Bayesian Nonparametrics for Sparse Dynamic Networks

Cian Naik[1][0000−0002−5774−4014] (✉ cian.naik@stats.ox.ac.uk), François Caron[1][0000−0002−3952−224X], Judith Rousseau[1][0000−0002−0998−6174], Yee Whye Teh[1,3][0000−0001−5365−6933], and Konstantina Palla[2][0000−0001−7066−0539]

[1] Department of Statistics, University of Oxford, Oxford, United Kingdom
[2] Microsoft Research, Cambridge, United Kingdom
[3] Google Deepmind, London, United Kingdom

Abstract. In this paper we propose a Bayesian nonparametric approach to modelling sparse time-varying networks. A positive parameter is associated to each node of a network, which models the sociability of that node. Sociabilities are assumed to evolve over time, and are modelled via a dynamic point process model. The model is able to capture long term evolution of the sociabilities. Moreover, it yields sparse graphs, where the number of edges grows subquadratically with the number of nodes. The evolution of the sociabilities is described by a tractable time-varying generalised gamma process. We provide some theoretical insights into the model and apply it to three datasets: a simulated network, a network of hyperlinks between communities on Reddit, and a network of co-occurences of words in Reuters news articles after the September 11th attacks.

Keywords: Bayesian nonparametrics · Poisson random measures · networks · random graphs · sparsity · point processes

1 Introduction

This article is concerned with the analysis of dynamic networks, where one observes the evolution of links among a set of objects over time. As an example, links may represent interactions between individuals on social media platforms across different days, or the co-occurrence of words across a series of newspaper articles. In each case the pattern of these interactions will generally vary over different time steps. Probabilistic approaches treat the dynamic networks of interest as random graphs, where the vertices (nodes) and edges correspond to objects and links respectively. In the graph setting, sparsity is defined in terms of the rate in which the numbers of edges grows as the number of nodes increases. In a *sparse* graph the number of edges grows sub-quadratically in the number of nodes. Hence, in a large graphs, two nodes chosen at random are very unlikely to be linked. Sparsity is a property found in many real-world network datasets [34, 36], and in our work we are concerned with modelling sparse networks.

Bayesian approaches play an important role in the modelling of random graphs, providing a framework for parameter estimation and uncertainty quantification. However, most of the popular Bayesian random graph models result in dense graphs, i.e. where the number of edges grows quadratically in the number of nodes, see [36] for a review. A recent Bayesian nonparametric approach, proposed by [7] and later developed in a number of articles [40, 20, 4, 39, 33], seeks to solve this problem by representing a graph as an infinite point process on $\mathbb{R}_+^2$, giving rise to a class of sparse random graphs. This class of sparse models is projective and admits a representation theorem due to [25].

In this paper, we are interested in the dynamic domain and aim to probabilistically model the evolution of sparse graphs over time, where edges may appear and disappear, and the node popularity may change over time. We build on the sparse graph model of [7] and extend it to deal with time series of network data. We describe a fully generative and projective approach for the construction of sparse dynamic graphs. It is challenging to perform exact inference using the framework we introduce, and thus we consider an approximate inference method, using a finite-dimensional approximation introduced by [30].

The rest of the article is structured as follows. In Section 2 we give some background on the sparse network model of [7]. Section 3 describes the novel statistical dynamic network model we introduce in detail, as well as its sparsity properties. The approximate inference method, based on a truncation of the infinite-dimensional model, is described in Section 4. In Section 5 we present illustrations of our approach to three different dynamic networks with thousands of nodes and edges.

2   Background: model of Caron and Fox for sparse static networks

Bayesian nonparametrics provides a natural setting for the study of sparse graphs. Parameters can be infinite dimensional, and thus the complexity of models can adapt to data in question. In the context of network modelling, this allows for the consideration of graphs that may have infinitely many nodes, only finitely many of which form connections.

To this end, instead of the standard approach of representing a graph G by a finite dimensional adjacency matrix, [7] instead represent it by a point process. Letting $\alpha > 0$ be a positive parameter tuning the size of the network, a finite multigraph of size $\alpha > 0$ is represented by a point process on $[0, \alpha]^2$

$$N = \sum_{i,j} n_{ij} \delta_{(\theta_i, \theta_j)},$$

where $n_{ij} = n_{ji} \in \{0, 1, 2, \ldots\}$, $i \leq j$, represents the number of interactions between individuals i and j. The $\theta_i \in [0, \alpha]$ can be interpreted as node labels, and $\delta_{(\theta_i, \theta_j)}$ denotes a point mass at location (θ_i, θ_j). The node labels are introduced for the model's construction, but are not observed nor inferred. Each node i is
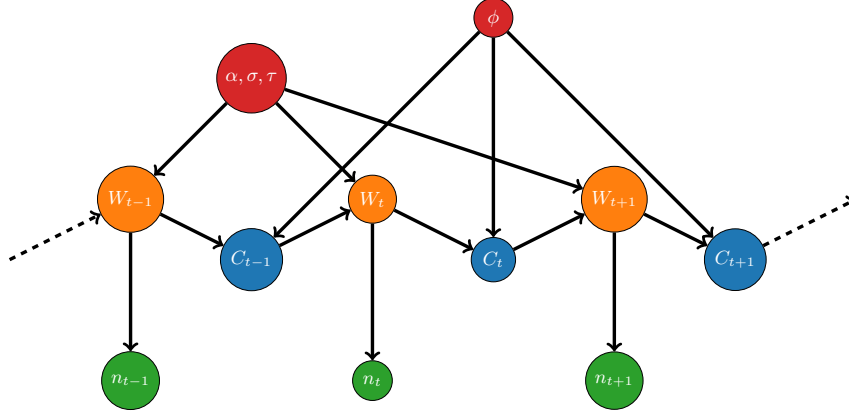
Fig. 1: Graphical representation of the model. The counts n_t are derived from the sociabilities W_t, whose time-evolution depends on the counts C_t and hyperparameters α, σ, τ and ϕ.

assigned a sociability parameter $w_i > 0$. Let $W = \sum_i w_i \delta_{\theta_i}$ be the corresponding random measure (CRM) on $[0, \alpha]$. We assume that W is a generalised gamma completely random measure [27, 1, 5, 31]. That is, $\{(w_i, \theta_i)_{i \geq 1}\}$ are the points of a Poisson point process with mean measure $\nu(w)dw \mathbb{1}_{\theta \leq \alpha}d\theta$ where $\mathbb{1}_A = 1$ if the statement A is true and 0 otherwise, and ν is a Lévy intensity on $(0, \infty)$ defined as

$$\nu(w) = \frac{1}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-\tau w} \tag{1}$$

with hyperparameters $\sigma < 1$ and $\tau > 0$. We write simply $W \sim \mathrm{GG}(\alpha, \sigma, \tau)$ The GGP is a CRM with two interpretable parameters and useful conjugacy properties [31, 10]. Importantly, with this GGP construction, [7, 8] show that this model yields sparse graphs with a power-law degree distribution when $\sigma > 0$. The advantage of using this construction over a standard gamma process [28] is that the parameter σ allows us to control the sparsity properties of the model, and thus fit to networks with different power-law degree distributions.

To each pair of nodes i, j, we assign a number of latent interactions n_{ij}, where

$$n_{ij} \mid w_i, w_j \sim \begin{cases} \mathrm{Poisson}(2w_i w_j) \ i < j, & n_{ji} = n_{ij} \\ \mathrm{Poisson}(w_i w_j) \ \ \ i = j \end{cases} \tag{2}$$

Finally, two nodes are said to be connected if they have at least one interaction; let $z_{ij} = \mathbb{1}_{n_{ij} > 0}$ be the binary variable indicating if two nodes are connected.

3   Dynamic statistical network model

In order to study dynamically evolving networks, we assume that at each time $t = 1, 2, \ldots, T$, we observe a set of interactions between a number of nodes.

This set of interaction is represented by a point process N_t over $[0, \alpha]^2$ as in Equation (3), where α tunes the size of the graphs.

$$N_t = \sum_{i,j} n_{tij} \delta_{(\theta_i, \theta_j)}. \tag{3}$$

Here, n_{tij} is the number of interactions between i and j at time t, and the θ_i are unique node labels as before.

The dynamic point process N_t is obtained as follows. We assume that each node i at time t has a *sociability* parameter $w_{ti} \in \mathbb{R}_+$, that can be thought of as a measure of the node's willingness to interact with other nodes at time t. We consider the associated collection of random measures on $\mathbb{R}_+$, for $t = 1, \ldots, T$

$$W_t = \sum_i w_{ti} \delta_{\theta_i}, \quad t = 1, \ldots, T.$$

We first describe in Section 3.1 the model for the latent interactions. Then we describe in Section 3.2 the model for the time-varying sociability parameters $(W_t)_{t \geq 1}$. The overall probabilistic model is summarised in Figure 1.

3.1   Dynamic network model based on observed interactions

In the dynamic setting, what we observe in practice is often counts of interactions between nodes, e.g. hyperlinks, emails or co-occurrences, rather than a binary indicator of whether there is a connection between them. So for each pair of nodes $i \leq j$, we let $(n_{tij})_{t=1,2,\ldots T, j \geq i}$ be the interaction count between them at time t. We assume that n_{tij} can be modelled as

$$n_{tij} \mid w_{ti}, w_{tj} \sim \begin{cases} \text{Poisson}(2w_{ti}w_{tj}) & i < j, \quad n_{tji} = n_{tij} \\ \text{Poisson}(w_{ti}w_{tj}) & i = j \end{cases} \tag{4}$$

This model can be easily adapted to graphs with directed edges, by modifying the Equation (4) to

$$n_{tij} \sim Poisson(w_{ti}w_{tj})$$

for all $i \neq j$, where n_{tij} now represents the number of interactions from i to j. The resulting inference algorithm essentially remains the same, and from now on we assume we are in the undirected edge setting. As in the static case, we can reconstruct the binary graph by letting $z_{tij} = \mathbb{1}_{(n_{tij} > 0)}$ be the binary variable indicating if nodes i and j are connected at time t, i.e. two nodes are connected at time t if and only if $n_{tij} > 0$. To avoid ambiguity, we say that the number of edges in the graph at time t is $\sum_{i > j} z_{tij}$, rather than counting the number of interactions between pairs of nodes. Marginalizing out the interaction counts n_{tij}, we have for $i \neq j$:

$$\Pr(z_{tij} = 1 \mid (w_{t-k,i}, w_{t-k,j})_{k=0,\ldots,t-1}) = 1 - e^{-2w_{ti}w_{tj}}. \tag{5}$$

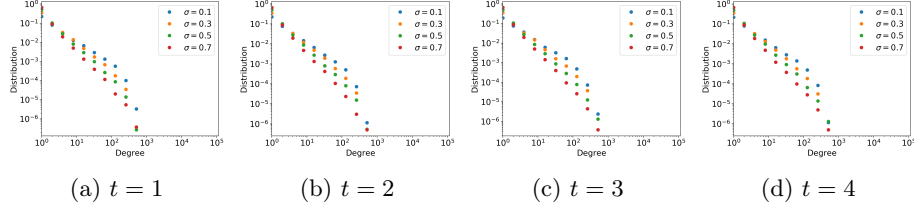(a) $t = 1$ (b) $t = 2$ (c) $t = 3$ (d) $t = 4$

Fig. 2: Degree distributions over time, for a network simulated from the GG model with $T = 4$, $\alpha = 200$, $\tau = 1$, $\phi = 1$ and varying values of σ.

3.2 A dependent generalised gamma process for the sociability parameters

When modelling the sociability parameters in the dynamic setting, we have two goals. Firstly, we want the resulting graphs at each time to be sparse (ideally fitting within the framework of [7]). Secondly, we want sociability parameters to be dependent over time, so that we may model the smooth evolution of the sociabilities of the nodes. To this end, we consider here that the sequence of random measures $(W_t)_{t=1,2,\ldots}$ follows a Markov model, such that W_t is marginally distributed as $\mathrm{GG}(\alpha, \sigma, \tau)$. In order to do this, we build on the generic construction of [37]. A similar model has been derived by [9] for dependent gamma processes (corresponding to the case $\sigma = 0$ here). As in [7], we use the generalised gamma process here because of the flexibility the sparsity parameter σ gives us. In particular, this setup allows us to capture power-law degree distributions, unlike with the gamma process.

For a sequence of additional latent variables $(C_t)_{t=1,2,\ldots}$, we consider a Markov chain $W_t \to C_t \to W_{t+1}$ starting with $W_1 \sim \mathrm{GG}(\alpha, \sigma, \tau)$ that leaves W_t marginally $\mathrm{GG}(\alpha, \sigma, \tau)$. For $t = 1, \ldots, T - 1$, define

$$C_t = \sum_{i=1}^{\infty} c_{ti} \delta_{\theta_i} \quad c_{ti}|W_t \sim \mathrm{Poisson}(\phi w_{ti}) \tag{6}$$

where $\phi > 0$ is a parameter tuning the correlation. Given C_t, the measure W_{t+1} is then constructed as a combination of masses defined by C_t and GG innovation:

$$W_{t+1} = W_{t+1}^* + \sum_{i=1}^{\infty} w_{t+1,i}^* \delta_{\theta_i} \tag{7}$$

with

$$W_{t+1}^* \sim \mathrm{GG}(\alpha, \sigma, \tau + \phi) \tag{8}$$

$$w_{t+1,i}^*|C_t \sim \mathrm{Gamma}(\max(c_{ti} - \sigma, 0), \tau + \phi). \tag{9}$$

By convention, $\mathrm{Gamma}(0, \tau) = \delta_0$ (a point mass at 0). Hence, $w_{t+1,i}^* = 0$ if $c_{ti} = 0$, else $w_{t+1,i}^* > 0$. Because the conditional laws of $W_{t+1}|C_t$ coincide with that of

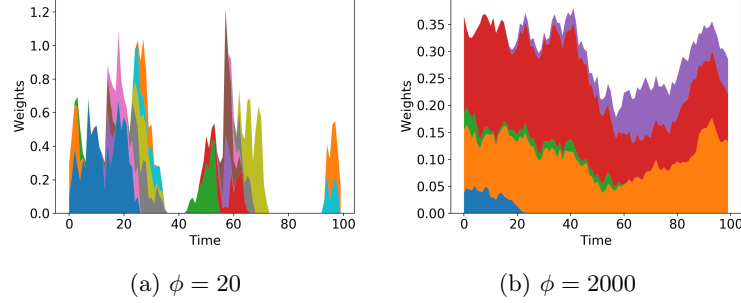(a) $\phi = 20$ (b) $\phi = 2000$

Fig. 3: Evolution of weights over time, for a network simulated from the GG model with $T = 100$, $\alpha = 1$, $\sigma = 0.01$, $\tau = 1$ and (a) $\phi = 20$ (b) $\phi = 2000$. In each case we plot the largest weights, with the respective nodes represented by different colours.

$W_t|C_t$ [38, 23, 24], the construction guarantees that W_{t+1} has the same marginal distribution as W_t, i.e., they are both distributed as $\mathrm{GG}(\alpha, \sigma, \tau)$. Moreover, as proved in Section 1.1 of the Supplementary Material, the conditional mean of W_{t+1} given $W_t = \sum_i w_{ti}\delta_{\theta_i}$ has the form

$$E[W_{t+1}|W_t] = \left(\frac{\tau}{\tau + \phi}\right)^{1-\sigma} E[W_t]$$
$$+ \frac{1}{\tau + \phi} \sum_{i=1}^{\infty} [\phi w_{ti} - \sigma(1 - e^{-\phi w_{ti}})]\delta_{\theta_i} \tag{10}$$

In the gamma process case ($\sigma = 0$), the above expression reduces to

$$E[W_{t+1}|W_t] = \frac{\tau}{\tau + \phi} E[W_t] + \frac{\phi}{\tau + \phi} W_t.$$

3.3 Summary of the model's hyperparameters

The model is parameterised by $(\alpha, \sigma, \tau, \phi)$, where:

- α tunes the overall size of the networks, with a larger value of α corresponding to larger networks.
- σ controls the sparsity and power-law properties of the graph, as will be shown in Section 3.4. In Figure 2 we see that different values of σ give rise to different power-law degree distributions.
- τ induces an exponential tilting of large degrees in the degree distribution.
- ϕ tunes the correlation of the sociabilities of each node over time. As we see in Figure 3, larger values correspond to higher correlation and smoother evolution of the weights.
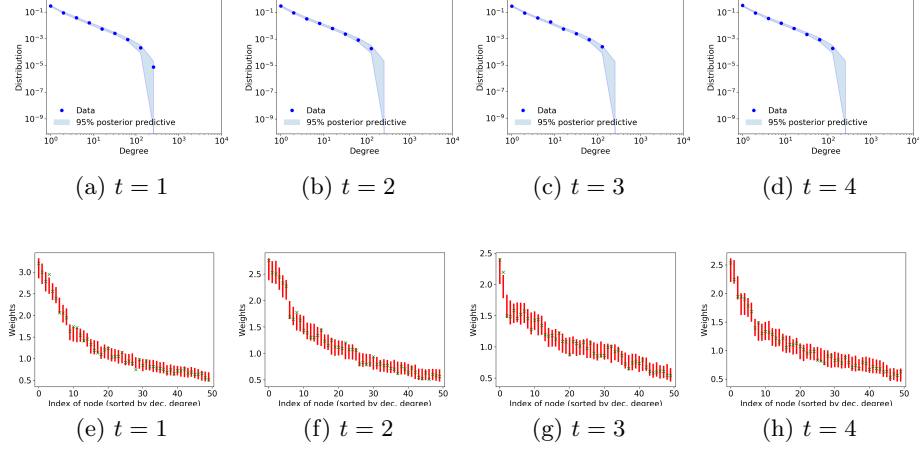
Fig. 4: (Top row) Posterior Predictive Degree Distribution and (bottom row) 95% credible intervals for sociabilites of the nodes with the highest degrees, for a network simulated from the GG model. True weights are represented by a green cross.

3.4 Sparsity and power-law properties of the model

By construction, the interactions at time t, n_{tij}, are drawn from the same (static) model as in [7], using a generalised gamma process for the Lévy intensity, and so applying Proposition 18 in [8] we obtain the following asymptotic properties

Proposition 1. *Let $N_{t,\alpha}$ be the number of active nodes at time t, $N_{t,\alpha}^{(e)} = \sum_{i \leq j} z_{tij}$ be the number of edges and $N_{t,\alpha,j}$ the number of nodes of degree j in the graph at time t, then as α tends to infinity, almost surely, we have for any t: if $\sigma > 0$, $N_{t,\alpha}^{(e)} \asymp N_{t,\alpha}^{2/(1+\sigma)}$, if $\sigma = 0$, $N_{t,\alpha}^{(e)} \asymp N_{t,\alpha}^2 / \log^2(N_{t,\alpha})$ and if $\sigma < 0$, $N_{t,\alpha}^{(e)} \asymp N_{t,\alpha}^2$. Also, almost surely, for any $t \geq 1$ and $j \geq 1$, if $\sigma \in (0,1)$,*

$$\frac{N_{t,\alpha,j}}{N_{t,\alpha}} \to p_j, \quad p_j = \frac{\sigma \Gamma(j - \sigma)}{j! \Gamma(1 - \sigma)},$$

while if $\sigma \leq 0$, $N_{t,\alpha,j}/N_{t,\alpha} \to 0$ for all $j \geq 1$.

Hence the graphs are sparse if $\sigma \geq 0$ and dense if $\sigma < 0$. This distinction stems from the properties of the GGP that we use to define the distribution of the sociabilities, as for $\sigma < 0$, the GGP is finite-activity, and for $\sigma \geq 0$ the GGP is infinite activity. For further details see [7, 8].
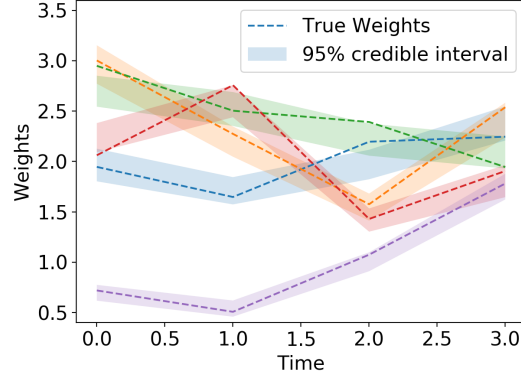
4 Approximate inference

Fig. 5: Evolution of weights of high degree nodes for the network simulated from the GG model. The dotted line shows the true value of the weights in each case.

4.1   Finite-dimensional approximation

Performing exact inference using this model is quite challenging, and we consider instead an approximate inference method, using a finite-dimensional approximation to the GG prior, introduced by [30]. This approximation gives rise to a particularly simple conjugate construction, enabling posterior inference to be performed.

Let $\mathrm{BFRY}(\eta, \tau, \sigma)$ denote a (scaled and exponentially tilted) BFRY random variable[4] on $(0, \infty)$ with probability density function

$$g_{\eta,\tau,\sigma}(w) = \frac{\sigma w^{-1-\sigma} e^{-\tau w} \left(1 - e^{-(\sigma/\eta)^{1/\sigma} w}\right)}{\Gamma(1 - \sigma) \left\{\left(\tau + (\sigma/\eta)^{1/\sigma}\right)^\sigma - \tau^\sigma\right\}}$$

with parameters $\sigma \in (0, 1)$, $\tau > 0$ and $\eta > 0$.

At time 1, consider the finite-dimensional measure

$$W_1 = \sum_{i=1}^{K} w_{1i} \delta_{\theta_i}$$

where $w_{1i} \sim \mathrm{BFRY}(\alpha/K, \tau, \sigma)$ and $K < \infty$ is the truncation level. As shown by [30], for $\sigma \in (0, 1)$

$$W_1 \xrightarrow{d} \mathrm{GG}(\alpha, \sigma, \tau)$$

as the truncation level K tends to infinity. By this we mean that, if $W \sim GG(\alpha, \sigma, \tau)$, then

$$\lim_{K \to \infty} \mathcal{L}_f(W') = \mathcal{L}_f(W)$$

---

[4] The name was coined by [13] after [3].

(a) $t = 2$ (b) $t = 6$ (c) $t = 10$ (d) $t = 12$
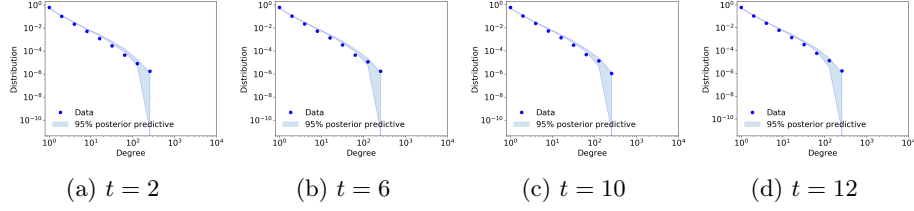
Fig. 6: Posterior Predictive Degree Distribution over time for the Reddit hyperlink network

for an arbitrary measurable and positive f, where $\mathcal{L}_f(W) := \mathbb{E}\left[e^{-W(f)}\right]$ is the Laplace functional of W as defined by [30].

To use this finite approximation with our dynamic model, we consider Poisson latent variables as in Equation (6). The measure W_{t+1} is then constructed as:

$$W_{t+1} = \sum_{i=1}^{K} w_{t+1,i} \delta_{\theta_i}$$
$$w_{t+1,i} | C_t \sim \mathrm{BFRY}\left(\alpha'_t/K, \tau + \phi, \sigma - c_{ti}\right).$$

where

$$\alpha'_t = K(\sigma - c_{ti})(\sigma K/\alpha)^{\frac{c_{ti} - \sigma}{\sigma}}.$$

This construction mirrors that of Section 3.2, with the key difference being that we now obtain a stationary $\mathrm{BFRY}\left(\alpha/K, \tau, \sigma\right)$ distribution for the w_{ti}. We can easily see this by noting that if

$$w_{ti} \sim \mathrm{BFRY}\left(\alpha/K, \tau, \sigma\right)$$
$$c_{ti} | w_{ti} \sim \mathrm{Poisson}(\phi w_{ti})$$

then

$$p(w_{ti}|c_{ti}) \propto w_{ti}^{-1-\sigma+c_{ti}} e^{-(\tau+\phi)w_{ti}} \left(1 - e^{-(\sigma K/\alpha)^{1/\sigma} w_{ti}}\right) \qquad (11)$$

which we then recognise as a $\mathrm{BFRY}\left(\alpha'_t/K, \tau + \phi, \sigma - c_{ti}\right)$ with α'_t as above.

Thus, the use of a finite-dimensional approximation BFRY random variables gives us the simple conjugate construction that we desired. The reason that we introduce this non-standard distribution is that, as far as we know, it is not possible to approximate the generalised gamma process using a finite measure with i.i.d. gamma random weights. In the specific case $\sigma = 0$, we could use the simpler finite approximation using $Gamma(\alpha/K, \tau)$ random variables. However, this would preclude us from modelling networks with power-law degree distributions, as discussed previously.
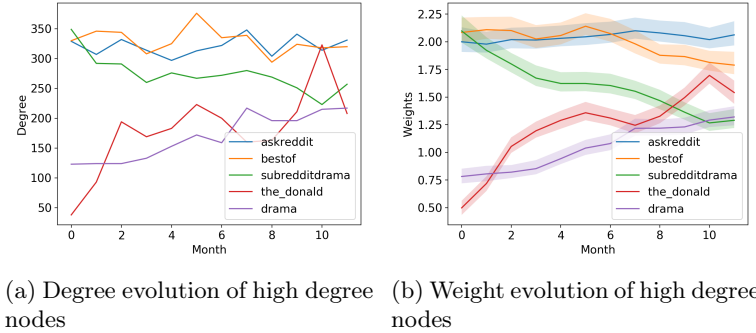
(a) Degree evolution of high degree nodes

(b) Weight evolution of high degree nodes

Fig. 7: Evolution of (a) degrees and (b) weights of high degree nodes for the Reddit hyperlink network

4.2   Posterior Inference Algorithm

In order to perform posterior inference with the approximate method, we use a Gibbs sampler. We introduce auxiliary variables $\{u_{ti}\}_{i=1}^{K}, t = 1, \ldots, T$ following a truncated exponential distribution. The overall sampler is as follows (see the Supplementary Material for full details):

1. Update the weights w_{ti} given the rest using Hamiltonian Monte Carlo (HMC).
2. Update the latent c_{ti} given the rest using Metropolis Hastings (MH).
3. Update the hyperparameters α, σ, ϕ and τ and the latent variables u_{ti} given the rest. We place gamma priors on α, τ and ϕ, and a beta prior on σ.

5   Experiments

In this section, we use our model to study three dynamic networks. The code used for all experiments is available at `https://github.com/ciannaik/SDyNet`.

5.1   Simulated Data

In order to assess the approximate inference scheme, we first consider a synthetic dataset. We simulate a network with $T = 4$, $\alpha = 100$, $\sigma = 0.2$, $\tau = 1$, $\phi = 10$ from the exact model (see Section 3), and then estimate it using our approximate inference scheme described in Section 4. The generated network has $3,096$ nodes and $101,571$ edges. We run an MCMC chain of length $600,000$, with $300,000$ samples discarded as burn-in. We set the truncation threshold to $K = 15,000$.

The approximate inference scheme estimates the model's parameters well. This can be seen in terms of the fit of the posterior predictive degree distribution to the empirical as seen in Figures 4a-4d, and the coverage of the credible intervals for the weights, as we see in Figures 4e-4h for the 50 nodes with the

(a) Weights evolution of the subreddits 'the_donald' and 'sandersforpresident'

(b) Weights evolution of the subreddits 'worldnews' and 'politics'

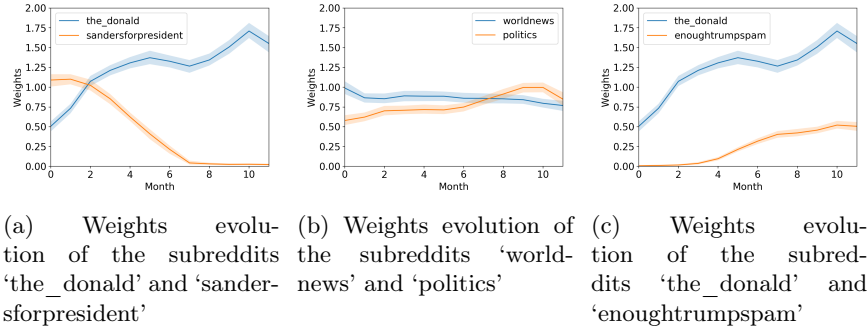(c) Weights evolution of the subreddits 'the_donald' and 'enoughtrumpspam'

Fig. 8: Evolution of the weights, for the Reddit hyperlink network.

highest degree. In Figure 5 we see that the model is also able to capture the weight evolution of the highest degree nodes over time.

In this simulated data experiment, as well as when looking at the real data experiments, we need to choose the truncation level K. The trade-off here is between a better approximation for larger K, but slower mixing and a longer running time of the MCMC algorithm. We examine the effect that our choice of K has in the Supplementary Material.

5.2   Real Data

We illustrate the use of our model on two more dynamic network datasets: the Reddit Hyperlink Network [29][5] and the Reuters Terror dataset[6].

Reddit Hyperlink Network The Reddit hyperlink network represents hyperlink connections between subreddits (communities on Reddit) over a period of $T = 12$ consecutive months in 2016. Nodes are subreddits (communities) and the symmetric edges represent hyperlinks originating in a post in one community and linking to a post in another community. The network has $N = 28,810$ nodes and $388,574$ interactions. The observations here are hyperlinks between the pair of subreddits i, j at time t. The dataset has been made symmetric by placing an edge between nodes i and j if there is a hyperlink between them in either direction. We also assume that there are no loops in the network, that is $n_{tij} = 0$ for $i = j$. We run the Gibbs sampler $400,000$ samples, with the first $200,000$ discarded as burn in. In this case, we choose a truncation level of $K = 40,000$.

From Figure 6 we see that our model is capturing the empirical degree distribution well. Furthermore, in Figure 7 we see that the model is able to capture the evolution of weights associated with each subreddit in a fashion that agrees

---

[5] https://snap.stanford.edu/data/soc-RedditHyperlinks.html

[6] http://vlado.fmf.uni-lj.si/pub/networks/data/CRA/terror.htm

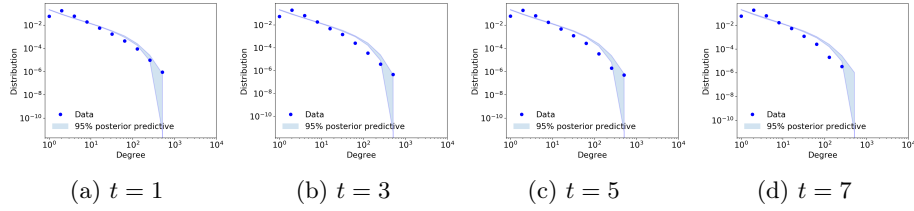(a) $t = 1$ (b) $t = 3$ (c) $t = 5$ (d) $t = 7$

Fig. 9: Posterior Predictive Degree Distribution for the Reuters terror dataset

with the observed frequency of interactions. The high degree nodes here are interpretable as either communities with a very large number of followers - such as "askreddit" or "bestof", or communities which frequently link to others - such as "drama" or "subredditdrama".

In particular, we see in Figure 8a that the weights of the controversial but popular political subreddit "The Donald" increases to a peak in November, corresponding to the 2016 U.S. presidential election. Conversely, the weights of the subreddit "Sandersforpresident" decrease as the year goes on, corresponding to the end of Senator Bernie Sanders' presidential campaign, a trend that again agrees with the evolution of the corresponding observed degrees.

Reuters Terror Dataset The final dataset we consider is the Reuters terror news network dataset. It is based on all stories released during $T = 7$ consecutive weeks (the original data was day-by-day, but was shortened and collated for our purposes) by the Reuters news agency concerning the 09/11/01 attack on the U.S.. Nodes are words and edges represent co-occurence of words in a sentence, in news. The network has $N = 13,332$ nodes (different words) and $473,382$ interactions. The observations here are the frequency of co-occurence between the pair of words i, j at time t. We assume that there are no loops in the network, that is $n_{tij} = 0$ for $i = j$. We run the Gibbs sampler with $200,000$ samples, with the first $100,000$ discarded as burn in. In this case, we choose a truncation level of $K = 20,000$.

Figure 9 suggests that the empirical degree distribution does not follow a power law distribution. In particular there are significantly fewer nodes of degree one than we would expect. Our model therefore provides a moderate fit to the empirical degree distribution. The model is however able to capture the evolution of the popularity of the different words, as shown in Figure 10a. For example, the weights of the words "plane" and "attack" decrease over time after 9/11, while the words "letter" and "anthrax" show a peak a few weeks after the attack. These correspond to the anthrax attacks that occurred over several weeks starting a week after 9/11.

Due to the empirical degree distribution not following a power-law, the estimated value of σ is very close to 0. This causes a slow convergence of the MCMC algorithm, which we see in the Supplementary Material.

(a) Weights evolution of the words 'plane' and 'attack'

(b) Weights evolution of the words 'al quaeda', 'taliban' and 'bin laden'

(c) Weights evolution of the words 'anthrax' and 'letter'
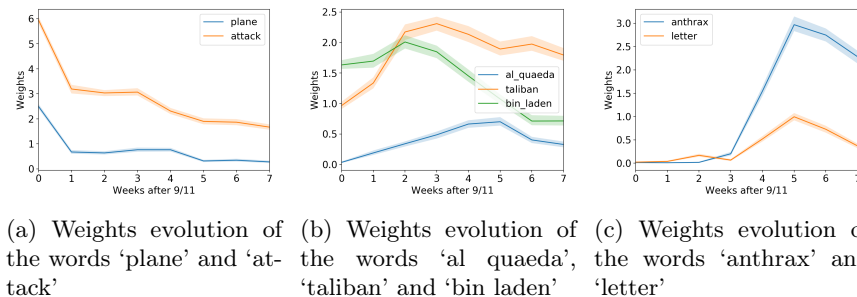
Fig. 10: Evolution of the weights, for the Reuters terror dataset

6   Discussion and Extensions

A lot of work has been done on modelling dynamic networks; we restrict ourselves to focussing on Bayesian approaches. Much of this work has centred around extending static models. For example, [44] and [14] extend the stochastic block model to the dynamic setting by allowing for parameters that evolve over time. There has also been work on extending the mixed membership stochastic blockmodel [16, 21, 43], the infinite relational model [35] and the latent feature relational model [15, 19, 26]. The problem inherent in these models is that they lead to networks which are dense almost surely [2, 22], a property considered unrealistic for many real-world networks [36] such as social and internet networks.

In order to build models for sparse dynamic networks, [17] build on the framework of edge-exchangeable networks [6, 11, 42], in which graphs are constructed based on an infinitely exchangeable sequence of edges. As in our case, this framework allows for sparse networks with power-law degree distributions. This work, along with others in this framework [35, 18], utilises the mixture of Dirichlet network distributions (MDND) to introduce struture in the networks.

Conversely, our works builds on a different notion of exchangeability [25, 7]. Within this framework, [32] use mutually-exciting Hawkes processes to model temporal interaction data. The difference between our work and theirs is that the sociabilities of the nodes are constant throughout time, with the time evolving element driven by previous interactions via the Hawkes process. Their work also builds on that of [39], incorporating community structure to the network. Exploring how communities appear, evolve and merge could have many practical uses. Thus, expanding our model to capture both evolving node popularity and dynamically changing community structure could be a useful extension.

Furthermore, our model assumes that the time between observations of the network is constant. If this is not the case, it may be helpful to use a continuous-time version of our model. This could be done by considering a birth-death process for the interactions between nodes, where each interaction has a certain lifetime distribution. The continuously evolving node sociabilities could then by described by the Dawson-Watanabe superprocess [41, 12].

The goal of our work is to extend the Bayesian nonparametric framework for network modelling of [7] to the dynamic setting. Since exact inference is intractable in this regime, we also introduce an approximate inference method. Further work is needed to apply this framework in scenarios such as dynamic community detection and link prediction for unseen time steps. We leave this to future work.

Acknowledgements

Bibliography

[1] Aalen, O.: Modelling heterogeneity in survival analysis by the compound Poisson distribution. The Annals of Applied Probability pp. 951–972 (1992)

[2] Aldous, D.J.: Representations for partially exchangeable arrays of random variables. Journal of Multivariate Analysis **11**(4), 581–598 (1981)

[3] Bertoin, J., Fujita, T., Roynette, B., Yor, M.: On a particular class of self-decomposable random variables : the durations of Bessel excursions straddling independent exponential times. Probability and Mathematical Statistics **26**(2), 315–366 (2006)

[4] Borgs, C., Chayes, J., Cohn, H., Holden, N.: Sparse exchangeable graphs and their limits via graphon processes. The Journal of Machine Learning Research **18**(1), 1–71 (2018)

[5] Brix, A.: Generalized gamma measures and shot-noise Cox processes. Advances in Applied Probability **31**(4), 929–953 (1999)

[6] Cai, D., Campbell, T., Broderick, T.: Edge-exchangeable graphs and sparsity. In: Advances in Neural Information Processing Systems. pp. 4249–4257 (2016)

[7] Caron, F., Fox, E.: Sparse graphs using exchangeable random measures. Journal of the Royal Statistical Society B **79**, 1–44 (2017), part 5

[8] Caron, F., Panero, F., Rousseau, J.: On sparsity, power-law and clustering properties of graphex processes. arXiv pp. arXiv–1708 (2017)

[9] Caron, F., Teh, Y.W.: Bayesian nonparametric models for ranked data. In: NIPS (2012)

[10] Caron, F., Teh, Y., Murphy, T.: Bayesian nonparametric plackett-luce models for the analysis of preferences for college degree programmes. The Annals of Applied Statistics **8**(2), 1145–1181 (2014)

[11] Crane, H., Dempsey, W.: Edge exchangeable models for interaction networks. Journal of the American Statistical Association **113**(523), 1311–1326 (2018)

[12] Dawson, D.A.: Stochastic evolution equations and related measure processes. Journal of Multivariate Analysis **5**(1), 1–52 (1975)

[13] Devroye, L., James, L.F.: On simulation and properties of the stable law. Statistical Methods and Applications **23**(3), 307–343 (2014)

[14] Durante, D., Dunson, D.: Bayesian logistic gaussian process models for dynamic networks. In: AISTATS. pp. 194–201 (2014)

[15] Foulds, J., DuBois, C., Asuncion, A., Butts, C., Smyth, P.: A dynamic relational infinite feature model for longitudinal social networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 287–295 (2011)

[16] Fu, W., Song, L., Xing, E.P.: Dynamic mixed membership blockmodel for evolving networks. In: Proceedings of the 26th annual international conference on machine learning. pp. 329–336 (2009)

[17] Ghalebi, E., Mahyar, H., Grosu, R., Taylor, G.W., Williamson, S.A.: A nonparametric bayesian model for sparse temporal multigraphs. CoRR (2019)

[18] Ghalebi, E., Mirzasoleiman, B., Grosu, R., Leskovec, J.: Dynamic network model from partial observations. In: Advances in Neural Information Processing Systems. pp. 9862–9872 (2018)

[19] Heaukulani, C., Ghahramani, Z.: Dynamic probabilistic models for latent feature propagation in social networks. In: International Conference on Machine Learning. pp. 275–283 (2013)

[20] Herlau, T., Schmidt, M.N., Mørup, M.: Completely random measures for modelling block-structured sparse networks. Advances in Neural Information Processing Systems **29** (2016)

[21] Ho, Q., Song, L., Xing, E.: Evolving cluster mixed-membership blockmodel for time-evolving networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. pp. 342–350 (2011)

[22] Hoover, D.N.: Relations on probability spaces and arrays of random variables. Preprint, Institute for Advanced Study, Princeton, NJ (1979)

[23] James, L.F.: Poisson process partition calculus with applications to exchangeable models and bayesian nonparametrics. arXiv preprint math/0205093 (2002)

[24] James, L.F., Lijoi, A., Prünster, I.: Posterior analysis for normalized random measures with independent increments. Scandinavian Journal of Statistics **36**(1), 76–97 (2009)

[25] Kallenberg, O.: Exchangeable random measures in the plane. Journal of Theoretical Probability **3**(1), 81–136 (1990)

[26] Kim, M., Leskovec, J.: Nonparametric multi-group membership model for dynamic networks. In: Advances in neural information processing systems. pp. 1385–1393 (2013)

[27] Kingman, J.: Completely random measures. Pacific Journal of Mathematics **21**(1), 59–78 (1967)

[28] Klenke, A.: Probability theory: a comprehensive course. Springer Science & Business Media (2013)

[29] Kumar, S., Hamilton, W.L., Leskovec, J., Jurafsky, D.: Community interaction and conflict on the web. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web. pp. 933–943. International World Wide Web Conferences Steering Committee (2018)

[30] Lee, J., James, L.F., Choi, S.: Finite-dimensional bfry priors and variational bayesian inference for power law models. In: Advances in Neural Information Processing Systems. pp. 3162–3170 (2016)

[31] Lijoi, A., Mena, R.H., Prünster, I.: Controlling the reinforcement in Bayesian non-parametric mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **69**(4), 715–740 (2007)

[32] Miscouridou, X., Caron, F., Teh, Y.W.: Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data. In: Advances in Neural Information Processing Systems. pp. 2343–2352 (2018)

[33] Naik, C., Caron, F., Rousseau, J.: Sparse networks with core-periphery structure. Electronic Journal of Statistics **15**(1), 1814–1868 (2021)

[34] Newman, M.: Networks: an introduction. OUP Oxford (2009)

[35] Ng, Y.C., Silva, R.: A dynamic edge exchangeable model for sparse temporal networks. arXiv preprint arXiv:1710.04008 (2017)

[36] Orbanz, P., Roy, D.M.: Bayesian models of graphs, arrays and other exchangeable random structures. IEEE Trans. Pattern Anal. Mach. Intelligence (PAMI) **37**(2), 437–461 (2015)

[37] Pitt, M.K., Walker, S.G.: Constructing stationary time series models using auxiliary variables with applications. Journal of the American Statistical Association **100**(470), 554–564 (2005)

[38] Prünster, I.: Random probability measures derived from increasing additive processes and their application to Bayesian statistics. Ph.D. thesis, University of Pavia (2002)

[39] Todeschini, A., Miscouridou, X., Caron, F.: Exchangeable random measures for sparse and modular graphs with overlapping communities. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **82**(2), 487–520 (2020)

[40] Veitch, V., Roy, D.M.: The class of random graphs arising from exchangeable random measures. arXiv preprint arXiv:1512.03099 (2015)

[41] Watanabe, S.: A limit theorem of branching processes and continuous state branching processes. Journal of Mathematics of Kyoto University **8**(1), 141–167 (1968)

[42] Williamson, S.A.: Nonparametric network models for link prediction. The Journal of Machine Learning Research **17**(1), 7102–7121 (2016)

[43] Xing, E.P., Fu, W., Song, L., et al.: A state-space mixed membership blockmodel for dynamic network tomography. The Annals of Applied Statistics **4**(2), 535–566 (2010)

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