Thinned random measures for sparse graphs with overlapping communities

Anonymous Author(s)
Affiliation
Address
email

Abstract

Network models for exchangeable arrays, including most stochastic block models, generate dense graphs with a limited ability to capture many characteristics of real-world social and biological networks. A class of models based on completely random measures like the generalized gamma process (GGP) have recently addressed some of these limitations. We propose a framework for thinning edges from realizations of GGP random graphs that models observed links by both nodes' overall propensity to interact, and on the similarity of node memberships within a large set of latent communities. Our formulation allows us to learn the number of communities from data, and enables efficient Monte Carlo methods that scale linearly with the number of observed edges, and thus (unlike dense block models) sub-quadratically with the number of entities or nodes. We compare to alternative models for both dense and sparse networks, and demonstrate effective recovery of latent community structure for real-world networks with thousands of nodes.

1 Introduction

Given observations of (often binary) relationships Y_{ij} between pairs of nodes or entities (i, j), many relational models [1] seek to uncover an underlying set of communities. Classic stochastic blockmodels [2] generalize mixture models for clustering non-relational data by assigning each entity to one of K communities (clusters). The *infinite relational model* (IRM) [3] instead uses a Dirichlet process prior [4] to partition entities into single communities. While the IRM allows the number of communities to be inferred from data, later work has shown that real-world social networks are better captured by models which allow individuals to participate in multiple communities [5], including applications of the *hierarchical Dirichlet process* (HDP) [6] to relational data [7].

There is an extensive literature on descriptive statistics of biological and social networks [8, 9], including degree distributions, path distances and "small world" phenomena [10], community structures and modularity, and notions of centrality and causality. In particular, sparsity is a ubiquitous phenomenon in real-world networks [8, 9]: as network size grows, the number of edges grows more slowly than the quadratic number of node pairs. However, the IRM and HDP relational models (and a large literature of related models [1]) generate *dense* graphs where the number of edges scales quadratically with the number of nodes. In fact, a classic representation theorem [11] shows that any generative model which samples edges from independent Bernoulli distributions, given latent node-specific parameters, generates dense graphs; most existing probabilistic network models have this form.

By representing graphs as a latent process (a completely random measure), Caron and Fox [12] showed that it is possible to create generative models that capture the sparsity of real-world networks. Related models, including certain infinite limits of graphs called *graphons*, have been studied by several authors [13, 14, 15]. However, these models mostly produce homogeneous graphs with sparsity and heavy-tailed degree distributions, but lacking the community structure of real networks.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Two notable exceptions are work by Herlau et al. [16] and Todeschini et al. [17] that augment the random-measure models of [12] with variables encoding latent community structure.

In this paper, we propose a novel random graph model that

2 Background: Stochastic Blockmodels for Dense and Sparse Networks

An undirected binary network with N nodes and E edges may be represented by an $N \times N$ binary adjacency matrix Y. $Y_{ij} = Y_{ji} = 1$ if there is an edge (link) between nodes $i \neq j$, otherwise $Y_{ij} = 0$.

2.1 Mixed Membership Stochastic Blockmodels

Stochastic blockmodels (SBMs) [2] assume that each node belongs to one of K latent communities, where the probability of an edge depends on how strongly their communities are connected. Let $c_i \in \{1, \ldots, K\}$ indicate the community of node i, where $c_i \overset{\text{ind}}{\sim} \text{Cat}(\beta)$ and β_k is the frequency of community k. Edges are then sampled independently as $Y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\eta_{c_i c_j})$, where $\eta_{k\ell} = \eta_{\ell k}$ is the probability of an edge between a node in community k and a node in community ℓ. These interaction probabilities are often assigned conjugate beta priors, $\eta_{k\ell} \sim \text{Beta}(\tau_a, \tau_b)$.

Mixed membership stochastic blockmodels (MMSBs) [5] extend block models to allow nodes to be members of multiple communities. Let $\pi_i = (\pi_{i1}, \ldots, \pi_{iK})$ denote a K-dimensional probability vector representing the strength of affiliation of node i to each of K communities. For every pair of nodes (i, j), the communities governing their interaction are sampled as $c_{ij} \sim \text{Cat}(\pi_i)$, $c_{ji} \sim \text{Cat}(\pi_j)$. Then like standard SBMs, edges are sampled independently as $Y_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\eta_{c_{ij} c_{ji}})$. Community memberships are typically assigned a hierarchical prior, such as

$$\pi_i \mid \beta \overset{\text{ind}}{\sim} \text{Dirichlet}(\zeta\beta_1, \ldots, \zeta\beta_K), \qquad \beta \sim \text{Dirichlet}\left(\frac{\gamma}{K}, \ldots, \frac{\gamma}{K}\right). \tag{1}$$

By setting $K \gg \gamma$, the model allows network-specific learning of the number of communities by favoring a sparse community frequency vector β. The hierarchical Dirichlet process [6, 7] is the limit of this prior as K approaches infinity [18].

2.2 Sparse Network Models via Completely Random Measures

For any fixed community frequencies β and interaction probabilities η, the SBM and MMSB may only generate *dense* graphs where the number of edges scales quadratically with the number of nodes [19]. In contrast, many real-world networks appear to be sparse [9]. Heuristics are often used to fit (mixed) SBMs to large but sparse networks, such as fixing (rather than learning) $\eta_{k\ell} = \varepsilon \approx 0$ for $k \neq \ell$; Kim et al. [7] fix $\varepsilon = e^{-30}$. We seek to avoid such heuristics by building models that simultaneously capture sparsity and (mixed) membership in latent communities.

Completely Random Measures. By representing the graph as a point process on the plane, Caron and Fox [12] showed that it is possible to generate sparse graphs by appropriately choosing the mean measure of the point process. According to their model, for $i \neq j$,

$$Y_{ij} \mid w_i, w_j \overset{\text{ind}}{\sim} \text{Bernoulli}\left(1 - \exp\left\{-2 w_i w_j\right\}\right), \tag{2}$$

where $w_i > 0$ represents the *sociability* of node i: nodes with higher w_i have higher probability to interact with other nodes, and hence greater expected degree. Node sociabilities are generated as the jumps of a *completely random measure* (CRM) [20] at real-valued locations ℓ_i uniformly distributed on an interval $[0, \alpha]$,

$$W_\alpha = \{w_i : \ell_i \in [0, \alpha]\}. \tag{3}$$

Here $\alpha > 0$ controls the (random) number of nodes N_α in the network by determining the size of the interval in which jumps are included. Depending on the distribution of the jumps, the model can capture both sparse and dense graphs. Intuitively, as detailed in Sec. 5.1 of [12], for sparse networks the distribution of the jumps needs to place almost all of its mass near zero.

This model requires CRMs for which the sum $\bar{W}_\alpha = \sum_{i:\ell_i \in [0,\alpha]} w_i$ of the jumps in $[0, \alpha]$ is finite. The observed network is then generated via a binary projection of an underlying (directed) multigraph,

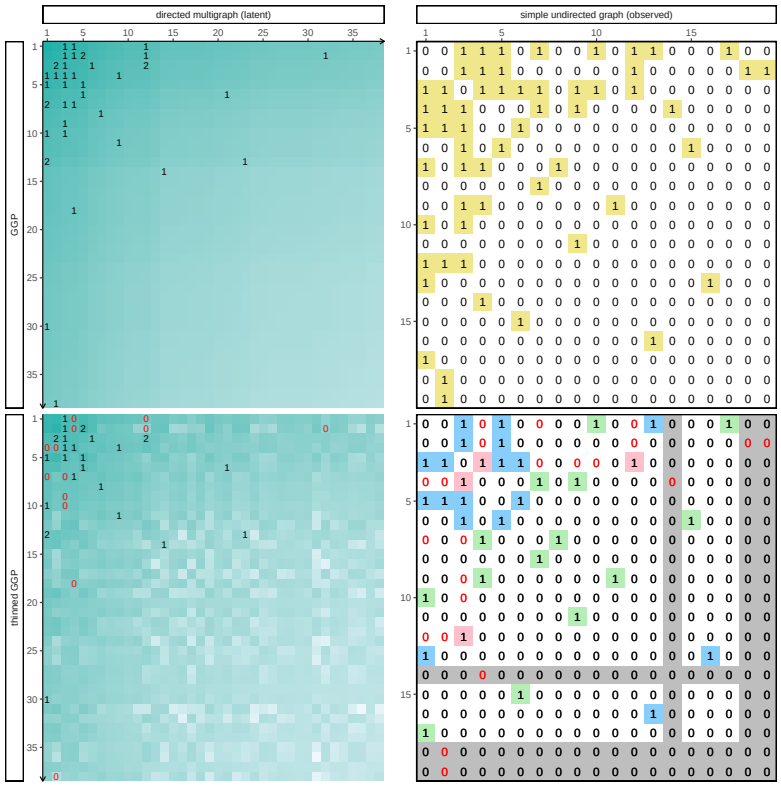


Figure 1: Generation of an adjacency matrix from the GGP model of Caron and Fox [12] (top) and via thinning via our proposed model with $K = 3$ communities (bottom). Rows and columns correspond to nodes in all plots, ordered in decreasing order of sociability w_i. *Left:* Latent directed multigraphs underlying the observed graphs on their right. The color of a cell is proportional to the Poisson rate: $2w_iw_j$ for the GGP, and $2w_iw_j \sum_{k=1}^{K} \pi_{ik}\pi_{jk}$ for the thinned GGP. Red zeros indicate thinned edges. *Right:* Binary adjacency matrix of the observed graph for the GGP (top) and thinned GGP (bottom) models. For the thinned GGP, edges are colored according to the community that generated it. Gray cells mark nodes whose edges in the latent multigraph were all thinned, and are thus observed under the GGP but unobserved under the thinned GGP. GGP hyperparameters were set as $\sigma = 0.1$, $\tau = 1$, $\alpha = 10$, and community memberships were sampled given $\beta = (1/3, 1/3, 1/3)$, $\zeta = 1$.

where the total number of edges n_{ij} from node i to node j is independently distributed as

$$n_{ij} \mid w_i, w_j \overset{\text{ind}}{\sim} \text{Poisson}(w_iw_j). \tag{4}$$

We then set $Y_{ij} = \mathbb{1}(n_{ij} + n_{ji} \geq 1)$ for nodes $i \neq j$. Because the sum of independent Poisson random variables is Poisson, Eq. (4) implies $P(n_{ij} + n_{ji} = 0) = \exp\{-2w_iw_j\}$, from which Eq. (2) follows. The number N of nodes in the observed network then equals the number N_α of nodes that have at least one edge in the underlying multigraph. This construction of the binary matrix from a multigraph [12] is visualized in Fig. 1.

The sum-property of the Poisson distribution also implies that, given $\bar{W}_\alpha$, the total number of edges $\bar{D}_\alpha$ in the multigraph has a Poisson$(\bar{W}_\alpha^2)$ distribution. Since n_{ij} has a high probability of being 0 for most node pairs (i, j), it is more efficient to first sample $\bar{D}_\alpha$, and then independently assign each edge to a pair of nodes based on their sociabilities. In more detail, Eq. (4) is equivalent to

$$\bar{D}_\alpha \mid \bar{W}_\alpha \sim \text{Poisson}(\bar{W}_\alpha^2), \quad P(x_{ev} = i \mid W_\alpha) = \frac{w_i}{\bar{W}_\alpha}, \quad n_{ij} = \sum_{e=1}^{\bar{D}_\alpha} \mathbb{1}(x_{e1} = i)\mathbb{1}(x_{e2} = j), \tag{5}$$

for $v = 1, 2$ and $e = 1, \ldots, \bar{D}_\alpha$. Caron and Fox [12] propose the *generalized gamma process* (GGP) [21, 22, 23] as a flexible but tractable CRM for W_α, with parameters $\tau \in (0, \infty)$, and $\sigma \in (-\infty, 0]$ for densely graphs or $\sigma \in (0, 1)$ for sparse graphs. We summarize this generative process via a directed graphical model in Fig. 2, and a simulated network is visualized in Fig. 3.

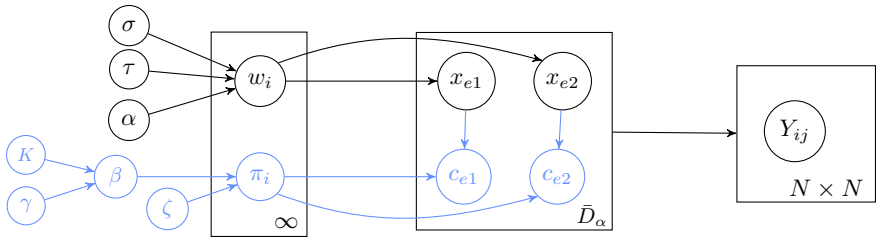


Figure 2: Directed graphical model representing the GGP random graphs of Caron and Fox [12] (black), and the additional variables (blue) in our thinned GGP. Settings of the GGP hyperparameters that induce sparse graphs ($\sigma > 0$) lead to an infinite number of *potential* nodes, with sociabilities w_i and community memberships π_i.

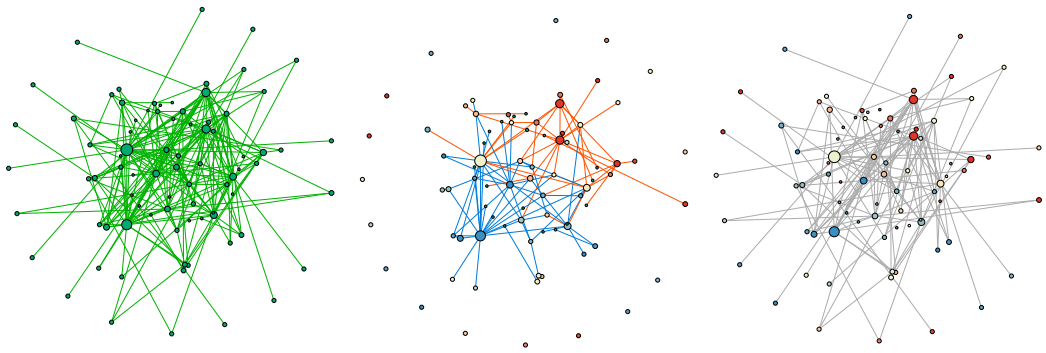
Sparse Block Models. A limitation of the framework described above is that it does not model the community (block) structure of the network – a well-recognized feature of complex networks. Herlau et al. [16] generalized the approach in [12] to accommodate both sparse and dense networks with community structure. More specifically, they introduce a latent assignment indicator, $c_i \in \{1, \ldots, K\}$, to indicate the assignment of node i to one of K communities. A bivariate CRM incorporates both the sociability weights and a set of parameters, say $\eta_{c_i c_j}$, capturing the interaction strength between two communities c_i and c_j in the underlying multigraph. As a result of their formulation, the total number of edges n_{ij} from node i to node j are independently distributed as $n_{ij}|c_i, c_j, w_i, w_j \overset{\text{ind}}{\sim} \text{Poisson}(\eta_{c_i c_j}\, w_i\, w_j)$, and the likelihood (2) of the observed nodes is modified as $Y_{ij} \mid c_i, c_j, w_i, w_j \overset{\text{ind}}{\sim} \text{Bernoulli}(1 - e^{-2\eta_{c_i c_j} w_i w_j})$.

Sparse Mixed Membership. The model proposed by Herlau et al. [16] accommodates sparse and dense networks with community structure but does not model overlapping community structures, i.e. their formulation results in a network partitioned into disjoint communities. In follow-up work, Todeschini et al. [17] extended the CRM-based framework discussed above by associating a vector $(w_{i1}, \ldots, w_{iK})$ of sociabilities to each node i, to represent different levels of affiliation of a node to the latent communities $k = 1, \ldots, K$. A node may have high levels of affiliation to more than a community, leading to the formation of edges across multiple communities. The vectors of nodes' sociabilities are distributed according to a compound CRM [24], specifically in their implementation they use a compound generalized gamma process (CGGP). The likelihood function is modified accordingly by specifying $Y_{ij} \mid \mathbf{w}_i \mathbf{w}_j \overset{\text{ind}}{\sim} \text{Bernoulli}(1 - \exp\{-2 \sum_{k=1}^{K} w_{ik} w_{jk}\})$. The latent community weights are further modeled as $w_{ik} = w_{i0} \beta_{ik}$, where w_{i0} is defined via a generalized gamma process and $\beta_{ik} \overset{\text{ind}}{\sim} \text{Gamma}(a_k, b_k)$. Thus, as [17] point out, the model exploits an (unconstrained) non-negative matrix factorization to define the Bernoulli probability link. In the Appendix, we show how this lack of constraints could hamper the ability of the model to recover the true community structure of the network. On the contrary, the regularization approaches to non-matrix factorization should improve identification of underlying structures and identifiability of the network parameters [25].

3 The thinned CRM model

In this work, we explore introducing overlapping community memberships in the innovative GGP model by Caron and Fox [12] as vectors of probabilities sampled from a hierarchical Dirichlet distribution. Compared to the formulation in Todeschini et al. [17], our model enables learning the number of communities from the data. Moreover, using probability vectors rather than unconstrained non-negative values to model community memberships, our model provides a regularized approach to inference in the GGP model, which appears to provide an increased accuracy in community detection in simulations.

Let a node i have both a sociability parameter w_i from the GGP as in (3) and a vector of probabilities $\pi_i = (\pi_{i1}, \ldots, \pi_{iK})$ drawn from a hierarchical Dirichlet distribution as in (1). Moreover, let the number of *potential* edges between nodes i and j in the latent multigraph depend only on their sociabilities, as in (4). For each of their n_{ij} potential edges, nodes i and j are each assigned



(a) Potential edges (b) Observed edges (c) Discarded edges

Figure 3: Visualization of networks simulated via a (thinned) GGP with $\alpha = 15, \sigma = 0.2, \tau = 1$. (a) Before thinning, *potential* edges are proposed according to nodes' sociabilities as in [12]. (b) After thinning, *observed* edges have colors (blue/red) corresponding to their assignment to $K = 2$ communities. (c) The *discarded* edges (gray) are those for which the connected nodes were assigned to different communities. In both (b) and (c), node colors represent their true community memberships, where lighter colors indicate more balanced memberships. Node sizes are proportional to betweenness centrality, and layout is determined by Gephi's Force Atlas 2 [26].

to a community according to their respective membership probabilities: only if the assignments correspond to the same community, the edge is retained, otherwise it is *thinned* (i.e., discarded). See Figures 1 and 3 for an illustration. In formulas, for every edge $e = 1, \dots, \bar{D}_\alpha$ in the multigraph, the number of edges between a pair of nodes i, j that is retained in the multigraph $\dot{n}_{ij}$ and the existence of an observed edge Y_{ij} in the undirected graph is determined by

$$c_{e1} \mid x_{e1}, (\pi_1, \pi_2, \dots) \overset{\text{ind}}{\sim} \text{Cat}(\pi_{x_{e1}}), \qquad c_{e2} \mid x_{e2}, (\pi_1, \pi_2, \dots) \overset{\text{ind}}{\sim} \text{Cat}(\pi_{x_{e2}}), \tag{6}$$

$$\dot{n}_{ij} = \sum_{e=1}^{\bar{D}_\alpha} \mathbb{1}(x_{e1} = i, x_{e2} = j, c_{e1} = c_{e2}), \qquad Y_{ij} = \mathbb{1}\left(\dot{n}_{ij} + \dot{n}_{ji} \geq 1\right), \tag{7}$$

where $\dot{n}_{ij}$ (full dot) indicates the number of (directed) edges from i to j in the latent multigraph that are retained, and x_{ev} are indicators in $\{1, 2, \dots\}$ as in (5), $v = 1, 2$. According to (7), we interpret an observed simple, undirected, graph as obtained from a binary projection of the possibly multiple retained edges generated in the latent multigraph. The probability of two nodes being assigned to the same community is $P(c_{ei} = c_{ej}) = \sum_{k=1}^{K} \pi_{ik}\pi_{jk}$ after marginalizing across communities. Therefore, $\dot{n}_{ij} \mid w_i, w_j, \pi_i, \pi_j \overset{\text{ind}}{\sim} \text{Poisson}(w_i w_j \sum_{k=1}^{K} \pi_{ik}\pi_{jk})$, marginalizing over the latent multigraph. Thus, the likelihood of the observed network is obtained obtained as

$$Y_{ij} \mid w_i, w_j \overset{\text{ind}}{\sim} \text{Bernoulli}\left(1 - \exp\left\{-2 w_i w_j \sum_{k=1}^{K} \pi_{ik}\pi_{jk}\right\}\right), \tag{8}$$

where we see that, differently than in the MMSB or in the GGP framework, our model favors edges between nodes that have *both* large sociabilities *and* similar community memberships. The model is summarized graphically in Figure 2 and is illustrated in Figures 4c, 4d and 1.

4 Posterior Inference

The formulation of the model as thinning of a multigraph sampled only according to the sociability parameters of the nodes (eq. (5) and (7)) leads to posterior distributions of $(w_1, w_2, \dots)$ and $(\pi_1, \pi_2, \dots)$ that are conditionally independent given the total number of proposed nodes n_{ij} between any two edges i and j in the multigraph. This characteristic of our model allows us to apply Theorem 6 in [12] to derive the posterior of nodes' sociabilities, provided we can condition on the (latent) multigraph. Therefore, we implement a Gibbs sampling strategy for posterior inference, where we sample both thinned and retained edges in the latent multigraph as well as their community assignments to allow for efficient updates of the variables of interest, $(w_1, w_2, \dots)$ and $(\pi_1, \pi_2, \dots)$.

Some steps in the sampler are straightforwardly derived, considering that we can borrow the approach detailed in [12] to update nodes' sociabilities and GGP hyperparameters (see the Appendix for an outline of the sampling scheme). However, the implementation of the sampler requires some additional careful development – which we detail below – in sampling the latent multigraph, due to the latent thinning process. These aspects exemplify some of the computational advantages that are specific to the sparse, as opposed to the dense block model approach. In the multigraph, let $\mathring{n}_{ij} = n_{ij} - \dot{n}_{ij}$ (empty dot) indicate the number of edges from i to j that are thinned due to mismatched community assignments. Here is the strategy adopted to sample the latent multigraph:

Sampling of $\dot{\mathbf{n}}_{\mathbf{ij}}$: The existence of retained edges between nodes i and j in the latent multigraph implies an edge in the observed graph according to (7). Thus, rather than considering each pair i, j, we only need to sample $\dot{n}_{ij}$ when $Y_{ij} = 1$, and so only as many as there are observed edges in the graph, and we can set the remaining ones to zero. Since there must be at least one latent edge when $Y_{ij} = 1$,

$$\dot{n}_{ij} \mid Y_{ij} = 1, w_i, w_j, \pi_i, \pi_j \sim \text{zero-truncated Poisson}(2w_i w_k \sum_{k=1}^{K} \pi_{ik}\pi_{jk}).$$

For retained edges we need to sample a single community assignment, since by definition these are the edges whose nodes have been assigned to the same communities. We can sample the community assignment of a retained edge between i and j easily as a draw from $\text{Cat}(\pi_{i1}\pi_{j1}, \ldots, \pi_{iK}\pi_{jK})$.

Sampling of $\mathring{\mathbf{n}}_{\mathbf{ij}}$: By construction, since thinned edges are unobserved, all pairs of nodes (i, j) may have edges that are thinned. These edges are auxiliary variables necessary to obtain the full conditional posterior distribution of W_α. Also, to update community memberships, it is necessary to assign every thinned edge to a pair of discordant communities. Thanks to the properties of Poisson processes, we do not need to sample $\mathring{n}_{ij}$ for all the pairs of nodes. Instead, we can resort to an efficient sampling strategy, that exploits the construction based on thinning. More specifically, we can first sample a *proposed* total number of thinned edges $\mathring{N}$ from a Poisson $\left(2\sum_{i,j=1}^{N} w_i w_j\right)$. Then, we can assign each of these independently to a pair of nodes according to their sociabilities, and then independently to a pair of commmunities given the memberships of the assigned nodes (as in 7). We will then accept those edges that were assigned to discordant communities.

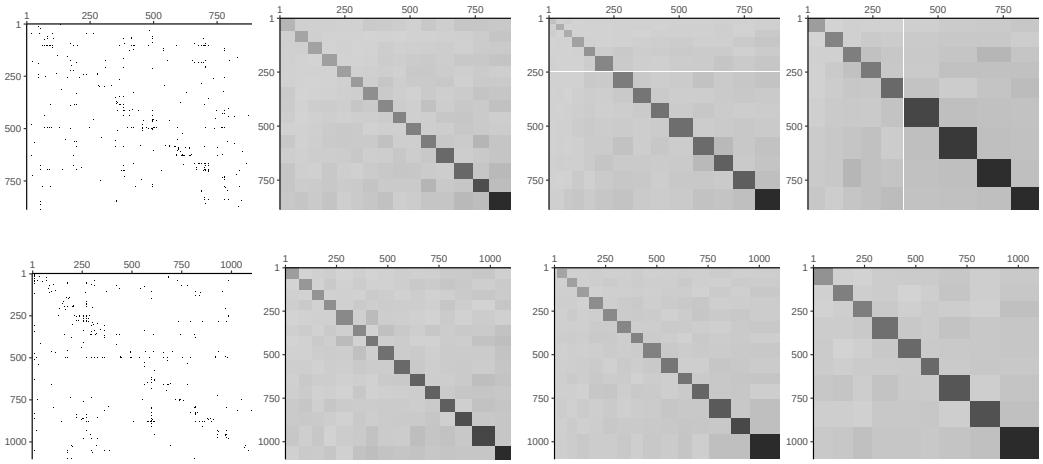
Finally, as illustrated in Figures 1 and 3, the total number of nodes with at least one edge in the multigraph *pre-* thinning N_α is likely to be different from both the total number of nodes with at least one edge *after* thinning and – as a consequence – from the total number of nodes N in the observed binary graph Y. Thus, there typically ought to be some latent number $N_\alpha - N \geq 0$ of nodes whose edges with the observed nodes have all been thinned. In order to learn such number, in our MCMC, we make an approximate update of N_α according to the mean ratio N_α/N from graphs simulated from the GGP prior given the latest samples of the hyperparameters α, σ and τ. We found that this heuristic method leads to convergence of the empirical MCMC-based estiamte of N_α across all simulated and real data that we considered (see Appendix).

5 Results

5.1 Simulation

We discuss a simulation study where we investigate the performances of our proposed method and the CGGP model of [17], based on simulated data generated from either model.

First, we present the results from a sparse graph with 15 communities, simulated from our TGGP model, by setting $\alpha = 250$, $\sigma = 0.1$, $\tau = 1$ for the distribution of the nodes' sociabilities and $\gamma = 10$ and $\zeta = 0.2$ for the distribution of nodes' membership. The adjacency matrix of the resulting simple, undirected, graph is shown in Figure 4 (a, bottom). Figure 4 (b, bottom) sorts the simulated nodes into blocks according to their main membership and plots the density of edges in each block. We see that the simulated graph has a clear block structure. We then run our MCMC sampler for 50,000 iterations, discarding the first 40,000 samples as burn-in. For our model fitting, we set the maximum number of communities to be 50 and we set $\gamma = 10$ and $\zeta = 0.5$ to allow for learning the number of communities. From a qualitative comparison of the (b) and (c) bottom panes in Figure 4, we see that



(a) adjacency matrix (b) true blocks (c) estimated, TGGP (d) estimated, CGGP

Figure 4: Results on two simulation experiments: a graph generated with the CGGP [17] (top) and a graph simulated with the proposed TGGP (bottom). Column (a) shows the true adjacency matrices, sorting nodes into blocks according to the strongest community membership; (b) shows the relative edge density in the true blocks, sorting blocks by intensity. Columns (c) and (d) show the blocks estimated, respectively, with the TGGP and the CGGP. In both simulations, we generate data setting $\alpha = 250$, $\sigma = 0.1$, $\tau = 1$ and $K = 15$ communities. For the CGGP, we set $a_k = 1/K$ and $b_k = 1$; for the TGGP, we set $\beta = (1/K, ..., 1/K)$ and $\zeta = 1$. In model fitting, for the TGGP, we set the maximum number of communities to 50, $\gamma = 10$ and $\zeta = 0.2$. The number of communities estimated by the CGGP is much smaller than the one estimated by the proposed TGGP, which is also closer to the truth.

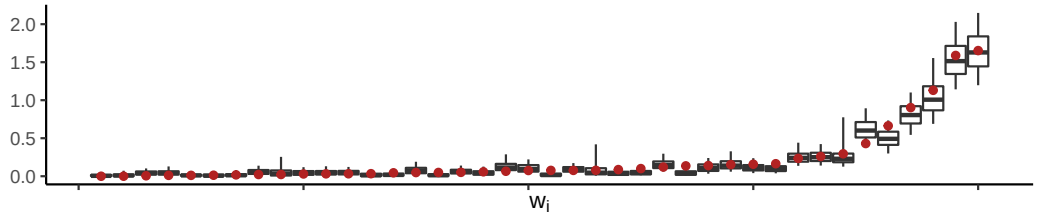


Figure 5: True simulated values (red) and 95% credible intervals for the sociability parameters of 40 randomly selected nodes.

the community memberships recovered from our model are close to the underlying truth. Figure 5 shows summaries of the posterior distributions of nodes' sociabilities for 40 randomly selected nodes spanning from low to relatively large sociability. We can see that the posterior distributions of all parameters tend to concentrate around the true values used for simulation (see, e.g., Figure 5 95% credible intervals for the sociability parameters). On the contrary, the CGGP model by [17] appears to struggle recovering the true community structure of the network (Figure 4, d bottom), despite setting the number of communities equal to the truth.

We also simulated a graph from the CGGP model of [17] with parameters set similarly as those used in the simulation above. In addition, we set $a_k = 1/K$ and $b_k = 1$ to make the prior on community memberships in [17] close to ours. We fit both the CGGP model and the TGGP model to the generated data. When fitting with the CGGP we still set the number of communities equal to the truth, i.e. 15. The results plotted in 4(d) suggests that the block structures learned by our model are closer to the true ones also under this mis-specification of the generative model.

5.2 Real network data

To assess the performance of our method for introducing overlapping communities in the GGP model for network as compared to the CGGP approach by [17], we evaluate the two methods on

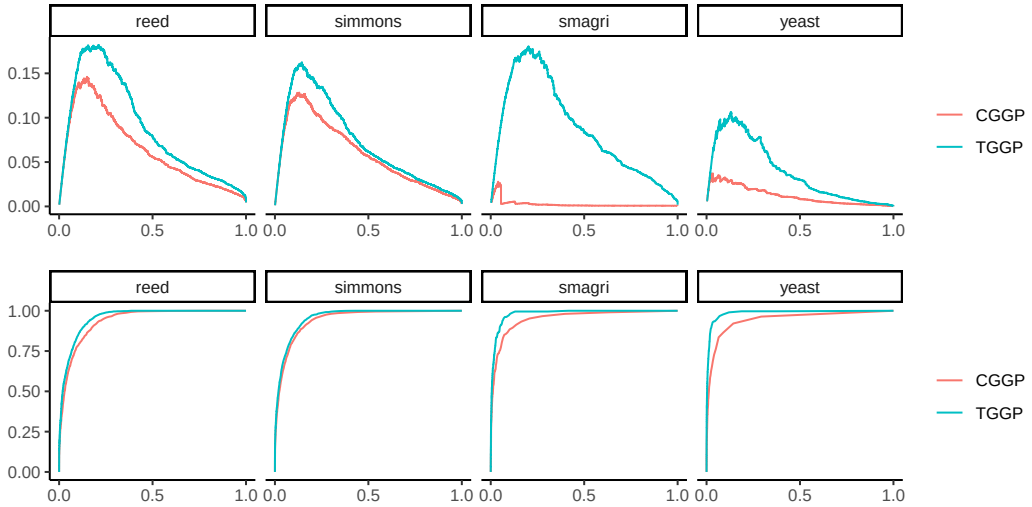


Figure 6: F-score vs. recall on prediction of missing edges (top) and ROC curve (bottom) on prediction of missing entries in the adjacency matrix.

two different forms of posterior predictive accuracy on 4 real world networks (a description of how the data was obtained and pre-processed can be found in the appendix). We run 50,000 iterations of the two models' MCMC on the fully observed data, and we used the model parameters estimated at the last iteration to infer (1) what entries of the adjacency matrix incorrectly classified as 0 are most likely to be edges and (2) the presence or absence of an edge if we consider 5% of entries of the adjacency matrix as missing. Note that the first task is more difficult than the second one, because most entries in the adjacency matrix are truly zeros (and so it is difficult to find the ones, while it is not as difficult to predict if an entry is a 0 or a 1). Figure 6 shows that our method does consistently better than the CGGP.

6 Discussion

We have proposed a framework for the analysis of binary network data that extends the GGP-based model of Caron and Fox [12] by allowing for overlapping community structures, depending both on the overall sociability of the nodes *and* the similarity of their community memberships. The assumed generative model proposes a novel latent multi-graph framework where nodes are allowed to connect between different communities, but the corresponding edges are hidden (thinned) in the projection giving rise to the observed network. With respect to alternative extensions of the original GGP-based model, e.g. the one in Todeschini et al. [17], the framework we propose allows for a regularized approach in the modeling of the latent strength of community memberships for each node, which results in improved inference and reconstruction of the true community structures both in simulation and data analysis. Furthermore, our model formulation allows for learning the number of communities directly from the data, as it does not require fixing this number *a priori*.

The proposed model is amenable to further extensions. Since the proposed mixed membership allocation framework relies on the well-studied Dirichlet-multinomial specification, it is possible to leverage the vast literature on (finite and infinite) hierarchical Dirichlet mixture models to model complex heterogeneous network data. It is also possible to include additional available information on covariates to guide the allocation of nodes to the communities. Finally, distributional and asymptotic properties of our model formulation can be investigated by exploiting recent results in the literature related to finite mixtures and mixtures of finite mixtures [see, e.g., 27, 28, 29, 18].

References

[1] A. Goldenberg, A.X. Zheng, S.E. Fienberg, and E.M. Airoldi. A survey of statistical network models. *Arxiv preprint arXiv:0912.5410*, 2009.

[2] Yuchung J Wang and George Y Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

[3] C. Kemp, J. Tenenbaum, T. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, 2006.

[4] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1(2):353–355, 1973.

[5] Edo M Airoldi, David Blei, Stephen Fienberg, and Eric Xing. Mixed membership stochastic blockmodels. *Advances in neural information processing systems*, 21, 2008.

[6] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, December 2006.

[7] Dae Il Kim, Prem K Gopalan, David Blei, and Erik Sudderth. Efficient online inference for bayesian nonparametric relational models. *Advances in neural information processing systems*, 26, 2013.

[8] S. H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, March 2001.

[9] Mark E. J. Newman. Networks: An introduction, 2010.

[10] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small–world' networks. *Nature*, 393: 440–442, June 1998.

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

[12] François Caron and Emily B Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1295–1366, 2017.

[13] Christian Borgs, Jennifer T Chayes, Henry Cohn, and Nina Holden. Sparse exchangeable graphs and their limits via graphon processes. *arXiv preprint arXiv:1601.07134*, 2016.

[14] Diana Cai, Trevor Campbell, and Tamara Broderick. Edge-exchangeable graphs and sparsity. *Advances in Neural Information Processing Systems*, 29, 2016.

[15] Victor Veitch and Daniel M Roy. Sampling and estimation for (sparse) exchangeable graphs. *The Annals of Statistics*, 47(6):3274–3299, 2019.

[16] Tue Herlau, Mikkel N Schmidt, and Morten Mørup. Completely random measures for modelling block-structured sparse networks. *Advances in Neural Information Processing Systems*, 29, 2016.

[17] Adrien Todeschini, Xenia Miscouridou, and François Caron. 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.

[18] Hemant Ishwaran and Mahmoud Zarepour. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, pages 941–963, 2002.

[19] Peter Orbanz and Daniel M Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence*, 37(2): 437–461, 2014.

[20] Olav Kallenberg. *Probabilistic symmetries and invariance principles*, volume 9. Springer, 2005.

[21] Philip Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396, 1986.

[22] Lancelot F James. Poisson process partition calculus with applications to exchangeable models and bayesian nonparametrics. *arXiv preprint math/0205093*, 2002.

[23] Antonio Lijoi, Ramsés H Mena, and Igor Prünster. Controlling the reinforcement in bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):715–740, 2007.

[24] Jim E Griffin and Fabrizio Leisen. Compound random measures and their use in bayesian non-parametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):525–545, 2017.

[25] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *arXiv [cs.LG]*, 5:1457–1469, 2004.

[26] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362, 2009.

[27] Sylvia Frühwirth-Schnatter and Gertraud Malsiner-Walli. From here to infinity: sparse finite versus dirichlet process mixtures in model-based clustering. *Advances in Data Analysis and Classification*, 13(1):33–64, 2019.

[28] Jeffrey W. Miller and Matthew T. Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

[29] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 3, 4, and 5
   (b) Did you describe the limitations of your work? [Yes] See Section 5 and 6
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We provided the assumptions and references, given the limitation of pages.
   (b) Did you include complete proofs of all theoretical results? [No] Most results we employ follow from analogous results in [12]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The code will be made publicly available on Github upon acceptance. The datasets are publicly available.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sections 4a nd 5
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 5, for example
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We will do in the resubmission
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? **[TODO]**
   (b) Did you mention the license of the assets? **[TODO]**
   (c) Did you include any new assets either in the supplemental material or as a URL? **[TODO]**
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[TODO]**
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[TODO]**
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[TODO]**
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[TODO]**
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[TODO]**