FULLY NEURAL-BASED OUT-OF-DISTRIBUTION DETECTION FOR TEMPORAL POINT PROCESSES

Rafael Lima *
Samsung R&D Institute Brazil
Avenida Cambacicas 1200
Campinas-SP, Brazil
rg.lima@samsung.com

Chris Solomou
University of York
Deramore Lane Heslington York
York, UK
cs2291@york.ac.uk

ABSTRACT

Temporal Point Processes have undergone increasing relevance in the modeling of continuous-time event streams. Regarding their applicability, one important aspect is that of detecting anomalous, or out-of-distribution, sequences. Recent works have focused on parametric models for this out-of-distribution detection. In the present work, we give a theoretical background treatment of the anomaly detection problem applied to TPPs, describe our fully neural-based strategy, show how a fully neural-based strategy of improved generalization outperforms traditional parametric approaches, and validate its effectiveness against a state-of-the-art approach on data of controlled generation.

1 INTRODUCTION

The ubiquity of asynchronous temporal behaviour in a myriad of both natural and social phenomena has prompted a surge of works investigating the applications of Temporal Point Process (TPP) (Daley & Vere-Jones, 2003) modeling to domains such as earthquake aftershock prediction (Ogata, 1999), retweeting behaviour modeling (Zhao et al., 2015; Rizoiu et al., 2018), academic citation counting (Xiao et al., 2016) and high-frequency financial transactions (Bacry et al., 2015b;a).

TPP modeling equates the problem of modeling one or more real-valued event time arrival sequences to that of finding an underlying corresponding Conditional Intensity Function (CIF) $\lambda(t)$, which is the expected arrival rate of new events as a function of time.

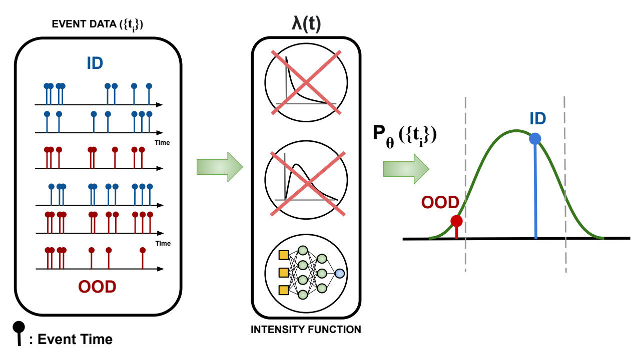

Figure 1: Our framework for fully neural-based out-of-distribution detection of temporal point processes.

Several strategies have been used to approximate a CIF best suited to a given set of sequences, ranging from simple parametric models (Ogata, 1981; Kobayashi & Lambiotte, 2016; Etesami et al., 2016) and grid-based methods (Mohler et al., 2012; Zhou et al., 2020; Achab et al., 2017; Bacry & Muzy, 2016; Lewis & Mohler, 2011; Zhou et al., 2013; Yang et al., 2017) to the more contemporary approaches using Recurrent Neural Networks (Du et al., 2016; Upadhyay et al., 2018; Xiao et al., 2018; 2017b; Yang et al., 2018), Generative Adversarial Networks (Xiao et al., 2017a; Goodfellow et al., 2014), as well as self-attentive models and Transformers (Zuo et al., 2020; Zhang et al., 2019).

A related problem to that of abstracting a set of sequences to a CIF model is that of detecting anomalous sequences (Shchur et al., 2021), i.e., those sequences which present a time arrival be-

*This work was not supported by any organization.

haviour rather uncommon w.r.t. the typical behaviour of the event arrivals corresponding to a given set of sequences. That is of major importance in situations as diagnosing server failures, identifying intrusions from malicious users in a system, and detecting frauds or shifts in a given market structure, to name a few examples.

The current approach makes use of a combination of a Goodness-of-Fit (GoF) statistic test with a TPP model which is learned over the distribution of inter-event times, and thus is insensitive to permutations of the given sequences. Time-clustering behaviour on natural and social phenomena, however, often possess a history-dependent behaviours which have been widely investigated with models such as self-exciting and self-damping point processes (Bacry et al., 2015b; Rizoiu et al., 2018).

In the present work, we propose a fully neural-based strategy (Omi et al., 2019) for the TPP model subjected to the GoF testing for anomaly detection problem in TPPs, as a means of capturing history-dependent information on the TPP modeling for improving the reliability of the detection tests. In the following, we give a theoretical background treatment of the anomaly detection problem applied to TPPs, describe our fully neural-based strategy, and validate its effectiveness against state-of-the-art approaches on real and simulated data.

2 Theoretical Background

In the following, we give a theoretical treatment of Temporal Point Processes (TPPs), the anomaly detection problem applied to TPPs, and describe our fully neural-based approach.

2.1 Temporal Point Processes

Temporal Point Process (TPP) modeling equates the problem of modeling one or more real-valued event time arrival sequences, such as $(t_0, t_1, ..., t_N)$ $(t \in \mathbf{R})$, to that of finding an underlying corresponding Conditional Intensity Function (CIF) $\lambda(t)$ such that

$$\lambda(t) = \mathbb{E}\{dN_t = 1|\mathcal{H}(t)\}, \tag{1}$$

where N_t is denoted the Counting Process, while $\mathcal{H}(t) = \{t_i\}$ $(t_i < t)$ is referred to as the *History* of the TPP up to time t, and $dN_t = 1$, if there is an event at time t, and $dN_t = 0$, otherwise.

2.2 Anomaly Detection of Temporal Point Processes

The Anomaly Detection for TPPs is equated to a type of an Out-of-Distribution (OoD) Detection problem, which aims to define if a given random instance of time-event sequence $\tilde{\int}$ belongs to an underlying unknown TPP which is manifested by a given set of sequences $\mathcal{S}$.

More formally, it is defined as a null hypothesis test:

$$\mathbb{H}_0 : \tilde{s} \sim \mathbb{P}_{\mathcal{S}} \qquad \mathbb{H}_1 : \tilde{s} \sim \mathbb{Q} \neq \mathbb{P}_{\mathcal{S}}, \tag{2}$$

where $\mathbb{P}_{\mathcal{S}}$ is the true underlying TPP generating the set $\mathcal{S}$, while $\mathbb{Q}$ is a distinct TPP.

Associated to this OoD Detection formulation is the Goodness-of-Fit (GoF) testing, which corresponds to a hypothesis test over a known generating probability distribution $\mathbb{P}_{\mathcal{M}}$:

$$\mathbb{H}_0 : \tilde{s} \sim \mathbb{P}_{\mathcal{M}} \qquad \mathbb{H}_1 : \tilde{s} \sim \mathbb{Q} \neq \mathbb{P}_{\mathcal{M}}, \tag{3}$$

where $\mathcal{M}$ corresponds to a known model for the TPP.

This knowledge of the model allows us to compute a test statistic $\int(\tilde{s})$ and its associated two-sided p-value $p_{\mathcal{S}}(\tilde{s})$

$$p_{\mathcal{S}}(\tilde{s}) = 2\text{min}\{\Pr(\int(\mathcal{S}) \leq \int(\tilde{s})|\mathbb{H}_0), 1 - \Pr(\int(\mathcal{S}) \leq \int(\tilde{s})|\mathbb{H}_0)\} \tag{4}$$

3 Fully Neural-Based Out-of-Distribution Detection for Temporal Point Processes

Several Neural-based variants have been recently proposed for modeling TPPs, as a way of leveraging modern Deep Learning techniques and approaches to increase the accuracy and variance of these time event-sequence models. Most notably, Recurrent Neural Networks (Du et al., 2016), Long Short Term Memory networks (Mei & Eisner, 2017), and Transformers (Zuo et al., 2020) have been applied to TPPs.

A highly performing approach, the fully neural-based TPP (Omi et al., 2019), proposes the use of a dense neural network to model the time integrated value of the CIF, also know as the *Compensator* function

$$\Phi_\theta(\tau) = \int_0^\tau \lambda(\tau)d\tau \quad t \in [0, T] \tag{5}$$

From that, by making use of the automatic differentiation techniques widely available in Machine Learning frameworks, it constructs a loss function equivalent to the Loglikelihood (LLH) of the TPP-realized sequence:

$$LLH(\tilde{s}) = \sum_{i=1}^{N_{\tilde{s}}} \log(\lambda(t_i)) - \int_0^{T_{\tilde{s}}} \lambda(t)dt = \sum_{i=1}^{N_{\tilde{s}}} \log\left(\frac{\partial\Phi_\theta(\tau = t_{i+1} - t_i)}{\partial\tau}\right) - \Phi_\theta(\tau = t_{i+1} - t_i) \tag{6}$$

By choosing $\theta^{\text{MAX}} \in \Theta$, the model parameters, such that

$$\theta^{\text{MAX}} = \arg\max_{\Theta} \text{LLH}(\Phi_\theta(\mathcal{S})), \tag{7}$$

we are left with a Maximum Likelihood Estimator which we may use as the known model for GoF testing.

The present work consists of, given θ^{MAX} as defined in Equation 7, we define a test statistic

$$\int(\cdot) = \mathrm{e}^{\text{LLH}(\Phi_{\theta^{\text{MAX}}}(\cdot))} \tag{8}$$

from where we can perform a two-sided test for a sequence $\tilde{s}$ as

$$p_{\mathcal{S}}(\tilde{s}) = 2\min\{\Pr(\mathrm{e}^{\text{LLH}(\Phi_{\theta^{\text{MAX}}}(\mathcal{S}))} \leq \mathrm{e}^{\text{LLH}(\Phi_{\theta^{\text{MAX}}}(\tilde{s}))}|\mathbb{H}_0), 1 - \Pr(\mathrm{e}^{\text{LLH}(\Phi_{\theta^{\text{MAX}}}(\mathcal{S}))} \leq \mathrm{e}^{\text{LLH}(\Phi_{\theta^{\text{MAX}}}(\tilde{s}))}|\mathbb{H}_0)\} \tag{9}$$

to classify those sequences as ID or OOD based on a p-value ($p_{\mathcal{S}}(\tilde{s})$) threshold of 0.05.

4 Discussion of Architecture Search

4.1 Architecture Description

In this section, we present the architecture of our model, which had its performance evaluated using a variety of synthetic data.

The core of our model consists of an RNN that is trained to learn temporal patterns by processing sequences of events. The sequences of events that are used as input to the RNN are windowed into subsequences of length 20 for avoiding potential gradient vanishing/explosion. Our architecture follows the proposed methodology by (Omi et al., 2019), where the first hidden layer receives the elapsed time (τ) and the hidden state of the RNN as inputs.

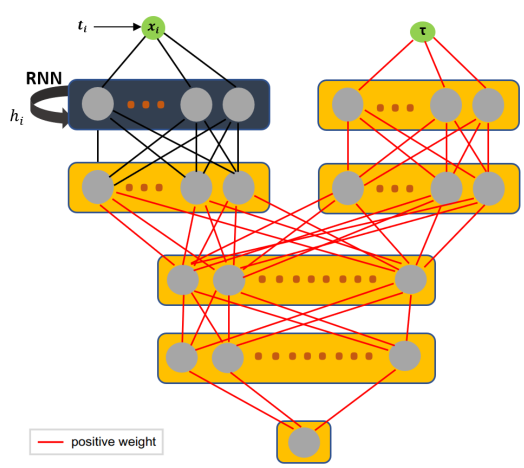

Figure 2: Our proposed neural architecture for out-of-distribution detection of time event sequences.

The units of the RNN and the Dense layers were determined using a grid search between $[16, 32, 64, 128]$, and the units that produced the best results were selected. Particularly the number of units for the RNN was set to 128 and the number of units for the Dense layer to 32. We initialize our weights using Glorot Uniform initialization and similar to (Omi et al., 2019), we constrain the weights to be positive. We apply a hyperbolic tangent activation function to the RNN and subsequent hidden layers. Constraining the weights to be positive ensures that the output of the tanh function is preserved, allowing the distinction of different sequences.

For the output layer, since the Cumulative Hazard Function is a monotonically increasing function, a single unit with an equally monotonically increasing activation function of exponential type is utilized for modelling this behavior.

4.2 Discussion

Throughout our experiments, we initially noticed mixed performance when detecting OoD sequences. Specifically, the model was successful in detecting the OoD sequences when trained on a particular distribution and tested on another. However, when the model was trained on the same distribution as the OoD sequences, it was unable to detect them well. In such cases, the model seemed to overfit to the training data, and consequently becoming unable to generalize when subjected to new sequences.

To remedy this issue, we subsequently performed a grid search for detecting the best units for each layer and experimented with different activation functions. For the prediction (output) layer we utilized a Dense layer with a single unit with an exponential activation function, for simulating the monotonically increasing behavior of the CHF.

5 Experiments

For evaluating the effectiveness of our fully neural-based approach for detecting OoD sequences in TPP, we perform experiments with data sets comprising 100 sequences of synthetic data. The data are simulated from seven different known types of point processes. The processes and their intensity functions are described below:

- **Stationary Poisson Process (SPP)**: The arrival rate of events remains constant over time, and is defined by a constant unitary intensity function
$$\lambda_{SPP}(t) = \lambda = 1, \forall t \in [0, T]. \tag{10}$$
- **Non-Stationary Poisson Process (NSPP)**: The arrival rate varies with time.Consists of a non-constant intensity value which is conditionally independent of past events:
$$\lambda(t) = \mathbb{E}\{dN_t = 1|\mathcal{H}(t)\} = \mathbb{E}\{dN_t = 1\} = A\sin\left(\frac{2\pi t}{L}\right) + 1 \in \mathbb{R}^+, \forall t \in [0, T], \tag{11}$$
with $A = 0.99$ and $L = 20000$.
- **Stationary Renewal Process (SRP)**: The inter-arrival time distribution remains constant over time. Each sequence $\{t_i\}_{i=1}^N$ is sampled by:
$$y_i \sim \frac{1}{s * y_i * \sqrt{2\pi}} \mathrm{e}^{-\dfrac{\log^2(y_i)}{2s^2}}, \text{ with } y_i = \frac{t_i}{\mathrm{e}^{-\dfrac{s^2}{2}}} \tag{12}$$
- **Non-Stationary Renewal Process (NSRP)**: The inter-arrival time distribution changes over time. Each sequence $\{t_i\}_{i=1}^N$ is sampled by:
$$\lambda(t - t_i) = \sin\left(\frac{2 * \pi * (t - t_i)}{20000}\right) * 0.99 + 1 \tag{13}$$
- **Self-Correcting Process (SCP)**: The inter-arrival time between events depends on the time elapsed since the last event. Each sequence $\{t_i\}_{i=1}^N$ is sampled by:
$$t_i = t_{i-1} + \left(\log \frac{e * \mu}{\mathrm{e}^x} + 1\right) / \mu, \text{ with } e \sim \mathrm{Exp}(\beta = 1) \tag{14}$$
where $t_0 = 0$ and $x_i = x_{i-1} - 1$.

- **Hawkes Process Type I (HP-I)**: The occurrence of an event increases the probability of the occurrence of another event. Its intensity function $\lambda(t)$ for a sequence $\{t_i\}_{i=1}^N$ is given by Hawkes (1971a;b):
$$\lambda_{HP-I}(t) = \mu + \sum_{t_i<t} \phi_1(t - t_i), \text{ with } \mu \in \mathbb{R}_+^* \text{ and } \phi_1(t) = 0.8 * \mathrm{e}^{-t}. \tag{15}$$
- **Hawkes Process Type II (HP-II)**: The intensity function $\lambda(t)$ for a sequence $\{t_i\}_{i=1}^N$ is given by:
$$\lambda_{HP-II}(t) = \mu + \sum_{t_i<t} \phi_2(t - t_i) + \sum_{t_i<t} \phi_3(t - t_i), \tag{16}$$
with $\mu \in \mathbb{R}_+^*$,$\phi_2(t) = 0.4 * \mathrm{e}^{-t}$ and $\phi_3(t) = 0.4 * \mathrm{e}^{-20*t}$.

The goal of our experiments was to gauge the ability of our fully-neural strategy to accurately distinguish between in-distribution (ID) and out-of-distribution (OoD) sequences. We evaluated the model's performance on sequences generated from the same process as the training data versus sequences generated from a different process. The results are presented in Figure 3 and are determined by the detection rate for each scenario. In this paradigm, the desired outcome is to have a low detection rate for ID sequences (generated from the same process) and a high detection rate for OoD sequences (generated from a different process).

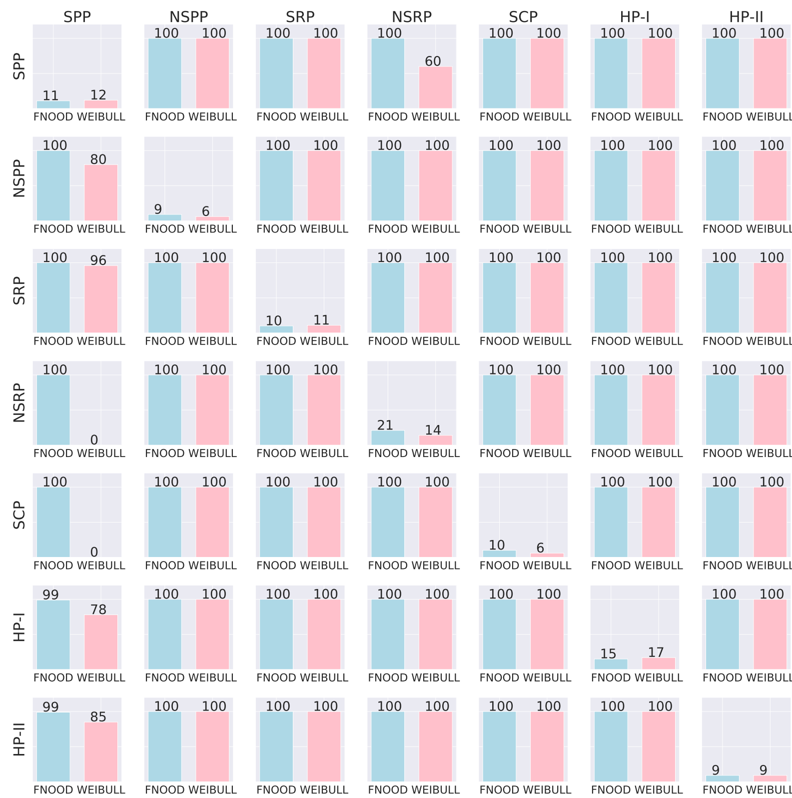

Figure 3: Comparison of Out-of-Distribution detection rate between our fully neural-based approach (FNOOD) and the baseline (Weibull distribution). Both models were trained on 100 sequences from each process, represented by the horizontal axis, and evaluated on sequences from all other processes, represented by the vertical axis. The performance is measured by the detection of 100 test sequences that were correctly classified as OoD.

6 Conclusion

In this work, we propose a fully-neural based approach for detecting Out-of-Distribution sequences in temporal point processes. We show the effectiveness of our proposal by testing it in a wide variety of synthetic data. The results are evaluated using a GoF test, allowing to compute a test statistic for detecting anomalous sequences. Our experiments show that our method consistently outperforms the Weibull distribution, which serves as a baseline, when both are evaluated on data of controlled generation.

REFERENCES

Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate hawkes integrated cumulants. *Journal of Machine Learning Research*, 18:192:1–192:28, 2017.

E. Bacry and J. Muzy. First- and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62(4):2184–2202, 2016.

Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Mean-field inference of hawkes point processes. *CoRR*, abs/1511.01512, 2015a. URL `http://arxiv.org/abs/1511.01512`.

Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *arXiv*, 2015b.

Daryl Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer, 2003.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016.

J. Etesami, N. Kiyavash, K. Zhang, and K. Singhal. Learning network of multivariate hawkes processes: A time series approach. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2016.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2672–2680, 2014.

Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, (1):201–213, 1971a.

Alan G. Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(3):438–443, 1971b. ISSN 00359246. URL `http://www.jstor.org/stable/2984686`.

Ryota Kobayashi and Renaud Lambiotte. Tideh: Time-dependent hawkes process for predicting retweet dynamics. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016.*, 2016.

Erik Lewis and George Mohler. A nonparametric em algorithm for multiscale hawkes processes. *Journal of Nonparametric Statistics*, 1(1):1–20, 2011.

Hongyuan Mei and Jason Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017.

George O. Mohler, Martin B. Short, P. Jeffrey Brantingham, Frederic P. Schoenberg, and George E. Tita. Self-exciting point process modelling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2012.

Yosihiko Ogata. On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

Yosihiko Ogata. Seismicity analysis through point-process modelling: A review. *Pure and Applied Geophysics*, 155(5):471–507, 1999.

Takahiro Omi, Naonori Ueda, and Kazuyuki Aihara. Fully neural network based model for general temporal point processes. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 2120–2129, 2019.

Marian-Andrei Rizoiu, Young Lee, and Swapnil Mishra. Hawkes processes for events in social media. In *Frontiers of Multimedia Research*, pp. 191–218. 2018.

Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, Jan Gasthaus, and Stephan Günnemann. Detecting anomalous event sequences with temporal point processes. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 13419–13431, 2021.

Utkarsh Upadhyay, Abir De, and Manuel Gomez-Rodriguez. Deep reinforcement learning of marked temporal point processes. *CoRR*, abs/1805.09360, 2018. URL `http://arxiv.org/abs/1805.09360`.

Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiangfeng Wang, Xiaokang Yang, Stephen M. Chu, and Hongyuan Zha. On modeling and predicting individual paper citation count over time. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pp. 2676–2682, 2016.

Shuai Xiao, Mehrdad Farajtabar, Xiaojing Ye, Junchi Yan, Xiaokang Yang, Le Song, and Hongyuan Zha. Wasserstein learning of deep generative point process models. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 3250–3259, 2017a.

Shuai Xiao, Junchi Yan, Xiaokang Yang, Hongyuan Zha, and Stephen M. Chu. Modeling the intensity function of point process via recurrent neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pp. 1597–1603, 2017b.

Shuai Xiao, Hongteng Xu, Junchi Yan, Mehrdad Farajtabar, Xiaokang Yang, Le Song, and Hongyuan Zha. Learning conditional generative models for temporal point processes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 6302–6310, 2018.

Guolei Yang, Ying Cai, and Chandan K. Reddy. Recurrent spatio-temporal point process for check-in time prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pp. 2203–2211. ACM, 2018.

Yingxiang Yang, Jalal Etesami, Niao He, and Negar Kiyavash. Online learning for multivariate hawkes processes. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 4944–4953, 2017.

Qiang Zhang, Aldo Lipani, Ömer Kirnap, and Emine Yilmaz. Self-attentive hawkes processes. *CoRR*, abs/1907.07561, 2019.

Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1513–1522, 2015.

Feng Zhou, Zhidong Li, Xuhui Fan, Yang Wang, Arcot Sowmya, and Fang Chen. Fast multi-resolution segmentation for nonstationary hawkes process using cumulants. *International Journal of Data Science and Analytics*, 10, 06 2020. doi: 10.1007/s41060-020-00223-3.

Ke Zhou, Hongyuan Zha, and Le Song. Learning triggering kernels for multi-dimensional hawkes processes. In *Proceedings of the International Conference on Machine Learning*, pp. 1301–1309, 2013.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. *CoRR*, abs/2002.09291, 2020. URL `https://arxiv.org/abs/2002.09291`.