Bayesian Nonparametric Hawkes Processes

Jaivardhan Kapoor*
IIT Kanpur, Kanpur, India
jkapoor@iitk.ac.in

Antonio Vergari
MPI-IS, Tuebingen, Germany
antonio.vergari@tue.mpg.de

Manuel Gomez Rodriguez
MPI-SWS, Kaiserslautern, Germany
manuelgr@mpi-sws.org

Isabel Valera
MPI-IS, Tuebingen, Germany
isabel.valera@tue.mpg.de

1 Introduction

Models based on *temporal point processes* [4, 10] are well-fitted to capture the dynamics of continuously generated streams of data, for which the classical i.i.d. assumption clearly does not hold. Among these, approaches based on *Hawkes Processes* (HPs) [11] have been shown to be particularly suitable to model the inherent self-excitatory nature of many real-world domains like sequences of user-generated content, (e.g., tweets [14, 5]), their behaviour (e.g., online learning patterns [16], popularity [18]) or interactions in social networks (e.g., sharing a post [7, 8]).

In all these application scenarios, the accuracy of modeling such temporal dynamics would depend on how effectively a model exploits the *latent structure* underlying the observed events [10]: e.g., understanding the dynamics underlying the topics over the generated user messages [6, 16] as well as discovering the hidden community structure behind users' interactions [3, 21, 17, 15] greatly improves modeling the dynamics of the whole time series of events.

In this context, *Bayesian nonparametric priors* (BNPs) have become the model-of-choice to flexibly represent the complex distributions over these latent structures. For instance, the popular *Dirichlet Process* (DP) [9], its hierarchical [22] and nested variants [19], have been frequently employed as BNPs to model a possibly infinite number of latent features in the form of topics, tasks, patterns or communities associated to streams of events from several application domains [3, 6, 16, 21].

However, a closer look into this fast-growing literature reveals that *many of these works do not actually rely on a valid BNP model*. Specifically, based on the restaurant metaphor generative process of BNPs, they derive an *ad-hoc prior* over hidden structure in the data. While this construction allows dealing with structures of increasing complexity as the amount of data grows, it does not result in a valid probability distribution on an infinite-dimensional space. As a result, the probability of observing a particular pattern in the latent structure *vanishes*, as the self-excitation of the HP tends to zero. We call this issue the "vanishing prior" problem.

The contribution in this work is threefold. First, we formalize the vanishing prior problem in Section 2. Second, in Section 3, we develop a formal methodology that enables us to overcome the vanishing prior problem. Our proposed methodology allows us to disentangle the BNP prior from the HP in the event generative process. As a result, we provide general and modular framework to plug-in any BNP (e.g., the nested [2] or the franchise [22] variants of the Chinese Restaurant Process (CRP) [9]) to model the latent structure of Hawkes events, while keeping inference straightforward by employing Sequential Monte Carlo schemes to "reverse" the generative process. The proposed framework is indeed general enough not only to model user activity but also users' interactions, as in [3]. Finally, in Section 4, we revisit the state-of-the-art on BNP priors for HPs, verifying whether current approaches suffer from the vanishing prior or not, and providing an intuition on how to use our methodology to "fix" them.

*Work done during an internship with the Probabilistic Learning group at the MPI-IS, Tuebingen, Germany.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

2 The "vanishing prior" problem

A marked temporal point process (TPP) is a stochastic process whose realization consists of a sequence of discrete events localized in continuous time, taking the form of $e : (t, u, m)$, where t is the event time, u the user performing it, and m is the mark characterizing the content of the event (e.g., the sentiment of a message [5], or the words of a post [6, 16]). A TPP is fully characterized by its *intensity function* $\lambda(t)$, such that the probability of a new event happening in the time window $[t, t + dt)$ is $\lambda(t)dt$. The intensity $\lambda(t)$ of a Hawkes process (HP) is history-dependent and captures the *self-excitatory* nature of the events [11]. By the superposition property of HPs [23], it holds that

$$\lambda(t) = \lambda^{\mathsf{exo}} + \lambda^{\mathsf{endo}}(t) = \mu + \sum\nolimits_{i:e_i \in \mathcal{H}(t)} \gamma(t - t_i)$$

where $\lambda^{\mathsf{exo}} = \mu > 0$ is a baseline intensity independent of the history $\mathcal{H}(t)$, modeling the *exogenous* event activities, and $\gamma(t)$ is a triggering kernel modeling the self-excitation phenomenon across events, i.e., the *endogenous* event activity λ^{endo} [11, 23].

The above event representation can be extended as $e : (t, u, m, z)$ to account for the latent nature of an event, being captured by a random *latent variable* z, which we refer to as *pattern*. The event pattern parametrizes the mark distribution $p(m|z)$ as well as the triggering kernel $\gamma_z(t)$, such that the endogenous intensity component λ^{endo} can be rewritten as: $\sum_{i:e_i \in \mathcal{H}(t)} \gamma_{z_i}(t - t_i)$.

Several works have attempted to combine HPs with BNPs into a single generative process—which we will refer here as a BNP+HP model— to place a nonparametric prior distribution on the pattern distribution, where the number of patterns K is potentially infinite. For example, the patterns may correspond to the *clusters* (e.g., document topics) that every event belongs to, as in [6, 16], or to a *set of (binary) features* characterizing every event, as in [20], or even to paths in a tree hierarchy over users, as in [21].

Most of these works rely on generative processes based on the *restaurant metaphor* (where the random discrete probability measure induced by the BNP is integrated out [2]) to come up with an intuitive combined process for the BNP+HP model. However, many of these BNP+HP models [6, 20] construct an *ad-hoc prior* by directly mapping the choice of sampling a new pattern $z = K + 1$, or an already used pattern $z \in \{1 \dots K\}$ to respectively the intensities λ^{exo} and λ^{endo}. More specifically, they assume that, given the event time t, the latent pattern z can be sampled as follows:

$$z = \begin{cases} k & \text{with prob. } \frac{\lambda_k^{\mathsf{endo}}(t)}{\lambda(t)} \quad \text{for already seen patterns } k \in \{1, \dots, K\} \\ K+1 & \text{with prob. } \frac{\lambda^{\mathsf{exo}}}{\lambda(t)} \quad \text{for new (unseen) pattern} \end{cases}, \tag{1}$$

where $\lambda_k^{\mathsf{endo}}(t) = \sum_{i:e_i \in \mathcal{H}(t)|z_i=k} \gamma_k(t - t_i)$ is the endogenous intensity of pattern k, and the parameters corresponding to a new pattern $K+1$ (e.g., the kernel parameters and, the word distribution associated to a topic) are sampled from a continuous base measure, e.g., H_0 in a DP.

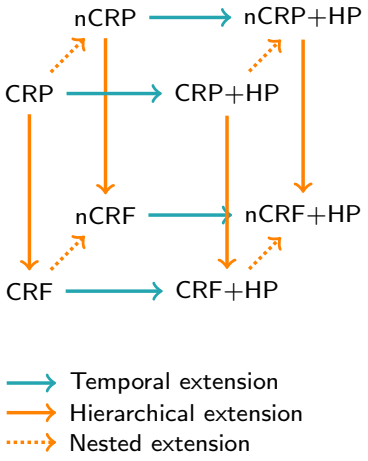
Although indeed the above equation allows for a potentially infinite number of patterns, this generative process does not lead to a valid probability distribution on an infinite-dimensional space, and therefore, to the *promised* BNP prior. More specifically, the above expression suffers from the fact that when $\lambda_k^{\mathsf{endo}}(t)$ becomes negligible with respect to the total intensity $\lambda(t)$ (or the exogenous one λ^{exo}), the probability of $z = k$ tends to zero. In other words, pattern k will *vanish*. As an extreme case, consider a memoryless system where $\gamma_{z=k}(t)$ is a delta function on zero, then the probability of observing the same pattern more than once is zero.

This problem has been briefly discussed in [16] for the CRP+HP model introduced in [6]. However, more recent approaches still suffer from the same issue [20]. For a model affected by the vanishing prior issue, patterns would vanish as their $\lambda_k^{\mathsf{endo}}(t)$ intensity tends to zero. The practical consequences of this are that i) the same pattern cannot be sampled twice from the continuous base measure, and, more importantly, ii) patterns cannot be shared across realizations of different stochastic processes, e.g., between users or contents, therefore making the benefits of modeling a shared latent structure negligible.

In the next section we explain how to avoid this issue and devise a methodology for distilling BNP+HP models, by employing any BNP from the literature.

3 Building-up BNP+HPs processes

In this section, we propose a general methodology to place a *generic* BNP prior over the parameters of a HP, avoiding the vanishing prior issue *by design*. To this end, we use the superposition property



Algorithm 1: BNP+HP: Generative process

Input: a BNP, triggering kernel function $\gamma(\cdot)$, N
Output: $\{e_n = (t_n, u_n, m_n, z_n)\}_{n=1}^N$

for $n = 1 \dots N$ **do**
  Compute $\lambda(t) = \sum_u \lambda_u(t|\mathcal{H}(t))$;
  Sample $t_n \sim \mathsf{Hawkes}(\lambda(t))$;
  Sample $u_n \sim \mathsf{Cat}(\{\lambda_u(t_n)/\lambda(t_n)\}_{u\in\mathcal{U}})$;
  Sample $b_n \sim \mathsf{Ber}(\lambda_{u_n}^{\mathsf{endo}}(t_n)/\lambda_{u_n}(t_n))$;
  if $b_n = 1$ **then**
    Sample $z_n \sim \mathsf{Cat}(\{\lambda_{u_n,k}^{\mathsf{endo}}(t_n)/\lambda_{u_n}^{\mathsf{endo}}(t_n)\}_{k=1}^K)$;
  else
    Sample $z_n \sim \mathsf{BNP}$;
  Update $\mathcal{H}_{u_n}(t)$;
  if $z_n = K + 1$ **then**
    Sample parameters for the new pattern;
  Sample mark $m_n \sim p(m_n|z_n)$;

Figure 1: Nested and arbitrarily deep restaurant-based BNPs (left) can be interchangeably combined with HPs via our methodology. The generic BNP+HP generative process (right) lets one distill a specific model from the lattice on the left, while providing a clean and valid way to sample between the temporal (blue) and BNP (orange) components.

of HPs to differentiate between the events that are triggered by the endogenous activity of the HP, from the events that are sampled directly from the BNP prior. More specifically, the distribution over the event pattern z given the event time t is a mixture distribution with two components, one corresponding to the HP and the other to the BNP prior, i.e.,

$$z|t \sim \begin{cases} \mathsf{Categorical}\left(\left[\frac{\lambda_1^{\mathsf{endo}}(t)}{\lambda^{\mathsf{endo}}(t)}, \dots, \frac{\lambda_K^{\mathsf{endo}}(t)}{\lambda^{\mathsf{endo}}(t)}\right]\right) & \text{with prob. } \frac{\lambda^{\mathsf{endo}}(t)}{\lambda(t)} \\ \mathsf{BNP} & \text{with prob. } \frac{\lambda^{\mathsf{exo}}}{\lambda(t)} \end{cases} . \tag{2}$$

Note that in the above expression, in contrast to Eq. 1, while events triggered by the endogenous activity of the HP are sampled from a categorical distribution over the normalized endogenous intensities associated to already observed patterns, i.e., $k \in 1, \dots, K$, events sampled from the BNP can be assigned to either an already seen pattern $k \in 1, \dots, K$ or a new (unseen) pattern $K + 1$. As a consequence, even in absence of active history (i.e, $\lambda_k^{\mathsf{endo}}(t) = 0$), the probability of reusing an observed pattern is greater than zero, avoiding therefore the vanishing prior problem.

Most importantly, note that by doing so, we create a *clean interface* between the history-dependent rate component of the BNP+HP process (i.e., the endogenous intensity λ^{endo}) and a history-invariant latent prior (i.e., the exogenous rate λ^{exo}). As a matter of fact, we can therefore use Eq. 2 to develop a general generative BNP+HP process, which allow plugging-in any desired BNP prior, while keeping the HP-related procedures for sampling and inference unchanged. For example, one may easily combine any restaurant-based BNP generative model, such as the Chinese Restaurant Process (CRP); its hierarchical version, the Chinese Restaurant Franchise (CRF); or their nested counterparts, the nCRP and nCRF, with an HP to develop its corresponding temporal extension (see Figure 1).

Algorithm 1 sketches the generative process for a generic BNP+HP model over N events representing user activities in the form $\{e_n = (t_n, u_n, m_n)\}_{n=1}^N$. Here, we split the sampling of the pattern z into two steps: 1) sampling from an auxiliary binary variable b_n indicating whether the event is triggered by endogenous activity (λ^{endo}) or by exogenous activity (λ^{exo}); and 2) sampling the event pattern z_n conditioned on the latent binary variable b_n. Note that the auxiliary variable b_n allows to clearly show that the temporal component part (blue) is not affected by the choice of a specific BNP model (orange), and vice-versa, preventing thus from the vanishing prior issue to happen.

Inference. Moreover, we would like to remark that when using restaurant-based BNP generative models, one can easily derive a general sequential Monte Carlo (SMC) algorithm, by "reversing" the generative process in Algorithm 1. This inference framework i) is suitable for a large number of BNP models, and ii) exploits the temporal dependencies in the observed data to sequentially sample the latent pattern associated to each event, therefore scaling to large datasets. Algorithm 2 in the Appendix sketches the SMC inference process. For a concrete particularization of our inference scheme, see the derivation for the the CRP+HP model in [16].

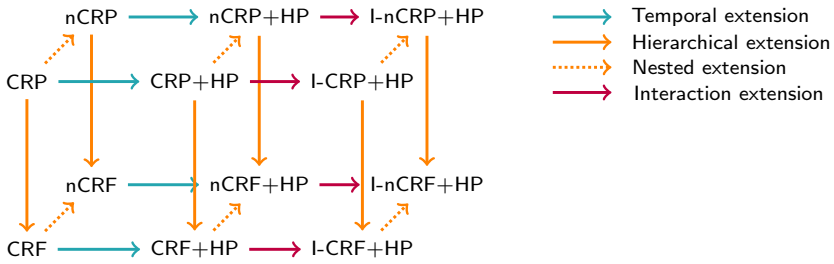


Figure 2: Extending the lattice over BNP+HP models with our methodology to I-BNP+HP models to deal with interactions between pairs of users.

3.1 Extension to interaction models

BNPs have been successfully used to model user *interactions*, e.g., to capture static user community structures [13, 1, 12]. Here, we refer to these models as I-BNPs. There also exist temporal extensions of I-BNP models that employ HPs to model interactions events between users, such as e.g. exchange of emails, likes, etc [3, 20, 17] by accounting for the *mutual-excitation* of pairs of users in addition to individual self-excitations. In these models every interaction event can be represented as $\{e_n = (t_n, s_n, d_n, m_n, z_n^{(s)}, z_n^{(d)}\}_{n=1}^N$, where s_n is the sender, d_n the receiver, and $z_n^{(s)}, z_n^{(d)}$ are the latent patterns associated to respectively the sender and receiver. As an example, $z_n^{(s)}, z_n^{(d)}$ may correspond to the communities that the users belong to.

The methodology proposed in the previous section can easily be extended to account for I-BNPs, and thus for user interactions. To this end, one simply needs to modify Algorithm 1 by

i) replacing line 4 with the sampling steps for the sender and receiver, i.e., $s_n \sim \mathsf{Cat}\left(\{\lambda_{u,\cdot}(t_n)/\lambda(t_n)\}_{u\in\mathcal{U}}\right)$ and $d_n \sim \mathsf{Cat}\left(\{\lambda_{s_n,u}(t_n)/\lambda_{s_n,\cdot}(t_n)\}_{u\in\mathcal{U}}\right)$; and
ii) substituting the BNP prior in line 9, by the corresponding I-BNP model.

Here, the interaction intensity is given by $\lambda_{s_n,d_n}(t) = \mu_{s_n,d_n} + \sum_{j:t_j\in\mathcal{H}_{s_n,d_n}(t)} \gamma_{z_j^{(s)} z_j^{(d)}}(t - t_j)$.

4 BNP+HP: Where are we now?

In this section, we navigate the lattice of possible combinations of (I-)BNPs and HPs (see Fig. 2), in order to pose into a common methodological framework the efforts performed in the literature. Specifically, we verify the validity of existing approaches, i.e., we check whether they suffer from the vanishing prior issue or not, in which case we discuss the possibility to solve it via our methodology.

In [6], a Hawkes extension of the CRP is unsuccessfully attempted, leading to a follow-up work [16] which finally manages to define a sound CRP+HP model. Note that, although the model in [16] is referred to as Hierarchical Dirichlet Hawkes, it in fact induces a CRP+HP. However, this model can be easily extended to account for a distribution over patterns (clusters) per user by replacing the CRP by a CRF in line 9 of Algorithm 1.

Similarly, one may replace the CRP by a nCRP, which induces a infinitely deep and wide tree structure. The work in [21] attempts to provide an I-nCRP+HP to model hierarchies of user communities. However, the nCRP is fully detached from the event generation process. In order to solve this issue via our methodology, an event pattern would correspond to a path in the tree induced by the (I-)nCRP.

The Hawkes IRM in [3] successfully derives a variant of the I-CRP+HP, where each user intensity is fully characterized by the unique community she belongs to. As such, all interaction events between two users are driven by their respective communities' intensities. In order to extend such model in the direction of mixed-membership methods, where a user may belong to several communities at the same time, one would just need to move to the I-CRF+HP in the lattice. Similar in spirit to an I-CRF+HP, the Hawkes-CCRM model of [17] proposes a completely random measure prior to enforce sparsity in the community structures, while focusing on non-marked HPs.

Finally, a temporal extension of the Indian Buffet process (IBP) via HPs has been proposed [20] to model user interactions via latent user features. Unfortunately, this approach also suffers from the vanishing prior issue and does not deliver a promised valid BNP.

Hence, although there is still work to do in order to complete the lattice (and extending it to other BNP models, e.g., the IBP), we believe that the proposed methodology paves the way moving forward.

References

[1] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

[2] D. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7, 2010.

[3] C. Blundell, J. Beck, and K. A. Heller. Modelling reciprocating relationships with hawkes processes. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2600–2608. Curran Associates, Inc., 2012.

[4] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

[5] A. De, I. Valera, N. Ganguly, S. Bhattacharya, and M. Gomez Rodriguez. Learning and forecasting opinion dynamics in social networks. In *Advances in Neural Information Processing Systems 29*, pages 397–405. 2016.

[6] N. Du, M. Farajtabar, A. Ahmed, A. J. Smola, and L. Song. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the 21th KDD Conference*, pages 219–228. ACM, 2015.

[7] M. Farajtabar, N. Du, M. Gomez Rodriguez, I. Valera, H. Zha, and Le Song. Shaping social activity by incentivizing users. In *Advances in Neural Information Processing Systems 27*, pages 2474–2482. 2014.

[8] M. Farajtabar, Y. Wang, M. Gomez Rodriguez, S. Li, H. Zha, and L. Song. Coevolve: A joint point process model for information diffusion and network co-evolution. In *Advances in Neural Information Processing Systems 28*, pages 1954–1962. 2015.

[9] T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[10] M. Gomez Rodriguez and I. Valera. Learning with temporal point processes. *Tutorial at ICML*, 2018.

[11] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[12] Q. Ho, A. Parikh, L. Song, and E. Xing. Multiscale community blockmodel for network exploration. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the 14th AISTATS*, volume 15, pages 333–341. PMLR, 2011.

[13] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5, 2006.

[14] R. Kobayashi and R. Lambiotte. Tideh: Time-dependent hawkes process for predicting retweet dynamics. In *ICWSM*, pages 191–200, 2016.

[15] Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *International Conference on Machine Learning*, pages 1413–1421, 2014.

[16] C. Mavroforakis, I. Valera, and M. Gomez Rodriguez. Modeling the dynamics of online learning activity. *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[17] Xenia Miscouridou, Francois Caron, and Yee Whye Teh. Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data. *arXiv preprint arXiv:1803.06070*, 2018.

[18] S. Mishra, M. Rizoiu, and L. Xie. Feature driven and point process approaches for popularity prediction. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1069–1078. ACM, 2016.

[19] J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan. Nested hierarchical dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270, Feb 2015.

[20] X. Tan, V. Rao, and J. Neville. The indian buffet hawkes process to model evolving latent influences. *Proceedings of UAI*, 2018.

[21] X. Tan, V. Rao, and J. Neville. Nested crp with hawkes-gaussian processes. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 1289–1298. PMLR, 09–11 Apr 2018.

[22] Y. W. Teh, M. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, pages 1566–1581, 2006.

[23] H. Xu, D. Luo, X. Chen, and L. Carin. Benefits from superposed hawkes processes. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 623–631. PMLR, 09–11 Apr 2018.

Algorithm 2: BNP+HP: inference

Input: A sequence of events $\{e_n = (t_n, u_n, m_n)\}_{n=1}^{N}$, P number of particles, θ threshold for particle resampling, triggering kernels prior $\Gamma(a_\alpha, b_\alpha)$, user intensities prior (a_μ, b_μ),τ

Output: user intensities $\{\mu_u\}_{u=1}^{U}$, triggering kernels $\{\alpha_l\}_{l=1}^{L}$, BNP-specific counts and parameters

1 Initialize $\mu_u \sim \Gamma(a_\mu, b_\mu) \forall u \in \mathcal{U}$, all counts to 0;
2 Initialize particle weights $w_n^{(p)} = \frac{1}{P} \;\; \forall p \in 1 \dots P$;
3 **for** $n = 1 \dots N$ **do**
  `// for all particles`
4 **for** $p = 1 \dots P$ **do**
5 Sample pattern z_n;
6 **if** z_n *is globally new* **then**
7 Sample $\alpha_{z_n} \sim \Gamma(a_\alpha, b_\alpha)$;
8 Update triggering kernel prior for pattern $a_{\alpha_{z_n}}, b_{\alpha_{z_n}}$;
9 Compute $p(m_n | z_{1:n-1})$;
10 Update user intensities μ_u parameters $\;\; \forall u \in 1 \dots U$;
11 Compute $p(t_n | u_n)$;
12 Update $w_n^{(p)} = w_{n-1}^{(p)} p(m_n | z_{1:n-1}) p(t_n | u_n)$;
13 Update BNP-specific counts;
14 Normalize particle weights;
15 **if** $\sum_{p=1}^{P} {w_n^{(p)}}^2 \leq \theta$ **then**
16 Resample particles using systematic resampling;

Appendix

A. Inference in user activity models

Algorithm 2 sketches how to perform inference via Sequential Monte Carlo (SMC) in our methodology. More specifically, updating the time kernel parameter α and base rate μ is performed by exploiting the conjugacy between the Gamma distribution (Γ) and Poisson process, when priors $\Gamma(a_\alpha, b_\alpha)$ and $\Gamma(a_\mu, b_\mu)$ are respectively used. Patterns are sampled in a similar fashion to the posterior sampling in Chinese Restaurant Franchise([22], Section 4.1).

B. Inference in user interaction models

SMC Inference for I-BNP+HP models in the interaction case can be derived in a similar way as for the activity case, discussed above. The only substantial difference would be in computing the Hawkes likelihood (line 11). Conditioned on both a sender s_n and destination d_n, such a term would become:

$$p(t_n | s_n, d_n) = \lambda_{s_n, d_n}(t_n) \exp\left(- \int_{t_{n-1}}^{t_n} \lambda_{s_n, d_n}(t) dt \right)$$

Depending on the concrete specifications of the generative model, the sampling of patterns would change accordingly, taking into account the case if the generative process only considers patterns for the sender, or a combination of patterns from the sender and the receiver.