Journal of the Royal Statistical Society Series A: Statistics in Society, 2023, **00**, 1–22
https://doi.org/10.1093/jrsssa/qnad028

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

Original Article

Networks with correlated edge processes

Maria Süveges[1] **and Sofia Charlotta Olhede**[2]

[1]University of Geneva, 24 rue du Général-Dufour, 1211 Genève, Switzerland
[2]École polytechnique fédérale de Lausanne, Station 8, CH-1015 Lausanne, Switzerland

Address for correspondence: Sofia Charlotta Olhede, École polytechnique fédérale de Lausanne, Station 8, CH-1015 Lausanne, Switzerland. Email: sofia.olhede@epfl.ch

Abstract

This article proposes methods to model non-stationary temporal graph processes motivated by a hospital interaction data set. This corresponds to modelling the observation of edge variables indicating interactions between pairs of nodes exhibiting dependence and evolution in time over interactions. This article thus blends (integer) time series models with flexible static network models to produce models of temporal graph data, and statistical fitting procedures for time-varying interaction data. We illustrate the power of our proposed fitting method by analysing a hospital contact network, and this shows the challenge in modelling and inferring correlation between a large number of variables.

Keywords: correlated Bernoulli time series, exchangeable networks, link communities, time-varying network

1 Introduction

This paper introduces time series models for observations of dynamic graphs over time, and methods to estimate these models. This set of developments is motivated by the increasing prevalence of temporal observations of interactions between entities in many application areas (Ahmed & Xing, 2009; Liu & Duyn, 2013; Ribeiro et al., 2013). We call such observations *dynamic graphs*, and the observations correspond to samples from a temporal graph process (Crane, 2016), rather like a time series can be viewed as samples of a continuous-time stochastic process. The aim of this paper is to introduce a natural generalised linear modelling framework for discretely regularly sampled temporal graph processes that can flexibly capture data features such as cyclostationary and dependence of edges in time.

The rising ubiquity of dynamic graphs has been matched by technical innovation for their analysis, see, for example, recent contributions in Matias et al. (2018), Ludkin et al. (2018), Pensky (2019), Jiang et al. (2020), Pamfil et al. (2020), Hoff (2015), and Krivitsky and Handcock (2014). We also note existing work that is less recent, illustrating the importance of temporal graphs, for example, Snijders (2001), Hanneke and Xing (2006), and Guo et al. (2007). In the aforementioned articles (like in our model), the conditional distribution of the graph may be specified between time steps using exponential random graph models and by seeking to directly capture evolving topology of a graph. Simultaneously, the realisation that networks should be described directly in terms of observed interactions or edges rather than in terms of describing the interactions between nodes, in a nodal view, has been gaining considerable traction (Crane & Dempsey, 2018). These theoretical developments are paralleled by the recognition that nodal clustering may not be sufficient to model a graph due to overlapping node communities, and this problem may be resolved by assuming the links themselves to form communities on their own. Papers dealing with the detection of link communities and characterising their behaviour are, for example, Ahn et al. (2010), Evans (2010), Kim and Jeong (2011), Nguyen et al. (2011), and Meng et al. (2016).

Key to understanding dynamic graphs is proposing models for their dependence *and* evolution. The basic building blocks must consider the natural invariances of entities and temporal processes,

Received: December 23, 2021. **Revised:** January 31, 2023. **Accepted:** February 3, 2023

permutation invariance of measure, and shift-invariance of measure for stochastic processes. In addition, for non-Euclidean observations such as graphs it no longer makes sense to put all dependence in a zero-mean perturbation, as is done for most time series problems. First, we still want to encode additional temporal structure. We do not want to pose models whose structure is constant over time, and the perturbations are purely random. Second, the temporal structure should be parameterisable, and estimable from one process realisation. Our choice of model will be motivated by a real data example. For this reason, we wish to consider models satisfying some permutation invariance at fixed time stamps, but that are not stationary in time. This will be necessary especially as the processes we study could at most be assumed to be cyclostationary.

To explain about our modelling framework, and to be sufficiently concrete, let us study an example of a dynamic graph, depicted in Figure 1 as a set of six link communities. Link communities assume the same set of parameters for every edge in the same link community (set of edges). This is in contrast to the group membership of a node, as per the stochastic blockmodel (SBM). An alternative to assuming data is generated by the SBM is the mixed membership of Airoldi et al. (2005). Link communities are not equivalent to the mixed membership model. This follows as the mixed membership model specifies a set of behaviours per edge, since for each edge the Bernoulli success probability is drawn as a mixture over a fixed blockmodel.

Figure 1 shows the interactions over time of various groups of personnel and patients in a hospital ward in Lyon, France (the data set, the model and the fit resulting in the clustering shown in the Figure is described in Section 4.3). As in many other social networks, people in a hospital have distinct and varying contact patterns over time, dictated by a common rhythm of work meetings, patient visits, medical examinations (temporally scheduled 'rounds'), care for patients, meals (also periodic) and so on. The different colours indicate communities of similarly behaving link probabilities over time. The panels show the instantaneous contact probabilities of the communities at different times during the day as intensity of colour, so that, for example, deep purple corresponds to the maximal contact probability of the 'purple' community, and white, to a near-zero contact probability. The upper row shows the morning hours from 6a.m. to 8a.m., the lower row, the afternoon from 3p.m. to 5p.m. We see that different link communities have very different contact probability at different times of the day. Whereas, for example, the 'green' and 'orange' edges are mostly switched on in the morning, the 'purple' edges activate preferably in the afternoon.

These empirical facts must be linked with our choice of graph process modelling framework. Our observations are that: first, it is obvious that a model that intends to gain a detailed insight into the dynamics of such a network must be able to account for (periodic/cyclical) time-varying contact patterns, ubiquitous in human contacts. Second, links in human society can show interesting clustering according to their time variation patterns, which can be quite different from a node clustering scheme. Third, it is also evident that temporal evolution and community structure of this network cannot be written in a separable form $\rho(t)f(\xi_k(t), \xi_j(t))$, where $f(\xi_k(t), \xi_j(t))$ is a fixed constant baseline probability of interaction between nodes k and j depending on a latent process ξ_k, and $\rho(t)$ as a term driving the common temporal variation of these probabilities (thus the type of interactions change over time, not just the density of them). Fourth, it also seems plausible that human interactions, especially when observed with high temporal resolution, are in general temporally correlated as interactions cannot come and go willy-nilly from one moment to the other.

It is important to model these patterns, as an alternative to a classical SBM independently generated at each time-point. There are situations where the temporal dynamics of the links is an important factor in the scientific question. An example can be the modelling of the spread of an infectious disease in an evolving community. In this case, a detailed model of the temporal patterns of the contacts may inform much better the public health policymakers about the intervention with optimal cost-efficiency ratio than alternatives. We could have calculated node centrality measures and average contact probabilities, or fitted an SBM putting emphasis on similarities between nodes, rather than link communities that are discriminated based on their different temporal dynamics. Stationary processes cannot reproduce all manner of cyclostationary processes common in observations of human activities. Other similar examples with clear cyclostationary patterns will be found as energy networks or mobile phone networks. The application of detailed dynamic link community models can give a deeper insight as to risks of system breakdowns or overloads.

Currently, there are many proposed methods to perform inference on dynamic networks, say, for example, Matias and Miele (2017), Matias et al. (2018), Ludkin et al. (2018), Pensky

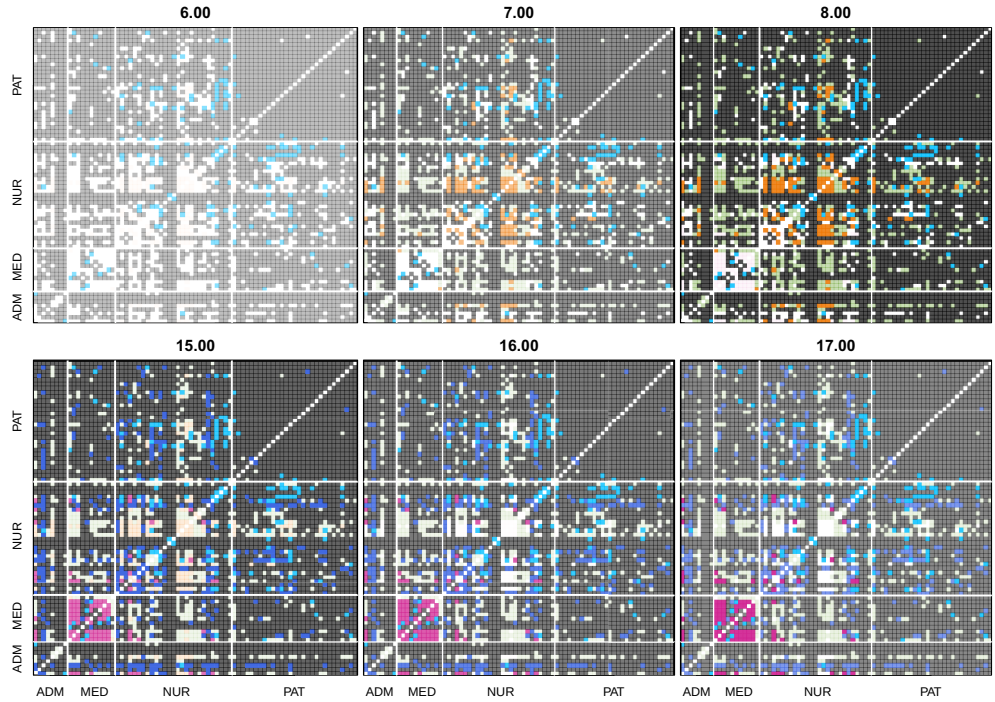

Figure 1. Snapshots of the varying daily activity levels of different link community groups in a hospital ward (see Section 4.3). The four groups of nodes according to status in the hospital is ADM (administrator), MED (medical staff), NUR (nursing staff), and PAT (patients). The hours of the day are shown in the main title of the panels. The colours indicate the membership of the links, while the intensity of the colours is proportional to the probability of the link being 'switched-on' at that particular time in the day. The maximal intensity of the colour corresponds to the maximal probability of the link to be 'on', which is different for all link communities (they are 0.063, 0.0004, 0.020, 0.011, 0.008, and 0.058 for the orange, black, green, dark blue, light blue, and purple clusters, respectively). White indicates 0 everywhere.

(2019), Jiang et al. (2020), and Olivella et al. (2021), each method coming with either explicit or implicit modelling assumptions. Most of these approaches use the SBM imposing clustering structure on the nodes (SBM; Anderson et al., 1992; Faust & Wasserman, 1992; Holland et al., 1983; Snijders & Nowicki, 1997) to model the underlying dynamic structure of the network, with clear choices on how parameters change or evolve across time. In the framework of the SBM, dynamics may arise from a latent process describing the evolution of the node memberships in the clusters over time, such as, for example, a set of independent discrete-state Markov processes for each node (e.g., Pensky, 2019) or a hierarchical model allowing for mixed memberships of the nodes and specifying a latent process on the membership distributions (e.g., Olivella et al., 2021). In most of these models, edges are generated independently, perhaps conditionally on the block memberships of the endpoints of the edges. Using latent variables to specify block membership will introduce marginal correlation in the adjacency matrix, but this correlation is not equivalent to modelling explicit correlation directly on observed edges. In this paper, we will instead argue that in the temporal setting, direct correlation in edges over time is natural, and this is why we prefer to use such relative to inducing correlation via latent variables. To this end, Figure 10 will later motivate our desire to model explicit correlation across edges. Finally, we note that many dynamic network observations, such as our example, cannot be taken as a series of temporally independent snapshots, especially in the context of human activities.

Exceptions to the description applied in those SBM-based papers cited above are Jiang et al. (2020) and Ludkin et al. (2018), where the former constructs correlated graphs by adding correlated noise that may erase or construct edges, that is, introduces correlation at the observed process, while the latter models community membership by a switching process, and directly imposes correlation on the edge variables. Thus over a given time interval, a correlation between edges is produced.

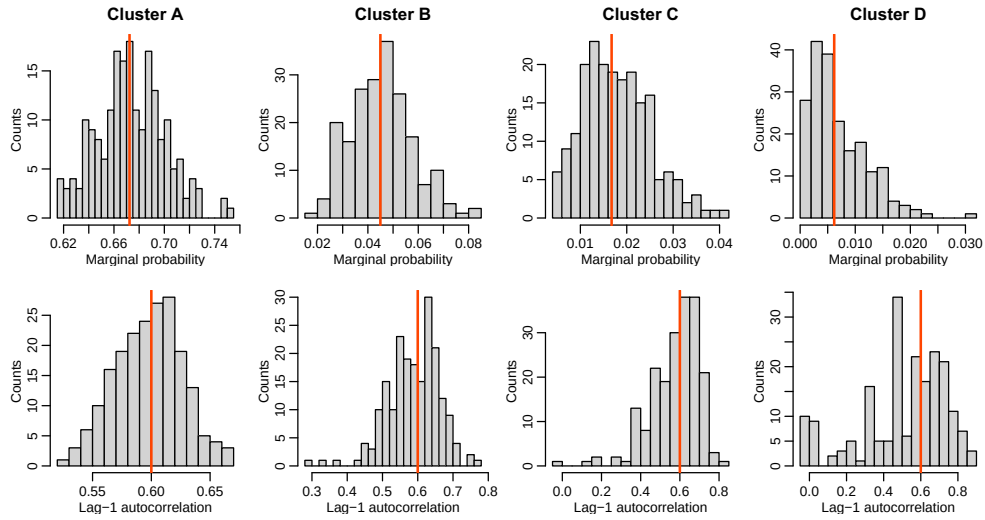

Figure 2. Exploring the sampling characteristics of fitted link communities in simulated data. The panels display the distribution of the estimates of the marginal probability (upper row) and the lag-1 autocorrelation (lower row) for the four simulated clusters. The vertical red line indicates the true value of the parameter (known as the data is simulated).

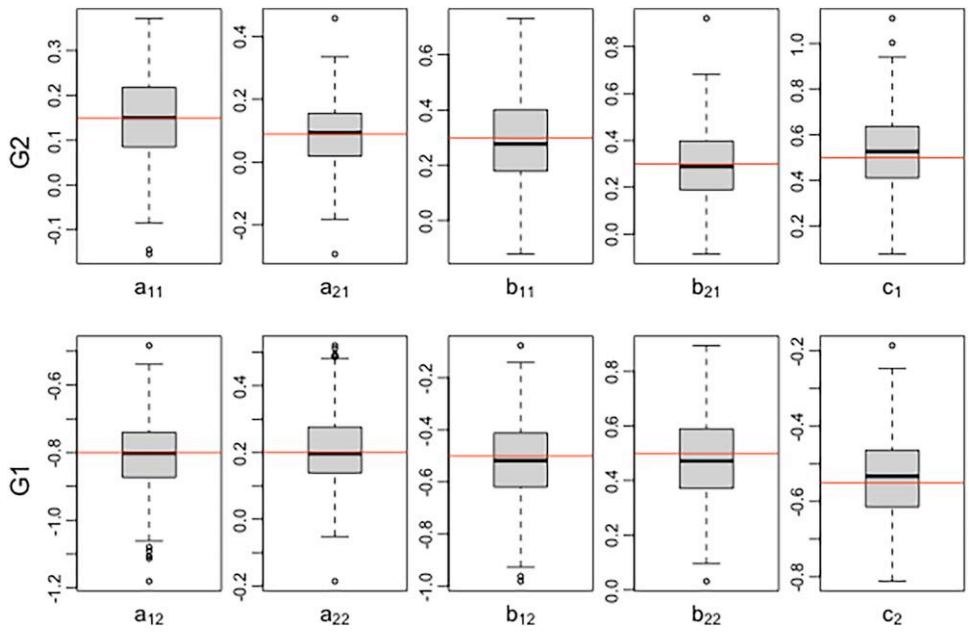

Figure 3. Exploring the sampling characteristics of fitted link communities in simulated data. The panels show estimates of model parameters of the simulated BALARM(2) model of Section 5 with two link communities G1 (top row) and G2 (bottom row). The red horizontal lines indicate the true parameter values (known as the data is simulated).

In this paper, we shall model correlation explicitly in the observed edges across time, based on using popular Bernoulli time series models. For the formulation, we reach back to generalised linear models (GLMs), using their well-known inferential characteristics to model the Bernoulli observations of simple graphs. This framework lends itself easily to extensions to Poisson time series

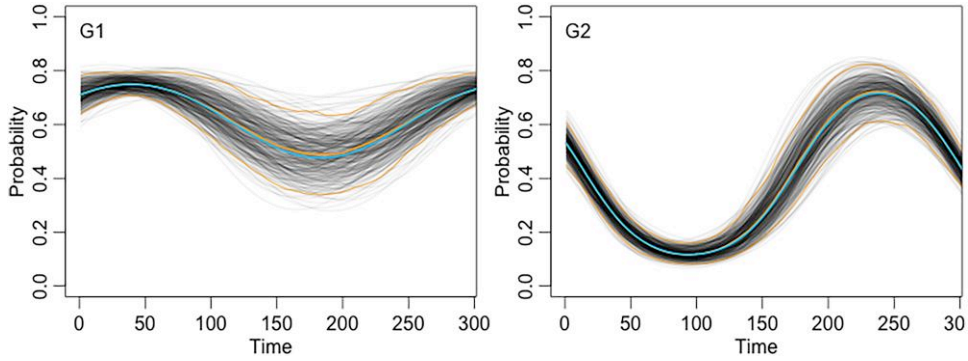

Figure 4. Time-varying contact probabilities conditioned on an immediate past $X_{t-1} = 0$, $X_{t-2} = 0$ for all t, in the simulated BALARM(2) model of Section 4.2. The curves have been computed using the estimated model parameters for each of the repetitions of the process. The blue line is the true contact probability, the thick orange line is the pointwise median estimate, the thin orange lines represent the pointwise 0.95 confidence band.

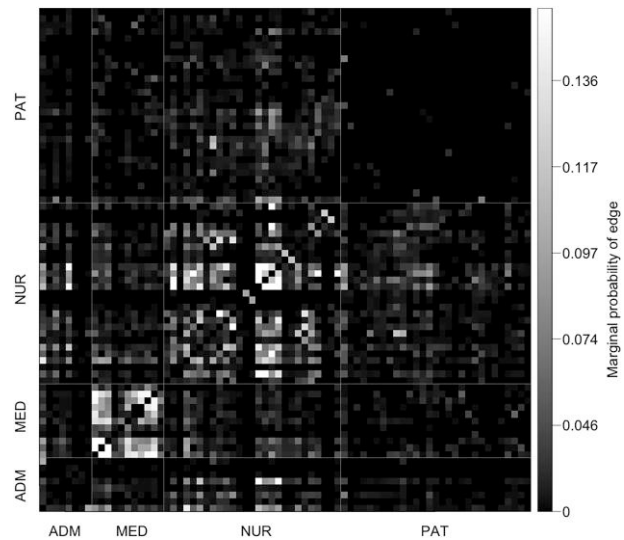

Figure 5. Link probabilities, averaged over time, between nodes of the hospital network. Shades of grey indicate the value of the probability, with black indicating 0 and white, 0.15. ADM, administrative staff; MED, medical staff; NUR, nursing staff; PAT, patients.

for counts of interactions, see, e.g., Hoff and Ward (2004), Minhas et al. (2016), Donnet and Robin (2019), and Schein et al. (2014). In terms of the latent edge variables defining the edge cluster memberships, we shall assume them fixed across time, drawing them at the temporal starting point of the process (assigning a community to each edge). Conditional on link community membership, for each edge, we then use the ALARM (A Logistic Autoregressive Model) generating mechanism (Agaskar & Lu, 2013), mainly because this allows the generation of positively and negatively correlated processes within the same model, as explained in Section 3.1. We then put the ALARM specification in the block model framework, introducing the block-ALARM specification (BALARM model), and give its likelihood, as described in Section 3.2. We use the EM algorithm to estimate the BALARM model. Whilst a simulation study lets us study the performance under correct model specification, see Sections 4.1 and 4.2, we also study the performance of this dynamic graph model when analysing temporal social interaction data from the geriatric short-stay ward of a university hospital in Lyon, France (Vanhems et al., 2013), mentioned already above. Using the BALARM model, we uncover groups of interactions between patients and staff,

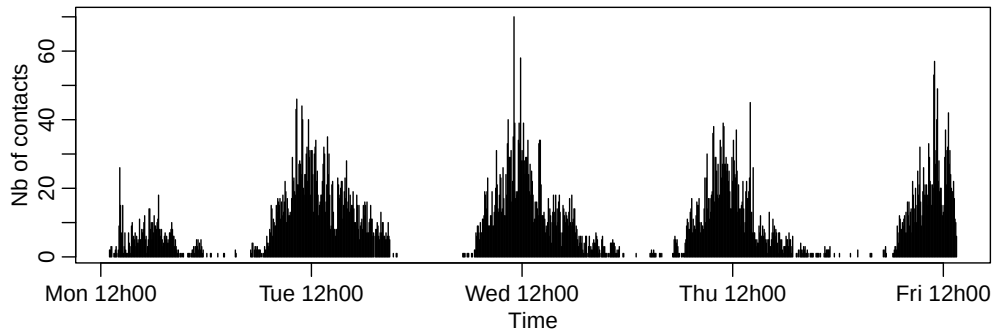

Figure 6. The total number of contacts $\sum_{k \neq j} A_{kj}(t_l)$ in each 5-min period within the observation span as a function of time.

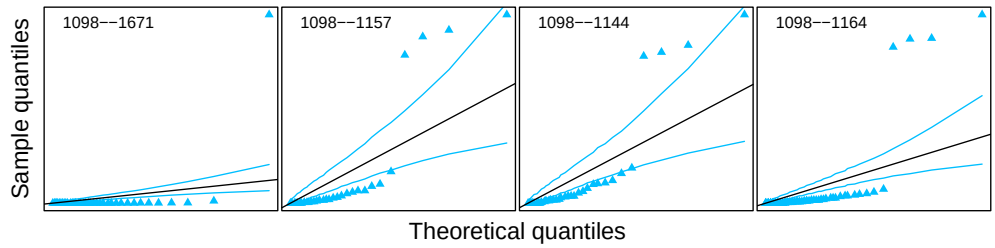

Figure 7. Geometric quantile–quantile plots for a few edges with link probability higher than 0.2. The black line has slope 1 and intersection 0, indicating the alignment of the expected positions of a sample from a geometric distribution.

with a clear daily temporal rhythm. This allows us to describe the graph process of temporally correlated edges in a compressed and simple, yet realistic manner.

2 Conditionally independent edge variable models

Consider a network with a fixed set of N nodes, without loss of generality, labelled by $\{k : 1, \ldots, N\}$. Let $A_{kj} \in \{0, 1\}$ denote the absence or presence of an edge (link) between nodes k and j, the value 0 indicating the 'switched-off' state of the edge, and 1, its 'switched-on' state.

The matrix $\{A_{kj}\}_{k,j \in \{1,\ldots,N\}}$ is called the adjacency matrix. We model A_{kj} as a Bernoulli variable with expectation p_{kj}:

$$A_{kj} \sim \text{Bernoulli}(p_{kj}), \quad 1 \leq i < j \leq N,$$

where $p_{kj} \in [0, 1]$ is the probability of an edge (kj). As the indexing k is arbitrary in order, without constraining p_{kj} this would be a high-dimensional model. To improve statistical efficiency it is standard practice to model A_{kj} as conditionally independent given some latent variables determining p_{kj}. There is more than one way to specify the number of latent variables, and the dimensionality of the latent variable model.

A common choice of specifying latent structure corresponds to the SBM. This assumes that each node k belongs to one of M possible clusters, and the probability p_{kj} of an edge being 'switched on' is determined by the $M \times M$ matrix θ_{ab} called the blockmatrix: if node k belongs to cluster a and node j to cluster b, then the probability of a link forming between them is equal to θ_{ab}. Introducing the random variable $Z_k \in \{1, \ldots, K\}$ to indicate the cluster membership of node k, we can then formulate the SBM as $A_{kj} \mid Z_k = a, Z_j = b \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta_{ab})$, and $Z_k \overset{\text{iid}}{\sim} \text{Multinom}(\pi_1, \ldots, \pi_M)$, where π_a is the probability that a node belongs to cluster a. With Z random, drawn from a distribution of labels, the marginal variability of an edge is greater than the inhomogeneous Bernoulli model that is specified conditionally. For a dynamical case where we observe the network over time, we need to also specify the temporal structure. There are various ways to do this. The possibility most often

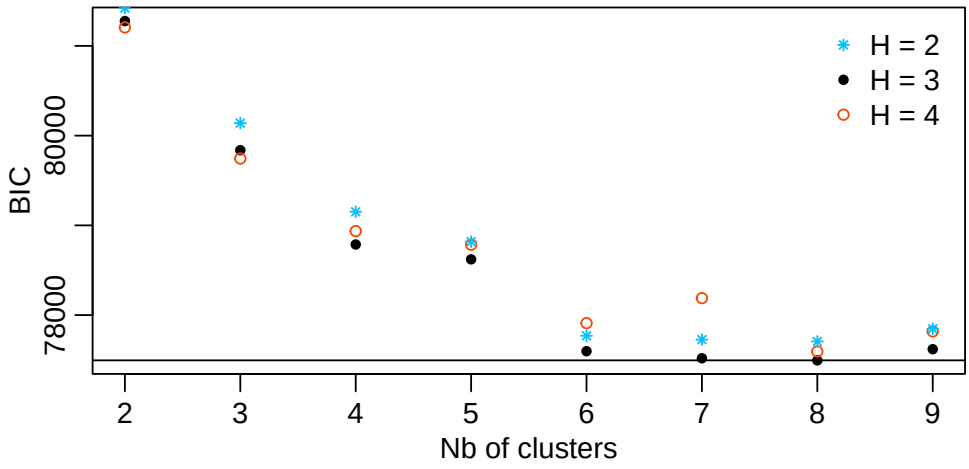

Figure 8. Bayes Information Criterion for the BALARM(1) models fitted to the hospital data. The different harmonic orders H are indicated with different colours and plotting symbols.

discussed in the literature is supposing the underlying blockmodel θ_{ab} stable over time, and assume that the nodes change cluster memberships over time according to some stochastic process, that is, suppose a specific time series structure for the indicators $Z_k(t)$, for example, a Markov process (e.g., Ludkin et al., 2018; Matias & Miele, 2017). A SBM varying in time is less frequently discussed, since this may make the model non-identifiable (Matias & Miele, 2017) if at the same time the nodes are allowed to change cluster. However, Jiang et al. (2020) proposes an autoregressive network model with changepoints over time in the blockmodel, and Pensky (2019) discusses the theoretical properties of models with smoothly varying connectivity probabilities. Finally, to obtain interesting dynamics, one can also relax the conditional independence in the generation of the Bernoulli variables A_{kj} over time, which is especially realistic if in a modern data acquisition setting we sample rather frequently, and expect edges not to flip very frequently. The simplest way of doing this is by introducing an autoregressive structure in the edge formation, making the value of $A_{kj}(t)$ directly depend on its previous measured value. This is done by imposing a first-order discrete autoregressive dependence in Jiang et al. (2020), and by imposing a continuous-time Markovian process CAR(1) in Ludkin et al. (2018).

The above-mentioned models estimate the network structure assuming nodal clusters and a unique membership of each node at any time. However, in real data, nodes may belong to more than one cluster (Palla et al., 2005, 2007). The first generalisation of the SBM is to allow the block structure to be relaxed, and this is called a mixed membership model (Airoldi et al., 2005). The mixed membership still references a latent set of nodal blocks, and so constrains the possible edge patterns we can observe. More generally, in order for edges to behave like groups of other edges, without referencing a latent nodal block structure, we shall use a link community model (Ahn et al., 2010).

Link communities, that is, when instead of nodes, edges are assumed to belong to one of a set of possible clusters in a data set, were originally proposed as a solution to this problem (Ahn et al., 2010), since in this way every node can maintain links belonging to different communities. Dynamical link communities are discussed in, for example, Meng et al. (2016) and Nguyen et al. (2011) using one-by-one updates of an initial link community state of the network, but no statistical inference is drawn about the network and its parameters. We shall use a model which groups generating mechanisms over edges, as is detailed in the next section.

3 Likelihood analysis of correlated edge models

3.1 The logistic autoregressive model

A class of basic models to deal with a regression with a binary response variable is the GLMs. Its definition consists of the specification of the response distribution (the Bernoulli distribution), the

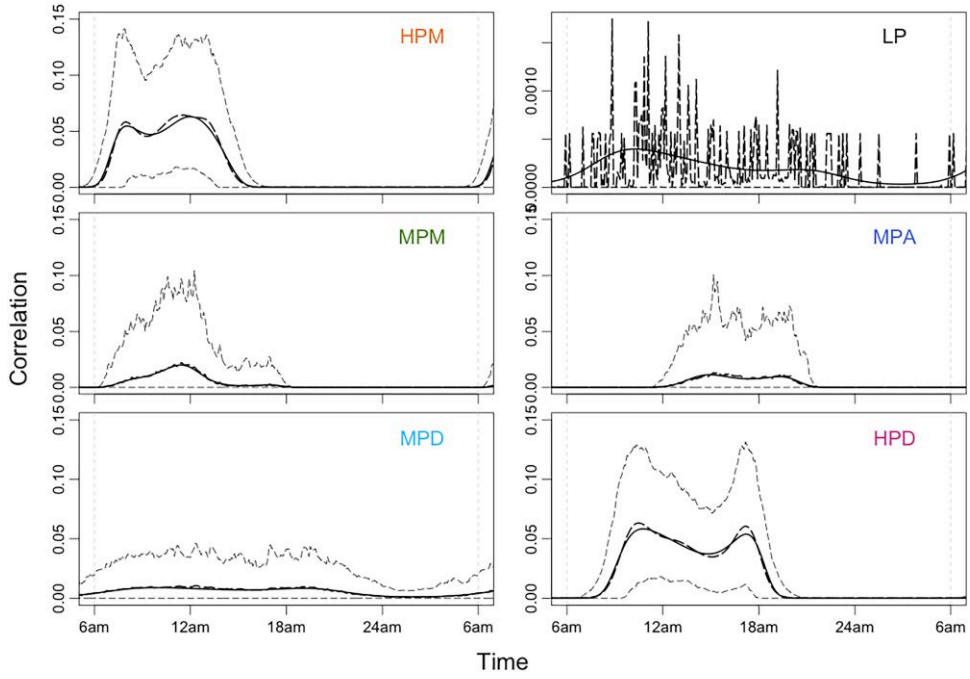

Figure 9. 95% pointwise bootstrap confidence intervals for the estimated daily variations of the link probabilities from the model with six components and three harmonic terms, for each identified cluster labelled according to Table 1, and colour-coded according to Figure 11. The thick solid line is the estimate on the real data. The heavy dashed line is the median of 500 bootstrap repetitions, the thin dashed lines represent the pointwise 0.025 and 0.975 quantiles. Note that while the five panels for the clusters HPM, MPM, MPD, MPA, and HPD have common *y*-axis limits [0, 0.15], the upper right panel showing the LP cluster has different limits, [0, 0.0025].

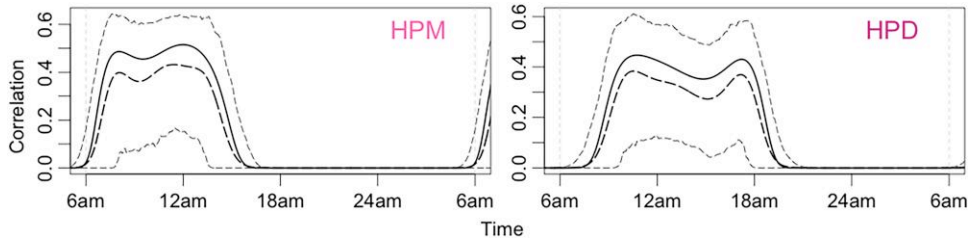

Figure 10. 95% pointwise bootstrap confidence intervals for the estimated daily variations of the lag-1 autocorrelation from the model with six components and three harmonic terms for the clusters with the highest link probabilities, for each identified cluster labelled according to Table 1, and colour-coded according to Figure 11. The line types correspond to the same quantities as presented in the caption of Figure 9.

linear predictor (comprising the influence of the covariates on the response), and the link function (the functional relationship between the linear predictor and the expected value of the response). We will use one such model to model the observed time series of an edge of a network, and combine these time series models into a blockmodel-type cluster structure based on the different dynamics of the time series.

Let $\boldsymbol{X}(t) = \{X_1(t), \ldots, X_P(t)\}^T$ denote a P-dimensional binary-valued vector with a multivariate Bernoulli dependence structure at time t. The collection of time series $\{\boldsymbol{X}(t)\}_{t=1,\ldots,n}$ satisfies an ALARM (*a l*ogistic *a*uto*r*egressive *m*odel Agaskar & Lu, 2013) if its conditional probability distribution can be given as

Table 1. Summaries of the estimated link communities

Cluster	Type	Percent	Max. link prob.
LP	*L*ow link *p*robability	72%	0.0004
MPM	*M*oderate link *p*rob., *m*orning	13%	0.02
MPA	*M*oderate link *p*rob., *a*fternoon	8%	0.011
MPD	*M*oderate link *p*rob., *d*ay-long	3%	0.008
HPM	*H*igh link *p*rob., *m*orning	3%	0.063
HPD	*H*igh link *p*rob., *d*ay-long	2%	0.058

$$X_i(t) \mid \boldsymbol{X}(t-1), \ldots, \boldsymbol{X}(t-K) \sim \text{Ber}\left\{\text{logit}^{-1}\left(\sum_{k=1}^{K}\sum_{d=1}^{P} b_{ikd} X_d(t-k) + c_i\right)\right\}, \quad b_{ikd}, c_i \in \mathbb{R} \quad \text{for all } i, k, d, \tag{1}$$

where $\text{logit}^{-1}(x) = \exp(x)/[1+\exp(x)]$. The coefficients b_{ikd} represent the temporal dependence of X_i on the previous values of the complete vector $\boldsymbol{X}(t-1), \ldots, \boldsymbol{X}(t-K)$, offering not only an autoregressive model for an edge, but also the possibility to model lagged cross-edge dependence. The ALARM model was originally introduced to model observed binary time series, whose covariance was governed by a graph. This graph would determine the linear term in our GLM, the $\{b_{ikd}\}$ terms of. We have modified the modelling strategy of Agaskar and Lu (2013) to generate block structure, but not allowed on the order of $(N^2/2)\cdot n$ (or higher) potential cross-dependencies, for n temporal observations. We, therefore, use the word *possibility*, as inferring $\binom{N}{2}^2$ random latent variables remains computationally unfeasible in most real examples. Depending on the value of the coefficients b_{ikd}, a wide range of associations can be modelled, including negatively correlated processes within the framework of one single model. The coefficients c_i adjust the overall marginal probabilities of the component Bernoulli processes.

The relationship between the parameters of the model (1) and observable features such as autocorrelation and stationary marginal probability of the realised process is not straightforward for higher autoregressive orders, but for an illustration, we show these for a range of parameter pairs for a first-order autoregressive ALARM model for $D = 1$, that is, a single binary time series:

$$X_t \mid X_{t-1} \sim \text{Ber}\left\{\text{logit}^{-1}(b\,X_{t-1} + c)\right\}, \quad b, c \in \mathbb{R}. \tag{2}$$

The stationary probability of an edge and the correlation between two consecutive edge values can be computed in this case either directly, or using the Markov character of the process as $\Pr(X_i = 1) = e^c(1+e^{b+c})/(1+2e^c+e^{b+c})$, and $\text{Corr}(X_i, X_{i-1}) = (e^{b+c}-e^c)/(1+e^c)(1+e^{b+c})$.

Although negative values of the linear coefficient b are associated with negatively correlated processes, and positive values with positively correlated ones, the relationship is not linear, and the value of the constant c also has an influence on the relationship. Moreover, maybe somewhat counter intuitively, both large negative and large positive b values can produce near-zero autocorrelations when they are associated with large constants c of the same sign. Note that while positively correlated binary time series can have any marginal probability, negatively correlated ones can have only much more restricted marginal probabilities as shown theoretically by Teugels (1990) and Chaganty and Joe (2006).

Statistical tests are needed to decide whether we should include autocorrelation into the model for a data set or not. One such test may be the comparison of the proportion of switched-on link states in the time series to an estimate of the edge probability based on the geometric distribution of the run lengths of the states. The two coincides only under independence, since the run lengths will no longer have a geometric distribution if the time series is dependent. Another possibility is to

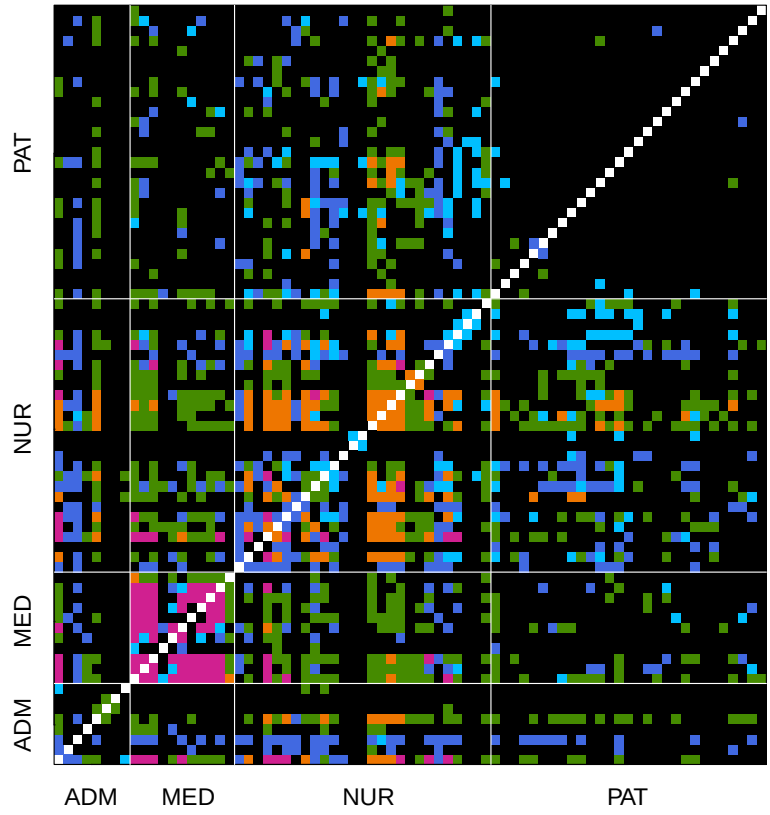

Figure 11. Cluster memberships of the edges (arranged in an adjacency matrix format) in the best model with six clusters and $H = 3$ for the hospital data. The colours indicate the different link communities the edges belong to (black, LP; red, HPD; orange, HPM; green, MPM; light blue, MPD; dark blue, MPA; for the naming, see Table 1). The arrangement of the nodes is the same as in Figure 5.

check the validity of the geometric distribution for the run lengths, either by a simple quantile–quantile plot or by an (approximate) Kolmogorov–Smirnov type distributional equivalence test. We will show an example of such a test for our data example in Section 4.3.

3.2 The block-ALARM model

Using the ALARM model, we can make the state of a link between two nodes depend directly on the state of the link at previous times. Link communities may then be assumed to follow distinctive temporal dependence models, with block-wise different parameters.

Let us suppose we are dealing with a series of snapshots from an undirected network, observed at times $t_1, \ldots, t_n$, all on the same node set consisting of N nodes. Let $A_{kj}(t_l)$ denote its (symmetric) adjacency matrix at time t_l. Define the (one-to-one) mapping $\sigma : \{(k, j) : k < j; k, j = 1, \ldots, N\} \mapsto \{1, \ldots, N(N-1)/2\}$. Define the collection of random variables $\{X_{il}\}$ by the induced mapping $X_{il} = X_{\sigma(kj),l} = A_{kj}(t_l)$. Assume that each edge can belong to one of G link communities, and let the variable $Z_i \in \{1, \ldots, G\}$ indicate the membership of edge X_{il} for all time indices l (we assume that the membership of the edge does not vary over time). Our model, which we call block-ALARM (BALARM) model, can then be written as

$$X_{il} \mid X_{i,l-1}, \ldots, X_{i,l-K}, \; Z_i = g \sim \mathrm{Ber}\left\{\mathrm{logit}^{-1}(\eta_{ilg})\right\}, \tag{3}$$

where η_{ilg} is a linear predictor containing the characterisation of the system such as autoregressive terms, temporal patterns expressed by explicit functions of time (for instance, a harmonic model), and covariates characterising the links.

In the case when the model is supposed to contain only autoregressive terms, but no covariates or temporal patterns, its form is

$$\eta_{ilg} = \sum_{k=1}^{K} b_{kg} x_{i,l-k} + c_g, \tag{4}$$

where $b_{kg} \in \mathbb{R}$ represents the order k autoregressive parameters in link community g, and $c_g \in \mathbb{R}$ determines the link probability value for link community g when all the preceding k time series values are zero. If the model is supposed to have a deterministic variation over time, such as in the presence of a typical daily pattern, this can be modified by adding terms containing time explicitly:

$$\eta_{ilg} = \sum_{d=1}^{D} a_{dg} f_d(t_l) + \sum_{k=1}^{K} b_{kg} x_{i,l-k} + c_g, \tag{5}$$

where $f_d(t)$ is an appropriate basis, for instance, harmonic functions in the case of a periodic temporal evolution. Here, we let $t_l = (l-1)\Delta$ for some sampling period Δ, which we shall not dwell on, and when possible we set $\Delta = 1$. The class of ALARM models falls into the class of Generalised Autoregressive Moving Average Models (Benjamin et al., 2003), for further discussion of this and related models, see Armillotta et al. (2022).

Observing a collection of edges with unknown memberships, and supposing that an edge can be a member of a single cluster, we also assume a multinomial model for memberships: $Z_i \sim \text{Multinom}(\pi_1, \ldots, \pi_G)$. Here, π_g is the probability of an edge to belong to cluster g. The complete-data likelihood of the model can then be written as

$$L(\theta; \{x_{il}\}, z_i) = \prod_{i=1}^{J} \prod_{g=1}^{G} \left\{ \pi_g \prod_{l=K+1}^{n} \left[\frac{\exp(\eta_{ilg})}{1+\exp(\eta_{ilg})} \right]^{x_{il}} \left[\frac{1}{1+\exp(\eta_{ilg})} \right]^{1-x_{il}} \right\}^{I(z_i=g)}, \tag{6}$$

where the parameter θ represents all parameters π_g, a_{dg}, b_{kg}, and c_g, and J is the number of edge variables in the observed network at a fixed time t_l. The corresponding log-likelihood is

$$\ell(\theta; \{x_{il}\}, z_i) = \sum_{i=1}^{J} \sum_{g=1}^{G} I(z_i = g) \left\{ \log \pi_g + \sum_{l=K+1}^{n} \left[x_{il} \eta_{ilg} - \log(1 + e^{\eta_{ilg}}) \right] \right\}, \tag{7}$$

with η_{ilg} defined in Equation (5). This model can be fitted using the EM algorithm (Dempster et al., 1977). We have here grouped edges that are governed by the same parameters, rather than nodes. This allows more flexibility in the behaviour of a node. A person may behave like others just in certain environments, and there are versions of this such as the mixed membership model (Airoldi et al., 2005). However, in the mixed membership model the grouping of behaviour is structurally constrained by an underlying nodal block structure that is not enforced here.

4 Simulation studies and data analysis

4.1 Simulation study of AR(1) processes

To explore the performance of the model described above when applied to autocorrelated network data with a mixture of various edge probabilities ranging from moderately high to extremely low (similar to our data example), we simulated a series of BALARM models using the following ALARM(1) processes:

$$\begin{aligned}
&\text{Cluster A:} \quad b_1 = 2.89, \quad c_1 = -1, \quad p_1 = 0.67;\\
&\text{Cluster B:} \quad b_2 = 4.48, \quad c_2 = -4, \quad p_2 = 0.045;\\
&\text{Cluster C:} \quad b_3 = 5.43, \quad c_3 = -5, \quad p_3 = 0.016;\\
&\text{Cluster D:} \quad b_4 = 6.42, \quad c_4 = -6, \quad p_4 = 0.006,
\end{aligned}$$

where p_i denotes the stationary marginal probability of the resulting Markov chain, and $\{b_i, c_i\}$ $(i = 1, \ldots, 4)$ corresponds to four different sets of coefficients in Equation (2). The lag-1 autocorrelation was set as 0.6 for all four processes. We created three two-component BALARM mixture models by combining Cluster A with each of the other three ALARM(1) models. This sequence represented a series of models in which all the edge processes had the same lag-1 autocorrelation, but they differed in their marginal link probabilities across a wide range, mimicking human interaction data in which link probabilities can range from very low to high, but where the persistence of edges is similarly high once switched-on. We generated $R = 200$ replicated data sets from each model, where each cluster contained 300 edge time series of a length of $n = 1{,}200$ (totalling 600 edges per model). The length of the time series and the low edge probabilities were chosen to mimic the situation with our data set. We fitted each data set using the procedure described in Section 3.2, using initial values randomly generated from a normal distribution centred on 0 and with a standard deviation of 0.5 for the EM algorithm.

Estimates of the marginal probabilities and the lag-1 autocorrelations from the fits are shown in Figure 2 (the histogram of cluster A, which occurred in all three models, is presented only once, although it was separately simulated and fitted for all three). Whereas the estimates for the clusters A and B appear to be reasonably good, the distribution of the estimated marginal probability for the two low-probability clusters is asymmetric. This is expected as an asymptotic normal theory of maximum likelihood estimates (which is approximated by the EM algorithm) breaks down for such low probabilities with the given time series length. Moreover, the estimate of the autocorrelation becomes unreliable, covering the whole [0, 1] interval, especially for cluster D with the lowest contact probability. This suggests that in data analysis with sparse contacts, which are quite typical in many applications of network analysis, we need to carefully consider the reliability of our estimates, and since asymptotic theory does not provide a sufficient quality of approximation, bootstrap methods are necessary.

4.2 Simulation study of an AR(2) process

We have also simulated a more complex case, with two different link communities, a periodic non-stationary component, and a second-order autoregressive temporal dependence (namely, a non-stationary BALARM(2)). The linear predictor of the model used was Equation (5), with parameter values chosen as $a_{11} = 0.15$, $a_{21} = 0.09$, $b_{11} = 0.3$, $b_{21} = 0.3$, $c_1 = 0.5$ for cluster 1 (denoted by G1 in Figures 3 and 4), and $a_{12} = -0.8$, $a_{22} = 0.2$, $b_{12} = -0.5$, $b_{22} = 0.5$, $c_2 = -0.55$) for cluster 2 (G2). We defined the base functions for the non-stationary time-varying term as $f_1 = \sin(2\pi t/P)$ and $f_2 = \cos(2\pi t/P)$ with period $P = 288$. One simulated dataset consisted of 300 independent realisations of length $n = 1{,}000$ from each of the clusters (that is, one dataset contained 600-time series, 300 belonging to each clusters). We produced 425 repetitions of this dataset, and performed estimation using the EM algorithm on each of them.

The distribution of the resulting parameter estimates is shown in Figure 3. We find that the fitting procedure is able to estimate the model parameters without large bias, including the second-order autoregressive parameters a_{21} and a_{22}. To judge the quality of the estimates better, we have reconstructed the time-varying contact probability patterns conditioned on the immediate lag-1 and lag-2 past of X_t being $X_{t-1} = X_{t-1} = 0$ for all t, both based on the estimated and the true parameter values. The 425 estimated curves are presented in Figure 4, together with the pointwise median of the estimates and with the true curve, illustrating convincingly the good quality of the estimates.

4.3 Data analysis

Our data set contains a high-resolution dynamic network of social interactions in a hospital ward, taken with the aim to identify crucial spreaders in a hypothetical epidemics (Vanhems et al., 2013). Social interactions between humans seem to be particularly in need to include correlations in their modelling, especially if observed at high temporal resolutions. Moreover, the strict daily schedules in a hospital imply daily varying contact probabilities between different groups in the hospital. The BALARM model, presented in Section 3.2, imposes a direct correlation between successive states of edges, and is adapted to provide a detailed model about the dynamics of the network over time, which can be particularly beneficial for modelling the unfolding of an epidemic.

4.3.1 University hospital social interaction data

Social interaction data over time from the geriatric short-stay ward of a university hospital in Lyon, France, was collected between Monday, 6 December 2010 at 1:00p.m. to Friday, 10 December 2010 at 2:00p.m., using RFID (radio frequency identification) devices attached to 29 patients (coded PAT in what follows), 27 nurses (NUR), 11 medical doctors (MED) and 8 administrative staff (ADM), in total $N = 75$ individuals (Vanhems et al., 2013). These node categories will be termed 'status', following Vanhems et al. (2013). The RFID gives a contact signal if it is able to exchange radio signals with another RFID, which happens when their owners stand closer than about 1.5 m from each other. This closeness was used as a proxy for a contact between two persons. Every 20 s, the presence or absence of these contact signals during the preceding 20 s period were recorded between each pair of devices.

For our use, we aggregated the data into 5 min snapshots, by defining the adjacency value $A_{kj}(t_l) = 1$ at time t_l between nodes $k, j \in \{1, \ldots, 75\}$ if there was at least one contact signal in the preceding 5 min between RFIDs i and j, and $A_{kj}(t_l) = 0$ if there was none. We used the time series $A_{kj}(t_l)$, $l \in 1, \ldots, n$ with $n = 1{,}159$ as our input data. The number of edge variables in the adjacency matrices is $J = N(N-1)/2 = 2{,}775$.

4.3.2 Choice of model

The plot of the average edge probabilities arranged in an adjacency matrix format, shown in Figure 5, indicates a block structure, which nevertheless does not fully coincide with the status of the nodes in the hospital, although a notable overlap exists in the case of doctors. This suggests that in these data, a hidden node cluster membership (related to but not identical with the hospital status) may explain at least part of the network structure within the framework of a time-varying SBM. However, a link community may, for example, provide the possibility of recognising subclusters of edges in different node communities with similar temporal patterns and correlations, or to scrutinise whether, in this data example, information on the status of the individuals in the hospital is sufficient to fully determine the patterns of the interactions.

The plot of the number of contacts $\sum_{k \neq j} A_{kj}(t_l)$ versus t_l, shown in Figure 6, suggests daily repeated patterns through the time span of the records, corresponding to the strict daily routine in a hospital. Since we are analysing human interaction data, which is typically autocorrelated, we should also test for the necessity of including an autoregressive term. For possibilities, we refer to Section 3.1. For our data, we use geometric quantile–quantile plots, where we estimated the marginal probability of each time series as $\hat{p}_{kj} = \sum_{l=1}^{T} A_{kj}(t_l)/T$. We show these for some edges in Figure 7, which indicates a discrepancy from the geometric distribution. However, in our case, the visible presence of the periodically varying contact probabilities can also cause this. In our model, we will include an autoregressive term, and will test for its significance using bootstrap.

We, therefore, complemented the linear predictor of the model (7) with a H-order harmonic series with a period P equal to a day ($P = 288$ in 5-min units):

$$\eta_{ilg} = \sum_{d=1}^{D} a_{dg} f_d(t_l) + \sum_{k'=1}^{K} b_{k'g} x_{i,l-k'} + c_g, \tag{8}$$

where $D = 2H$, $f_d(t) = \cos(2\pi \lceil d/2 \rceil P^{-1} t)$ for $d = 1, 3, \ldots, 2H-1$ and $f_d(t) = \sin(2\pi (d/2) P^{-1} t)$ for $d = 2, 4, \ldots, 2H$, and $\lceil \cdot \rceil$ stands for the function ceiling. Moreover, as we model the self-maintaining nature of human contacts, we suppose the autoregressive order to be $K = 1$.

The model was fitted using the EM algorithm (Dempster et al., 1977), for a range of different choices for the number of link communities ($G = 2, \ldots, 9$) and harmonic order ($H = 2, 3, 4$). The best model was selected by the Bayes Information Criterion (Schwarz, 1978), since BIC is a consistent selector of model complexity in clustering models and in linear modelling, and thus for our large data set, we can expect good performance. Moreover, according to Brewer et al. (2016), in cases like ours when the regression model is not expected to contain collinear variables, BIC slightly outperforms AIC and AIC_c in terms of its power to select the correct model. We present the selected model fit in the next sections.

4.3.3 Results

Figure 8 shows the resulting BIC values from the model fits. The overall best fit is the model with $H = 3$ and $K = 8$. However, the decrease in BIC values for $K > 6$ is small in comparison to the improvement on models with $K \leq 6$, and especially with the apparent best harmonic order $H = 3$, the models are practically equivalent above $K = 6$. Based on the principle of parsimony, we chose the model with $H = 3$ and $K = 6$, as the model representing the best compromise between quality of description and simplicity. A summary of the basic parameters of the communities in this model and our notation for the clusters is given in Table 1.

From the estimated model parameters, we can derive the average temporal variation of both the link probability and the lag-1 autocorrelation, and the most likely link community membership of each edge. The estimated memberships are shown in Figure 11, and the time-varying link probabilities and autocorrelations in Figures 9 and 10 (in heavy solid lines). The very low contact probabilities are reflecting the fact that in the data set, most (nearly 60%) of the edges have no contacts at all through the whole observation period. These links are appropriately attributed to the LP cluster, for which, accordingly, the maximal contact probability in any 5 min interval during a day is estimated at only 0.0004. Those links that do have at least one contact over this time have on average around three contacts during the data collection. The highest link forming probabilities belong to the HPM and HPD communities, still with a value of only about 0.06. With such low probabilities, we will resort to bootstrap for inference on the estimated daily patterns and autocorrelations.

We performed a parametric bootstrap analysis. We simulated 500 repetitions of the model using the estimated value of the parameters, and repeated the estimation on them. The initial values for the EM algorithm were fixed at the values used to perform the simulations (that is, the estimated model parameters), to facilitate the identification of clusters. Finally, from the estimated bootstrap parameters, we reconstructed the temporal pattern of the marginal probabilities for all six clusters and the correlations for the two clusters with the highest link probability levels, for which we can hope for a realistic correlation estimation. The results are shown in Figures 9 and 10.

Figure 9 shows that the 95% pointwise confidence bands are quite wide, but broadly support the estimated daily patterns, such as with the two-peaked aspect of clusters HPM and HPD, the low-activity tail of MPM stretching into the early afternoon, and the day-long, moderate-level activity of MPD (for the definitions, please see Table 1). The median of the bootstrap estimates matches very well the estimates on the real data, which indicates that the method estimates the contact probabilities in a reliable way. The broad uncertainty bands are not surprising, given the sparse contacts due to the extremely low contact forming probabilities together with a small degree of freedom. The lower limit of the bands, despite its appearance, does not include $p = 0$, since the inverse logit transform does not allow for the probabilities to be precisely 0 or 1. However, they can get arbitrarily small. Longer observational time would be necessary to put a more stringent lower limit on the estimates. The spiky look of the confidence bands in panel LP is due to the combination of the extraordinarily low link probability (at most 0.0004) of the cluster and the inverse logit transformation.

It is not reasonable to calculate the lag-1 autocorrelations for the four low-probability clusters MPM, MPD, MPA and LP using time series of this length, as our simulations in Section 4.1 illustrated. Nevertheless, it can be calculated and estimated for the two clusters with the highest link probabilities, as shown in Figure 10. The daily pattern of the correlations are again supported by the bootstrap estimation, and are markedly bounded away from 0, indicating that it is indeed necessary to include an autoregressive term into the modelling of this data set. However, the correlations appear to be estimated with a negative bias. This suggests an even stronger correlation of human relations in reality than that implied by our model estimates, and underlines the importance to incorporate this autoregressive nature into modelling efforts.

4.3.4 Interpretation

The model fit offers a very detailed insight into the social network of a hospital ward. Several interesting conclusions can be drawn.

Relationship to hospital status. The patterns discovered in Figure 11 are similar to the block patterns suggested by the time-averaged link probabilities in Figure 5. It appears thus that the link

clustering performed by the BALARM model is at least partly based on the time-averaged link probability of the different edges, and overlaps with what we would expect from a SBM which is based on the similarity of contact probabilities within blocks. However, the resulting link community model does not coincide fully with the clusters defined by status in the hospital.

Discrimination of link communities based on different edge dynamics. The model does not exclusively base its decision on the time-averaged contact probability of the edges. Many dynamic SBMs in the literature (Ludkin et al., 2018; Matias & Miele, 2017; Matias et al., 2018; Olivella et al., 2021; Pensky, 2019) suppose that the edge probabilities of the different blocks are constant over time, and the dynamics of the networks are determined by the latent process of the nodes moving among the blocks. Some models have conditionally independent edges with time-varying edge probabilities, while our model includes explicit correlation in the observed edges over time coupled with time-varying characteristics. In fact, the blocks in our model are discriminated based on their different time series characteristics, as our model definitions (7) and (8) imply. Figure 9 shows this very clearly: the daily variations of the contact probabilities of the six link communities are visibly different. For example, although the HPM and the HPD clusters have a qualitatively different two-peaked shape, the HPM cluster starts activity earlier than the HPD, at about 7a.m. when the HPD is still very close to 0 link probability confirmed by its bootstrap confidence bands. The activity of HPD cluster lasts longer than the HPM, being still highly active around 5p.m. when the HPM community has already ceased to be active. In addition, we are able to describe the dynamics of edges as we model their correlation.

Realistic picture of the dynamics. The structures in Figure 11, together with the dynamics in Figure 9 capture many realistic details from the life of a hospital, which lends credibility to the model fit.

- We have identified two clusters with the most frequent contacts, HPM (orange in the figures) and HPD (purple). Figure 11 shows that one of them corresponds mostly to interactions between doctors, and the other is mostly associated to nurse-nurse interactions. It appears that as far as doctor-doctor interactions are concerned, they form their own near-exclusive block. Those doctors not following the same daily pattern in their interactions (a mostly black and a mostly green row in the MED-MED block in Figure 11) may be a nutritionist and a physiotherapist who, according to the description of the data set in Vanhems et al. (2013), visited the ward occasionally, but were not present as regularly as the resident medical staff. They also have more sporadic interactions with the nursing staff than the other doctors, and less contact with patients and the administration too. Those doctors belonging to the main cluster HPD have not only very similar interaction patterns and link probabilities among themselves, but quite similar interaction patterns with nurses (those edges belonging mostly to the cluster MPM, green in the plots), and with patients too. Perhaps somewhat surprisingly, their edges with patients belong in majority to the cluster LP (black in the plots) and to the cluster MPM, which are the two clusters with the lowest contact probabilities. We can also draw the conclusion that for the modelling of the interaction of doctors with everybody else, a 2-component SBM might be an adequate model.
- Link structure within the block of nurses is far more complex. The NUR-NUR block is itself sub-divided into two large and at least one smaller block. (1) Some nurses interact with each other mostly in the morning (a mostly green and orange block along the diagonal, with HPM (orange) or MPM (green) contact probability patterns). Their interactions with patients also follow the morning patterns. (2) Another group interact within itself rather in the afternoon (predominantly dark blue block along the diagonal indicating the MPA cluster), though with more mixing of morning and afternoon patterns. Contacts with patients also belong to the MPA cluster.

The interaction of these two groups of nurses mostly goes by the morning patterns. This probably reflects both a division of the nurses into morning and afternoon shifts, and the existence of

an overlap between the shifts, which is a reasonable conclusion since flow of information about the patients will need some time for transmission.

- An anomalous group of nurses can be found as mostly black rows in the middle and at the top of the NUR-NUR block in Figure 11. Almost all their contacts with other staff members belong to the LP (black) cluster. With patients, however, many of their links belong to the cluster not mentioned so far, the MPD cluster (light blue in the plots). More complex models, such as the BIC-best 3-harmonic, 8-cluster model or the 4-harmonic, 6-cluster model, identify the NUR-PAT links of these nurses as a separate link cluster with a specific daily pattern consisting of a morning and an evening burst of activity. These bursts in these more complex models reach the highest link probabilities earlier in the morning and later in the evening than any other links. Based on this, we might guess that our model has identified the nurse-aides, those who have fewer medical tasks than the regular nurses or none at all, but care about the patients' basic needs such as getting dressed, washed or fed, possibly before or after the medical needs of the patients are satisfied during the workday.
- Doctors and nurses interact mostly in the morning, according to the MPM (green) link cluster pattern, with on average lower link probability than doctors have with other doctors, and a definitely different temporal pattern. Some links, however, belong to the MPA (dark blue) afternoon pattern. These two moderate link probability patterns make up those MED–PAT interactions too which do not belong to the low-probability LP (black) cluster.
- Patients have almost no contact with each other. They also have on average much fewer contacts with anybody else than the others with each other. They have the most frequent contacts with the nurses, but even that does not reach the average level of interactions which is observed between nurses and doctors. This, although striking at first sight, is perhaps expected. Patients in a geriatric short-stay ward may be seriously ill, affected by neurodegenerative diseases, and generally not in the mood of making contacts beyond the necessary (visitors were not tagged with an RFID (radio frequency identification device), and were not followed in the experiment).

4.3.5 Cross-correlations between edges

The question can be asked whether we need to include between-edge (possibly lagged) cross-correlations as well as the temporal autocorrelations. In the presence of periodically varying link probabilities of the clusters, apparent correlations may be found simply due to the similarities in the temporal patterns of some edges: contacts may be observed simultaneously simply because of their higher probability at some times at some lags, giving rise to spurious cross-correlations. To check for this, we simulated 5,000 time series of a length equal to the observed data, independently from each of the clusters of our model, and computed the cross-correlations between them.

The main results are presented in Figure 12. The only strong difference between the simulations with no cross-correlations (in grey) and the real data (red) was between two HPM or two HPD edges, as shown in the left two panels of Figure 12. This suggests that cross-correlation might exist between such link types. However, it is also possible that some insufficiently modelled temporal patterns give rise to these apparent excess correlations.

We found only small discrepancies for any other edge combination, of which two are shown in the right-hand panels (HPM-HPD and HPD-MPA). The existence of cross-correlations can, of course, not be ruled out. Indeed, our finding may just mean that, similarly to the autocorrelation, for such low contact probabilities the estimation of cross-correlation needs more observations to be reliable.

4.4 Comparison to other analysis methods

4.4.1 Comparison to Jiang et al. (2020)

The data set from a university hospital in Lyon was also analysed by Jiang et al. (2020). In that study, the researchers used a rougher aggregation of the data than we did, taking $A_{kj} = 1$ if there was at least one contact between nodes k and j during any given day. This was needed since their

model is stationary by construction, but the data at aggregations finer than one day are not stationary as they possess strong cyclostationary features. Jiang et al. (2020) found no evidence for the existence of nodal communities, and no evidence of significant autoregressivity. It is not surprising that our findings differ from those of Jiang et al. (2020), as the absence of autoregressivity at the timescale of a day is plausible in the data. We expect human contacts to be strongly correlated on shorter timescales (∼minutes), but much less so over days, if not for specific circumstances. As to the nodal clustering, our conclusion that there may be an approximate SBM-like block structure at least for a MED-(everyone else) division relies strongly on the estimated sub-daily contact probability patterns, and thus may remain undetectable with long aggregation times.

4.4.2 *Comparison to Olivella et al. (2021)*

For the sake of comparison, we modelled the data using another recent model. Among the many possibilities, for example Ludkin et al. (2018), Matias and Miele (2017) and Matias et al. (2018), we have chosen a method that permits us to include non-stationary terms in the contact probabilities, and thus providing an opportunity to model the daily patterns obvious in the data, but which (similarly to the majority of the statistical dynamic network models) is based on nodal clustering: the hierarchical mixed membership model of Olivella et al. (2021).

The model of Olivella et al. (2021) assumes not only latent mixed membership classes but also latent temporal classes. Thus given a latent state of 'day' or 'night' the slice of adjacency tensor have a distribution governed by a set of mixed membership probability vectors in G possible nodal clusters. It is also possible to incorporate nodal covariates in this model, such as in our dataset the status MED, ADM, NUR, or PAT of the study participants, by means of a logistic linear model. A Bernoulli variable is describing the existence or the absence of a contact between any two nodes. The model is implemented in the R package NetMix.

Initially, we assumed a two-state latent Markov chain (perhaps describing day and night shifts), using the status in the hospital as nodal covariates, and a harmonic model of order 3 as edge covariates, similarly to our fitted BALARM model. However, it turned out that the Markov chain switches between states very rarely. Indeed, intuitively the harmonic model is able to describe the variation of the activity between day and night shifts, as it repeats with a daily period. Thus, we dropped the assumption of a two-state latent Markov chain, and fitted the model using only one state and the harmonic model.

The results of the data analysis are summarised in Figure 13. The blockmodel is shown in the left panel, representing the temporally constant blockmodel structure of the network. The time-averaged share of three nodal clusters, G1, G2, and G3, in the population, is represented by the size of the circles corresponding to the clusters. Edge probabilities are indicated by the grey shade of the lines connecting the clusters to each other or to themselves. The cluster with the overall highest contact probability, G2, is instantiated most probably by ADM, MED, or NUR nodes, and much less likely by PAT nodes. Edges formed by nodes of which at least one belongs to this nodal cluster may roughly correspond to our link communities with the highest contact probability, namely, HPM and HPD, perhaps MPM, which are indeed most often formed between nurses, medical and administrative staff, but rarely including patients. It is also visible that we have edges of very low-probability contact probability, mainly when one or both nodes involved belong to G3, the one instantiated preferentially by patients. Contacts between nodes belonging to G1 and G2 or both belonging to G2 are more likely than contacts involving at least one node belonging to G3, but less likely than when both nodes are in G1.

The shape of the temporal patterns found, shown in the right panel of Figure 13, appear uniform. This is due to the model setup that does not allow the edges to have different temporal variations, unlike in the BALARM model. Thus, regardless of what nodal cluster the forming nodes belong to, all edges have the same temporal pattern, albeit possibly multiplied with a different scalar constant corresponding to a different average contact probability.

In contrast, our model is able to find the distinct time variations characterising the different roles in a hospital ward. Because of the strictly regulated schedule of the activities in a hospital, the BALARM model is better suited to give a detailed summary of the dynamics of the underlying hospital social network. Interpretation of the estimated nodal clusters (G1, G2 and G3) is also much less clear than the link communities in the BALARM model. In fact, G1, G2, and G3 are

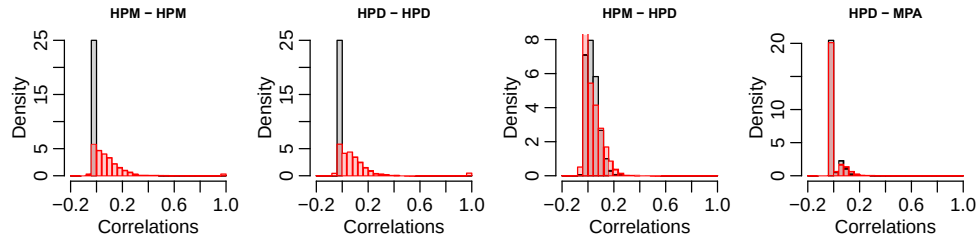

Figure 12. Histogram of cross-correlations between edge time series belonging to different link communities from the real data (red) and from simulations from the fitted model (grey).

discriminated based on not much more than their average contact probability. However, in this data problem (and in many more involving human interactions which usually have a daily rhythm) there is a far greater wealth of distinguishing information if one considers link communities and allows them to have their own clustering than if one considers solely nodal clustering. A combination of the two may be beneficial for modelling dynamic human societies.

5 Discussion

In our study, we proposed an approach combining binary-valued time series models with mixture modelling, in order to model the dynamics of networks describing human interactions. The nature of the data, which was collected in a hospital ward during four workdays with a high temporal resolution, raised an important modelling questions. Namely, how to account for the likely strong temporal autocorrelation and repetitive time-varying patterns present in the data, and in human interactions in general? Conditional independence and thus the connecting graphon framework may not be able to adequately model the strong and direct autoregressivity of human contacts: we do not decide randomly and independently at each moment whether we continue a conversation. This leads us out of the most often used family of models, the SBM (e.g., Matias & Miele, 2017; Olivella et al., 2021; Pensky, 2019; Xu & Hero, 2014), which are based on conditional independence. We proposed our temporal link community model (which we call BALARM) using the GLM framework in reply to this question, and in order to explore its potential in real-life situations, and applied it to hospital data.

Our BALARM modelled the time series of the elements of the adjacency matrix of the network as an autoregressive binary time series, assuming that each of these edges belonged to a single link cluster. The time series model for each edge included deterministic, time-varying linear components with a period of a day to account for the workday patterns at the hospital, and a linear autoregressive term, capturing the fact that, conditioned on the immediate past, the contacts' existence does not depend on the more distant past (a reasonable assumption for these contacts focusing mostly on current events and problems in the hospital), but it does depend on the immediate past. The logit transformation linked this linear predictor to the instantaneous probability of the edge to become switched-on. We further assumed that there is a finite number of link communities (clusters) the edges can belong to, and that within these clusters, the parameters of the time series are constant. Furthermore, we assumed that the model is identifiable in a sense that at least one parameter differs for two different link clusters.

We fitted the model using the EM algorithm for a range of model complexities, and selected the fit realising the best trade-off between simplicity and richness of interpretation according to the BIC. We obtained inference about the quantities of interest such as the daily contact probability patterns of the different communities and the autocorrelation over time produced by the fit by bootstrap. The results gave an unprecedentedly detailed insight into the time variations of the contacts of a social network. We could distinguish six clusters of different typical link probability variations over the day. These daily patterns could be put into realistic correspondence with the normal working day in a hospital, and also with the status of the persons in the hospital that formed them. We observed that although much of the link community structure found was clearly related to the status of the individuals in the hospital (namely, medical or administrative staff, nurse or patient), no perfect correspondence with an SBM could be found.

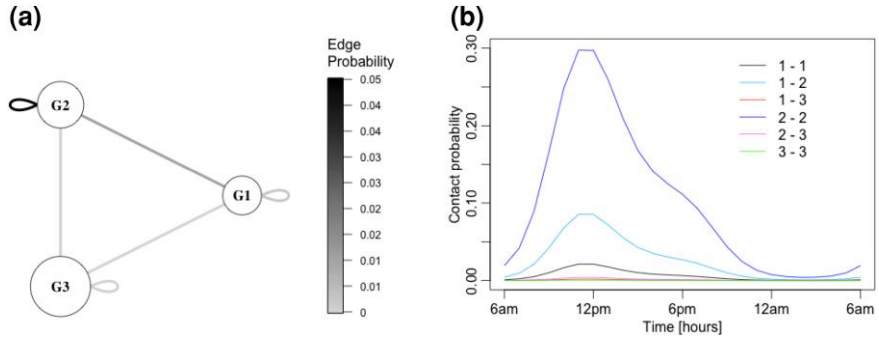

Figure 13. Model structure and edge probabilities in the model of Olivella et al. (2021). The left panel shows the time-independent component of the SBM, indicating probability levels of edge formation in shades of grey of the lines, and time-averaged membership proportion as the size of the circles. The right panel shows the temporally dependent component of the edge probabilities between the different groups.

Our approach differs in several aspects from the currently available models (Jiang et al., 2020; Ludkin et al., 2018) that explicitly take into account the autocorrelated nature of the links. The most important of these is that we have deliberately stepped out of the usually considered framework of node clustering, and instead, considered edges as the basic units of modelling. Both of the analysis methods of Jiang et al. (2020) and Ludkin et al. (2018) is based on node clustering (if with autoregressivity directly into the network's temporal dynamics). In terms of implementation for technical reasons Ludkin et al. (2018) does effectively use link clusters within their methodology. Our model thus provides an alternative to these models for social situations where the nature of the links themselves is the subject of investigation.

Another important difference is that the time series model at the core of our model is the classical GLM framework (McCullagh & Nelder, 1989), which has not yet been used in the modelling of dynamical networks, and which confers several significant advantages to our model. First, it offers the possibility to fit both negatively and positively correlated networks within one single model. This has strong relevance in practice, when trying to find link communities in real data. If it is possible that the data contain both positively and negatively correlated time series, most models currently in use need to set up two formally different models, which causes difficulties such as obtaining correct inference including uncertainty arising from the decision about positivity or negativity of correlation in the model. Our model accommodates naturally both possibilities within a single modelling framework. Another advantage provided by our framework is the flexibility to specify the temporal evolution of the network. Higher-order autoregressive terms and further relevant covariates, such as the harmonic terms in our model, can be straightforwardly added to the linear predictor of the model, and thus the detailed analysis of both stochastic and deterministic components of the dynamics becomes possible. This flexibility allowed us to decipher a realistic image of the life of the hospital ward under investigation from the network data, and to detect communities and distinguish their daily patterns, where the SBM-based model of Jiang et al. (2020) hit the problem of non-stationarity.

In principle, our model also allows for the inclusion of cross-correlation terms into the model, leading effectively to a model with similarities to spatio-temporal models. However, care needs to be taken then on how to select the basis in which we express the model's various temporal and spatial characteristics, and the large number of latent variables that have to be determined (we could hypothetically have dependence between all $\binom{N}{2}$ edge variables). The computational challenges with such an approach are unsurmountable. Identifiability issues can arise due to confounding between the cross-correlation terms and the deterministic time variations due to similarities in time patterns. Another possible generalisation is the adaptation of mixed membership models for link communities, which may be important and interesting for applications where edges cannot be supposed to belong to one single link community. An example of this may be the contacts between two people whose contacts are at times 'collaborator' type and at other times 'friendship' type.

This needs to model edges which change membership over time. Mixed membership models could be the correct way to model such social networks. However, for such models, identifiability issues must be considered carefully.

We believe that our model provides a level of insight into the link community evolution that has not been previously reached yet. Moreover, it offers a model-based, controllable simplification to compress useful information of observed complex networks, which can be used as a building block to gain insight into processes on the networks, and is amenable to statistical inference. Such realistic model fits can serve, for example, as the basis for in-depth investigations of the spread of an infectious disease within a social unit, providing a detailed insight into the evolution of the disease in the community, and helping identify the most efficient intervention points. Their practical use, we hope, will be a strong spur in the future to develop both more easy-to-apply, useful, realistic models and the theory behind dynamic link community models.

Acknowledgments

We would like to thank the anonymous referees for their thoughtful criticism and would also like to thank the European Research Council under Grant CoG 2015-682172NETS, within the Seventh European Union Framework Program.

Conflict of interest: None declared.

Data availability

The data used in this paper can be downloaded from https://networks.skewed.de/net/sp_hospital and is openly available there.

References

Agaskar A., & Lu Y. M. (2013). ALARM: A logistic auto-regressive model for binary processes on networks. In *2013 IEEE Global Conference on Signal and Information Processing* (pp. 305–308). Austin, TX, USA: IEEEConference.

Ahmed A., & Xing E. P. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, *106*(29), 11878–11883. https://doi.org/10.1073/pnas.0901910106

Ahn Y.-Y., Bagrow J. P., & Lehmann S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, *466*(7307), 761–764. https://doi.org/10.1038/nature09182

Airoldi E., Blei D., Xing E., & Fienberg S. (2005). A latent mixed membership model for relational data. In *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 82–89). New York, NY, USA: Association for Computing Machinery.

Anderson C. J., Wasserman S., & Faust K. (1992). Building stochastic blockmodels. *Social Networks*, *14*(1–2), 137–161. https://doi.org/10.1016/0378-8733(92)90017-2

Armillotta M., Luati A., & Lupparelli M. (2022). Observation-driven models for discrete-valued time series. *Electronic Journal of Statistics*, *16*(1), 1393–1433. https://doi.org/10.1214/22-EJS1989

Benjamin M. A., Rigby R. A., & Stasinopoulos D. M. (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association*, *98*(461), 214–223. https://doi.org/10.1198/016214503388619238

Brewer M. J., Butler A., & Cooksley S. L. (2016). The relative performance of AIC, AIC_c and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, *7*(6), 679–692. https://doi.org/10.1111/2041-210X.12541

Chaganty N. R., & Joe H. (2006). Range of correlation matrices for dependent Bernoulli random variables. *Biometrika*, *93*(1), 197–206. https://doi.org/10.1093/biomet/93.1.197

Crane H. (2016). Dynamic random networks and their graph limits. *The Annals of Applied Probability*, *26*(2), 691–721. https://doi.org/10.1214/15-AAP1098

Crane H., & Dempsey W. (2018). Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*, *113*(523), 1311–1326. https://doi.org/10.1080/01621459.2017.1341413

Dempster A. P., Laird N. M., & Rubin D. B. (1977). Maximum likelihood from in-complete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Donnet S., & Robin S. (2019). 'Bayesian inference for network Poisson models', arXiv, arXiv:1907.09771, preprint: not peer reviewed.

Evans T. S. (2010). Clique graphs and overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, *2010*(12), P12037. https://doi.org/10.1088/1742-5468/2010/12/P12037

Faust K., & Wasserman S. (1992). Blockmodels: Interpretation and evaluation. *Social Networks*, *14*(1–2), 5–61. https://doi.org/10.1016/0378-8733(92)90013-W

Guo F., Hanneke S., Fu W., & Xing E. P. (2007). Recovering temporally rewiring networks: A model-based approach. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 321–328). New York, NY, USA: Association for Computing Machinery.

Hanneke S., & Xing E. P. (2006). Discrete temporal models of social networks. In *ICML Workshop on Statistical Network Analysis* (pp. 115–125). Springer.

Hoff P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics*, *9*(3), 1169. https://doi.org/10.1214/15-AOAS839

Hoff P. D., & Ward M. D. (2004). Modeling dependencies in international relations networks. *Political Analysis*, *12*(2), 160–175. https://doi.org/10.1093/pan/mph012

Holland P. W., Laskey K. B., & Leinhardt S. (1983). Stochastic blockmodels: First steps. *Social Networks*, *5*(2), 109–137. https://doi.org/10.1016/0378-8733(83)90021-7

Jiang B., Li J., & Yao Q. (2020). 'Autoregressive networks', arXiv, arXiv:2010.04492, preprint: not peer reviewed.

Kim Y., & Jeong H. (2011). Map equation for link communities. *Physical Review E*, *84*(2), 026110. https://doi.org/10.1103/PhysRevE.84.026110

Krivitsky P. N., & Handcock M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 29–46. https://doi.org/10.1111/rssb.12014

Liu X., & Duyn J. H. (2013). Time-varying functional network information extracted from brief instances of spontaneous brain activity. *Proceedings of the National Academy of Sciences*, *110*(11), 4392–4397. https://doi.org/10.1073/pnas.1216856110

Ludkin M., Eckley I., & Neal P. (2018). Dynamic stochastic block models: parameter estimation and detection of changes in community structure. *Statistics and Computing*, *28*(6), 1201–1213. https://doi.org/10.1007/s11222-017-9788-9

Matias C., & Miele V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *79*(4), 1119–1141. https://doi.org/10.1111/rssb.12200

Matias C., Rebafka T., & Villers F. (2018). A semiparametric extension of the stochastic block model for longitudinal networks: Semiparametric estimation in PPSBM. *Biometrika*, *105*(3), 665–680. https://doi.org/10.1093/biomet/asy016

McCullagh P., & Nelder J. (1989). *Generalized linear models*. Chapman & Hall/CRC.

Meng F., Zhang F., Zhu M., Xing Y., Wang Z., & Shi J. (2016). Incremental density-based link clustering algorithm for community detection in dynamic networks. *Mathematical Problems in Engineering*, *2016*. Article ID 1873504, https://doi.org/10.1155/2016/1873504

Minhas S., Hoff P. D., & Ward M. D. (2016). A new approach to analyzing coevolving longitudinal networks in international relations. *Journal of Peace Research*, *53*(3), 491–505. https://doi.org/10.1177/0022343316630783

Nguyen N. P., Dinh T. N., Xuan Y., & Thai M. T. (2011). Adaptive algorithms for detecting community structure in dynamic social networks. In *2011 Proceedings IEEE INFOCOM* (pp. 2282–2290). IEEE.

Olivella S., Pratt T., & Imai K. (2021). 'Dynamic stochastic blockmodel regression for network data: Application to international militarized conflicts', arXiv, arXiv:2103.00702, preprint: not peer reviewed.

Palla G., Barabási A.-L., & Vicsek T. (2007). Quantifying social group evolution. *Nature*, *446*(7136), 664–667. https://doi.org/10.1038/nature05670

Palla G., Derényi I., Farkas I., & Vicsek T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814–818. https://doi.org/10.1038/nature03607

Pamfil A. R., Howison S. D., & Porter M. A. (2020). Inference of edge correlations in multilayer networks. *Physical Review E*, *102*(6), 062307. https://doi.org/10.1103/PhysRevE.102.062307

Pensky M. (2019). Dynamic network models and graphon estimation. *Annals of Statistics*, *47*(4), 2378–2403. https://doi.org/10.1214/18-AOS1751

Ribeiro B., Perra N., & Baronchelli A. (2013). Quantifying the effect of temporal resolution on time-varying networks. *Scientific Reports*, *3*(1), 1–5. https://doi.org/10.1038/srep03006

Schein A., Paisley J., Blei D. M., & Wallach H. (2014). Inferring polyadic events with Poisson tensor factorization. In *Proceedings of the NIPS 2014 Workshop on" Networks: From Graphs to Rich Data*. Montreal, QC, Canada: From Graphs to Rich Data.

Schwarz G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Snijders T. A. B. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology*, *31*(1), 361–395. https://doi.org/10.1111/0081-1750.00099

Snijders T. A. B., & Nowicki K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, *14*(1), 75–100. https://doi.org/10.1007/s003579900004

Teugels J. L. (1990). Some representations of the multivariate Bernoulli and binomial distributions. *Journal of Multivariate Analysis*, *32*(2), 256–268. https://doi.org/10.1016/0047-259X(90)90084-U

Vanhems P., Barrat A., Cattuto C., Pinton J.-F., Khanafer N., Régis C., Kim B., Comte B., & Voirin N. (2013). Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLOS ONE*, *8*(9), e73970.

Xu K. S., & Hero A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing*, *8*(4), 552–562. https://doi.org/10.1109/JSTSP.2014.2310294