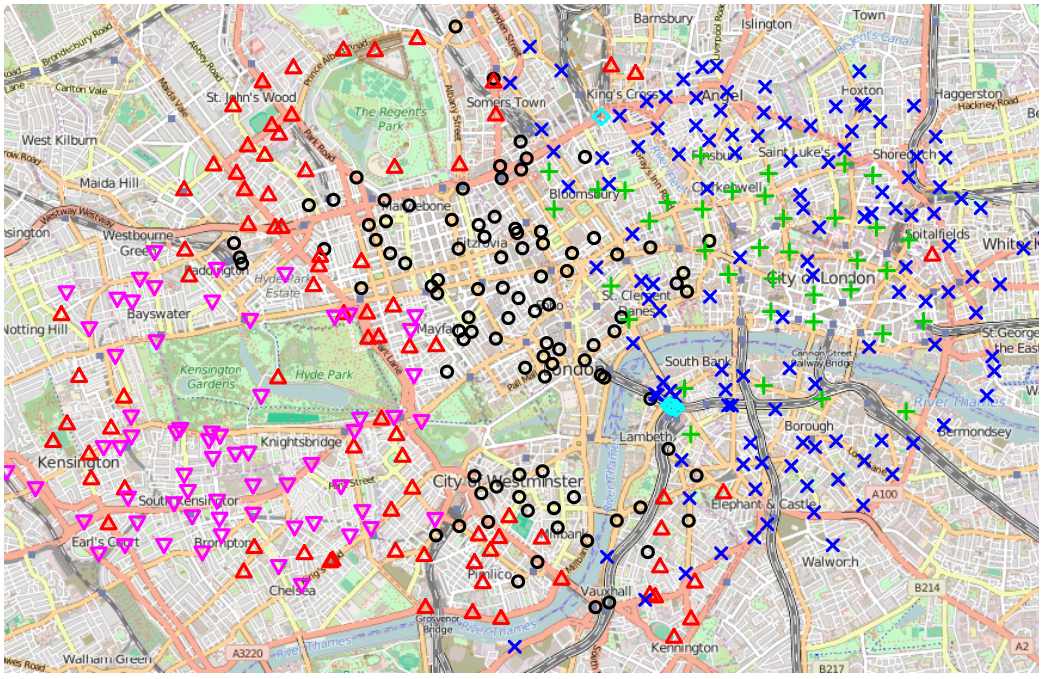

Figure S.5: London bike sharing system: Geographic positions of the stations and clustering into six clusters (represented by different colors and symbols) for day 2.

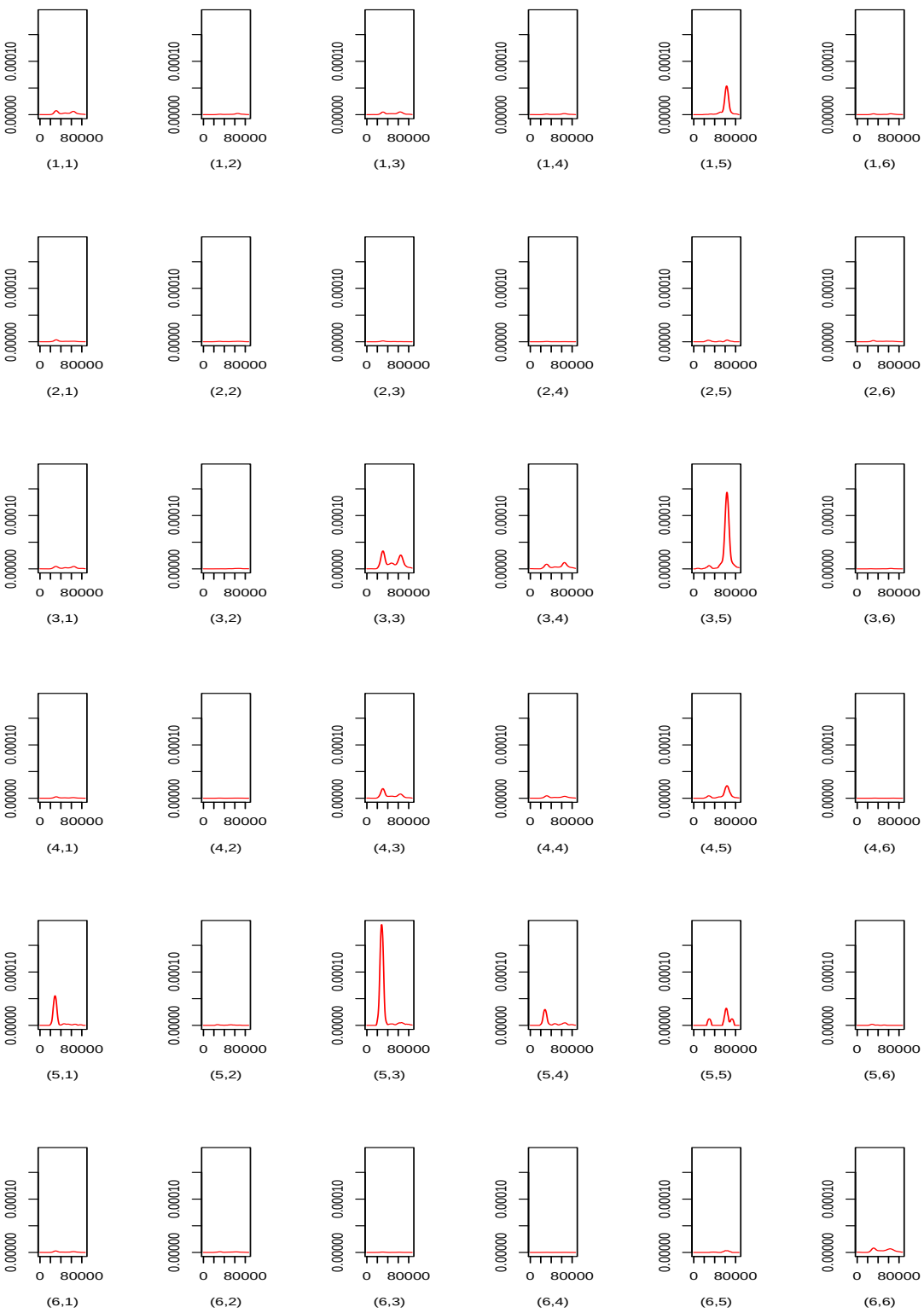

Figure S.6: London bike sharing system: Kernel intensity estimates for the six clusters (day 2), plotted on the same y-scale.

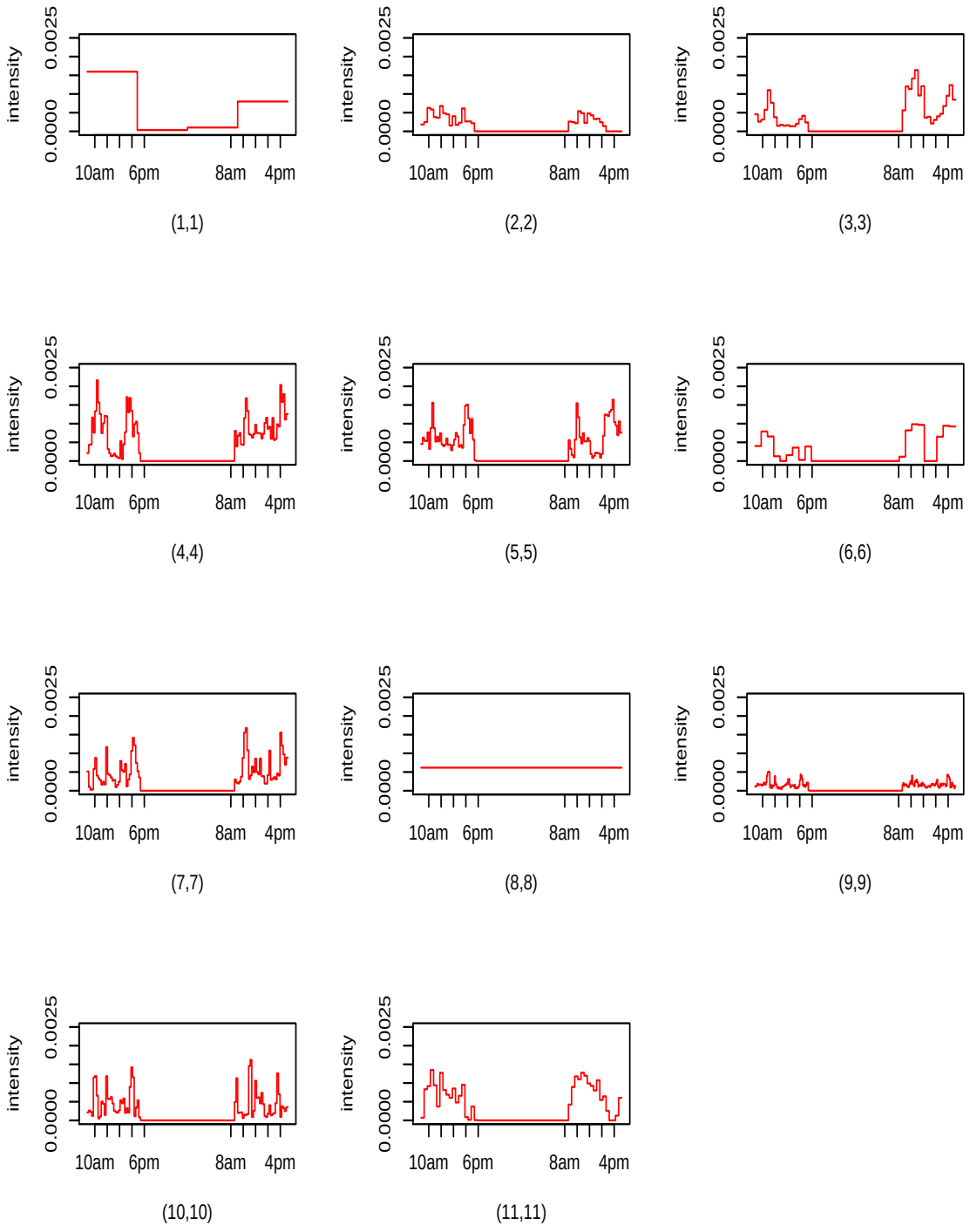

Figure S.7: Primary school: Estimated intra-group intensities for $Q = 11$ groups (plotted on the same y-scale).

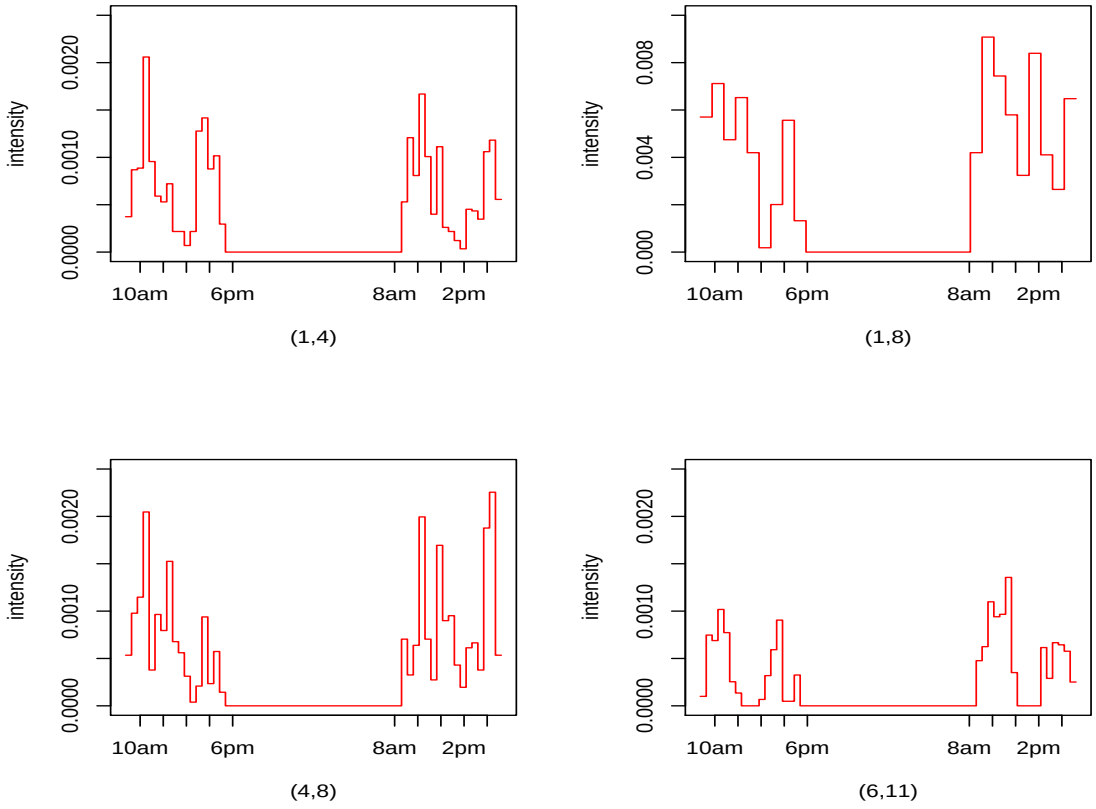

Figure S.8: Primary school: Inter-group interactions that correspond to intra-class interactions with $Q = 11$ groups.

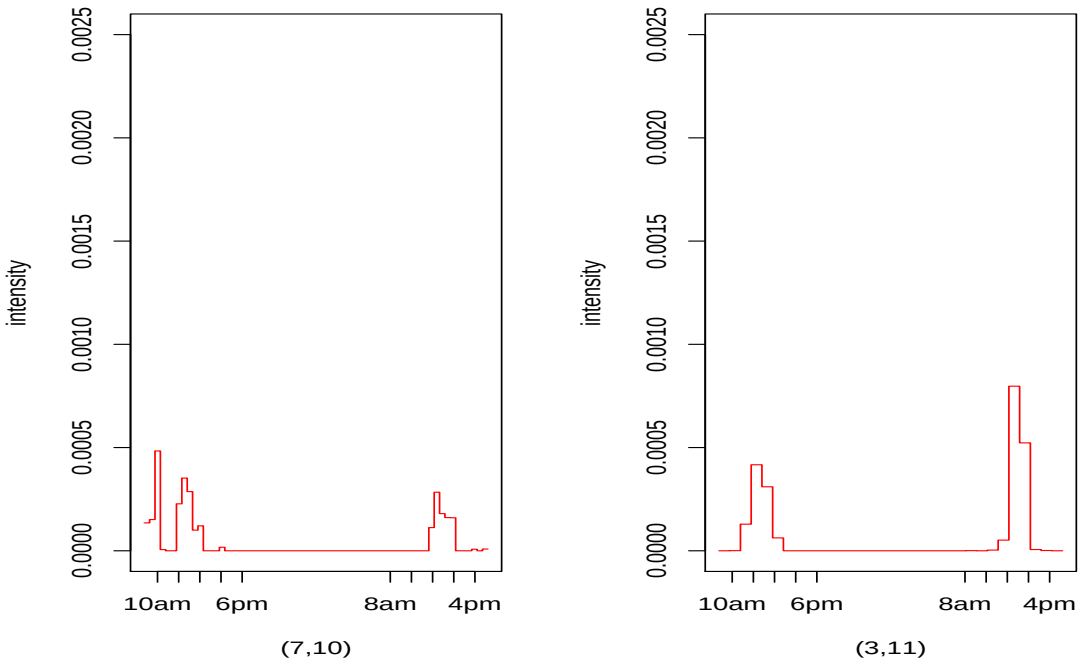

Figure S.9: Primary school: Estimated inter-group intensities between classes $5A$ (group 7) and $5B$ (group 10) and between classes $2A$ (group 3) and $2B$ (group 11) for $Q = 11$ groups.