ADVERSARIAL ANOMALY DETECTION FOR MARKED SPATIO-TEMPORAL STREAMING DATA

Shixiang Zhu, Henry Shaowu Yuchi, Yao Xie

Georgia Institute of Technology
H. Milton Stewart School of Industrial and Systems Engineering
Atlanta, GA, USA

ABSTRACT

Spatio-temporal event data are becoming increasingly commonplace in a wide variety of applications, such as electronic transaction records, social network data, and crime incident reports. How to efficiently detect anomalies in these dynamic systems using these streaming event data? This work proposes a novel anomaly detection framework for such event data combining the Long Short-Term Memory (LSTM) and marked spatio-temporal point processes. The detection procedure can be computed in an online and distributed fashion via feeding the streaming data through an LSTM and a neural network-based discriminator. This work studies the false-alarm-rate and detection delay using theory and simulation and shows that it can achieve weak signal detection by aggregating local statistics over time and networks. Finally, we demonstrate the good performance using real-world data sets.

***Index Terms*—** Anomaly detection, adversarial learning, long short-term memory, marked spatio-temporal point processes.

1. INTRODUCTION

Data recording spatio-temporal events exist ubiquitously in our daily life. These spatio-temporal data range from electronic transaction records at large chain department stores, earthquake records, to criminal history records by the police. In many cases when an abnormal incident takes place, it will result in anomalies in the recorded sequence (Figure 1). Such sequential anomaly data usually have a distinctive pattern compared to normal data, but they are difficult to obtain. Therefore, it has been a challenge to capture the anomalous pattern and detect such anomalies in the dynamic systems efficiently and accurately, especially when only insufficient one-class anomaly data are available.

In this work we propose a novel anomaly detection framework for streaming event data leveraging the power of adversarial learning [1, 2, 3]. The anomaly events are modeled using marked spatio-temporal point processes, where the historical information is specified as the last hidden state of the Long Short-Term Memory (LSTM). The neural network-based discriminator in the adversarial framework can be naturally used as an anomaly detector. The detection procedure can be carried out in an online and distributed fashion via feeding the streaming data through the LSTM and the discriminator.

The major contribution of this work is two-fold: (1) The work has obtained a robust anomaly detector based on a limited amount of training real data. It is proposed to generate "realistic" fake samples using an adversarial framework to improve the discriminator; (2) The work proposes modeling the event sequence data by integrating the versatile point process framework with LSTM. This gives the model better interpretability and flexibility in capturing the true underlying pattern.

We focus on the prediction accuracy and detection delay from application data and simulations. Our study shows the proposed approach is capable of achieving sequential anomaly detection for weak signals by aggregating local statistics over time and networks over time, location, and mark space. Finally, we demonstrate the satisfactory performance using data from real-world applications.

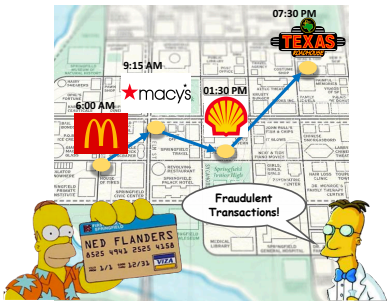


Fig. 1: An example of anomaly detection for marked spatio-temporal streaming data.

Related work. In the field of signal processing, there have been numerous articles tackling anomaly/outlier detection problem from other perspectives. [4] employs a statistic based on log-likelihood ratio or log-posterior density ratio, which is claimed to be a good estimator of goodness-to-fit. This statistic is used for anomaly test. In [5], outlier detection in sequential online data is looked into. This work adopts incremental decision trees for multi-model density estima-

tion, which forms the basis of an online anomaly detector. The l_1 norm principle component analysis and its variations are introduced in [6], which is able to catch change in signal subspace, effectively detecting the anomaly. These novel and rigorous approaches prove to be performing well for detecting signal anomalies. However, they are not an ideal fit in this problem setting, since the streaming data in our problem contain extra categorical information such as location.

Recently, there have been a number of attempts in integrating the idea of General Adversarial Network (GAN) and anomaly detection. Several existing methods are analyzed in [7]. These methods discussed stem directly from the structure of GAN without modeling of the streaming data. In the work of [8], Long Short Term-Recurrent Neural Network (LSTM-RNN) is applied in GAN to capture the distribution of multivariate time series of streaming data for cyber-physical systems. The work of [9] uses a GAN-based approach to carry out anomaly detection for medical image scans, taking advantage of discriminator feature information. Similarly looking into image anomaly detection, the work in [10] attempts to obtain the latent space representation from the generator of GAN to identify anomalies.

Typical adversarial problems draw attention to obtaining a high-quality fake data generator using the adversarial network. It aims to detect anomalies when normal data pattern is available and can be exploited for a traditional two-class change detection. However, in the setting of this work, only anomaly data are available, which calls for a one-class detection approach. Therefore in our work, more emphasis is placed on learning an efficient discriminator which is able to detect anomaly data accurately.

A few pieces of more recent research are closely related to this work in the effort of exploiting the LSTM structure or adversarial learning setting. In [11, 12], Hawkes process is combined with recurrent neural networks. Both make use of the LSTM structure to model the conditional intensity function similar to our work. However, the detailed structure of the two approached differ. In [12], the model consists of standard discrete-time LSTM which results in event intervals being coded into the model. The model also only contains a single intensity function $\lambda(t)$ with simple exponential decay. On the other hand, [11] proposed a continuous-time LSTM model by constructing a modular model with separate intensity functions, allowing more flexibility. Both of the papers look into temporal settings only. Our work further extends the modeling to a marked spatio-temporal setting. Additionally, in the work of [13], an adversarial learning strategy similar to this work is deployed. The objective is to improve maximum likelihood estimation (MLE) of predictive point process models. The aforementioned papers focus on modeling using point processes with adversarial learning strategies. But these papers do not take anomaly detection into consideration. A final relevant paper [14] applies the adversarial network to the classification problem, but it is limited to images only.

2. MODELING

Assume we have a set of marked spatio-temporal anomalous sequences $\mathcal{X} = \{\boldsymbol{x}\}$. Let $\boldsymbol{x} = \{x_1, x_2, \ldots, x_{N_T}\}$ be a single sequence where N_T is the number of events in the time window T. Each tuple $x_i = \{t_i, \boldsymbol{s}_i, \boldsymbol{m}_i\}$ denotes a single event occurred at time $t_i \in [0, T)$, at location $\boldsymbol{s}_i \in \mathcal{S} \subseteq \mathbb{R}^2$, associated with marks $\boldsymbol{m}_i \in \Omega^* = \mathbb{R}^d \times \Omega$ where Ω is a categorical mark space.

Our goal is to devise a discriminator $D : [0, T) \times \mathcal{S} \times \Omega^* \to (0, 1)$. Given a sequence $\boldsymbol{x}$ with arbitrary length, this discriminator $D(\boldsymbol{x}) > b$ if the sequence $\boldsymbol{x}$ is anomalous. Otherwise $D(\boldsymbol{x}) \leq b$, where b is a preset threshold.

Now we present our sequential data model, which is adapted from the neural Hawkes model in [11, 12, 15, 16]. The model represents the historical information of a sequence $\boldsymbol{x}$ using the final hidden state $\boldsymbol{h}_{N_T+1}$ after feeding the data sequentially into the LSTM.

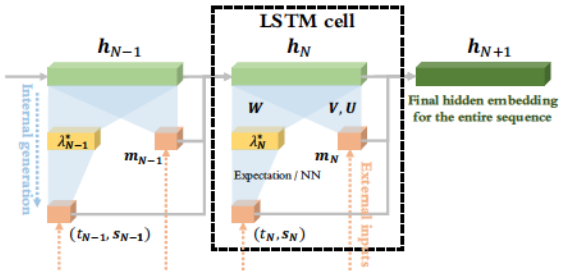


Fig. 2: An illustration of our proposed LSTM based model. The model can either take external points as input or generate points internally. The final hidden embedding after the last moment contains the information for the entire input or generated sequence.

We characterize the anomaly behavior of events in a sequence using a conditional intensity function $\lambda(t, \boldsymbol{s}, \boldsymbol{m}|\mathcal{H}_t)$, which is the probability of observing an event in the marked spatio-temporal space $[0, T) \times \mathcal{S} \times \Omega^*$ given the history of past events $\mathcal{H}_t$: $\lambda(t, \boldsymbol{s}, \boldsymbol{m}|\mathcal{H}_t)dtd\boldsymbol{s}d\boldsymbol{m} = \mathbb{P}\{x_{i+1} \in [t, t + dt] \times |B(\boldsymbol{s}, \Delta\boldsymbol{s})| \times |B(\boldsymbol{m}, \Delta\boldsymbol{m})||\mathcal{H}_t\}$, where $|B(\boldsymbol{\nu}, \Delta\boldsymbol{\nu})|$ are the Lebesgue measure of the ball $B(\boldsymbol{\nu}, \Delta\boldsymbol{\nu})$ with radius $\Delta\boldsymbol{\nu}$.

We model the nonlinear dependency of current event from past events using the LSTM. As shown in Figure 2, for the i^{th} event occurring at the time t_i, the data tuple $(t_i, \boldsymbol{s}_i)$ is fed as input into the LSTM unfolded up to the $i + 1^{\text{th}}$ event. The embedding $\boldsymbol{h}_i \in \mathbb{R}^p$ represents the memory effect, which is the influence from the past events. The LSTM updates $\boldsymbol{h}_{i-1}$ to $\boldsymbol{h}_i$ by taking into account the impact of the current event x_i. Finally, we use $\boldsymbol{h}_i$ to represent the influence of the history up to time $t, \forall t : t_i < t < t_{i+1}$.

Given the i^{th} input x_i and the last hidden state $\boldsymbol{h}_{i-1}$, we can obtain the hidden state $\boldsymbol{h}_i$ of LSTM (p is the dimension of the hidden state), which is defined as

$$\boldsymbol{h}_i = \boldsymbol{o}_i(x_i) \circ \sigma_h(\boldsymbol{c}_i(x_i)) = \boldsymbol{h}_{i-1}(x_i),$$

where σ_h is the hyperbolic tangent function, $\boldsymbol{o}_i$ is the current

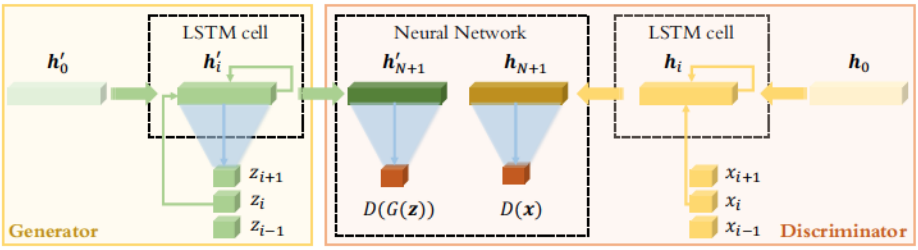


Fig. 3: Architecture of the adversarial learning framework, which consists of a LSTM structure and a fully connected neural network. For generator, the LSTM generates fake samples $\boldsymbol{z} = \{z_i\}$ given the initial hidden state h'_0. For discriminator, it takes real samples $\boldsymbol{x} = \{x_i\}$ as input given the initial hidden state h_0. A fully-connected neural network takes the last hidden state h_{N+1} or h'_{N+1} of a sequence of samples as input, and yields the probability of being anomaly.

output gate's activation state, and c_i is the current cell state. Both of o_i and c_i depend on the last hidden state h_{i-1}, and take x_i as the current input. It differs from the conventional LSTM structure because the external input of hidden state is the LSTM's output at the last moment. For more details of the LSTM structure, please refer to Appendix A[1].

Since we have assumed the mark is conditionally independent of location and time of events, the conditional intensity function is defined as

$$\begin{aligned}\lambda(t, s, \boldsymbol{m}|\mathcal{H}_t) &\approx \lambda(t, s, \boldsymbol{m}|\boldsymbol{h}_i) \\ &= \lambda_g(t, s|\boldsymbol{h}_i) \cdot p(\boldsymbol{m}|\boldsymbol{h}_i),\end{aligned} \tag{1}$$

where $\lambda_g(t, s|\boldsymbol{h}_i)$ is the spatio-temporal conditional intensity and $p(\boldsymbol{m}|\boldsymbol{h}_i)$ is the conditional probability density of marks.

Spatio-temporal representation. The spatio-temporal conditional intensity function $\lambda_g(t, s|\boldsymbol{h}_i)$ can be specified as

$$\lambda_g(t, s|\boldsymbol{h}_i) = f\left(\boldsymbol{W}_h^\intercal \boldsymbol{h}_i(t, s) + b_h\right), \tag{2}$$

where $f(\cdot)$ is the softplus function. $\boldsymbol{h}_i(t, s)$ is the hidden state with partial input t, s (see Appendix A). Weights $\boldsymbol{W}_h$ and bias b_h are the trainable parameters. To attain more expressiveness for the conditional intensity function, it can also be extended to multi-layers structure without too much efforts.

For simplicity of notation, we denote the conditional intensity $\lambda(\cdot|\boldsymbol{h}_i), \lambda_g(\cdot|\boldsymbol{h}_i)$ as $\lambda^*(\cdot), \lambda_g^*(\cdot)$. Let $f_g^*(t, s)$ be the corresponding conditional spatio-temporal probability density. Note the conditional spatio-temporal intensity $\lambda_g^*(t, s)$. The conditional probability density $f_g^*(t, s)$ is defined as:

$$f_g^*(t, s) = \lambda_g^*(t, s) \cdot \exp\left\{-\int_{t_n}^{t}\int_{\mathcal{S}} \lambda_g^*(\tau, \boldsymbol{\nu}) d\tau d\boldsymbol{\nu}\right\}.$$

Then we can estimate the time and location for the next event using the expectation:

$$\begin{bmatrix}\hat{t}_{i+1} \\ \hat{\boldsymbol{s}}_{i+1}\end{bmatrix} = \begin{bmatrix}\int_{t_i}^{T} \tau \int_{\mathcal{S}} f_g^*(\tau, \boldsymbol{\nu}) d\tau d\boldsymbol{\nu} \\ \int_{\mathcal{S}} \boldsymbol{\nu} \int_{t_i}^{T} f_g^*(\tau, \boldsymbol{\nu}) d\tau d\boldsymbol{\nu}\end{bmatrix}.$$

In general, the integration above cannot be obtained analytically. Therefore, numerical integration techniques which are commonly utilized are applied here to compute the expectation.

Mark representation. Given the hidden representation $\boldsymbol{h}_i$ up to the i^{th} event, we consider each type of the marks conditional independent of each other, *i.e.*, $p(\boldsymbol{m}|\boldsymbol{h}_i) = \prod_{\ell=0}^{d} p(\boldsymbol{m}[\ell]|\boldsymbol{h}_i)$. For the categorical mark $\boldsymbol{m}[0] \in \{1, \ldots, K\}$, we model the conditional probability of mark $\boldsymbol{m}[0]$ as a multinomial distribution defined as:

$$p(\boldsymbol{m}[0] = k|\boldsymbol{h}_i) = \frac{\exp(\boldsymbol{V}_k \boldsymbol{h}_i + b_k)}{\sum_{\kappa=1}^{K} \exp(\boldsymbol{V}_\kappa \boldsymbol{h}_i + b_\kappa)},$$

where weights $\boldsymbol{V} = \{\boldsymbol{V}_k\}$ and bias $\{b_k\}$ are the trained parameters.

For continuous marks $\boldsymbol{m}[\ell] \in \mathbb{R}, \forall 1 \leq \ell \leq d$, we model their conditional probability density $\boldsymbol{m}[\ell]$ as a Gaussian distribution defined by:

$$\boldsymbol{m}[\ell]|\boldsymbol{h}_i \sim \mathcal{N}(\boldsymbol{U}_\ell \boldsymbol{h}_i + b_\ell, \sigma),$$

where weights $\boldsymbol{U} = \{\boldsymbol{U}_\ell\}$ and bias $\{b_\ell\}$ are the trained parameters.

3. ADVERSARIAL ANOMALY DETECTION

The anomaly data is assumed to be generated by a real data distribution denoted by p_d. The fake data is generated by a fake data distribution denoted by p_z. Denote $G(\boldsymbol{z})$ the LSTM generator and $D(\boldsymbol{x})$ the discriminator. The generator G implicitly defines a probability distribution $\boldsymbol{z} \sim p_z$ as the distribution of fake trajectories obtained by $G(\boldsymbol{z})$ when $\boldsymbol{z}$ is a random initialization of the LSTM state. The discriminator D is a fully connected multi-layer neural network where the input layer of D is the last LSTM hidden state with external input $\boldsymbol{x}$, and the output layer is a softmax which yields the probability that the sequence is an anomaly trajectory.

To learn the discriminator while improving the generator, we follow [2, 3] to play an adversarial game by minimax the following objective function as shown in Figure 3.

$$\begin{aligned}\max_D \min_G \quad & \mathbb{E}_{\boldsymbol{x}\sim p_d(\boldsymbol{x})}\left[\log D(\boldsymbol{x})\right] + \\ & \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}\left[\log(1 - D(G(\boldsymbol{z})))\right]\end{aligned} \tag{3}$$

Performing the anomaly detection, we feed the data points of an unknown sequence $\boldsymbol{x}$ into our well-trained LSTM one at

[1] https://arxiv.org/abs/1910.09161

a time. Denote the first i events in $\boldsymbol{x}$ as $\boldsymbol{x}_i$. The alarm would be raised at step i once $D(\boldsymbol{x}_i)$ is larger than a preset threshold b. A general threshold of $b = 0.5$ can be adopted here.

4. NUMERICAL EXAMPLES

To evaluate the performance of the proposed anomaly detection approach, we apply it to two real applications as below:

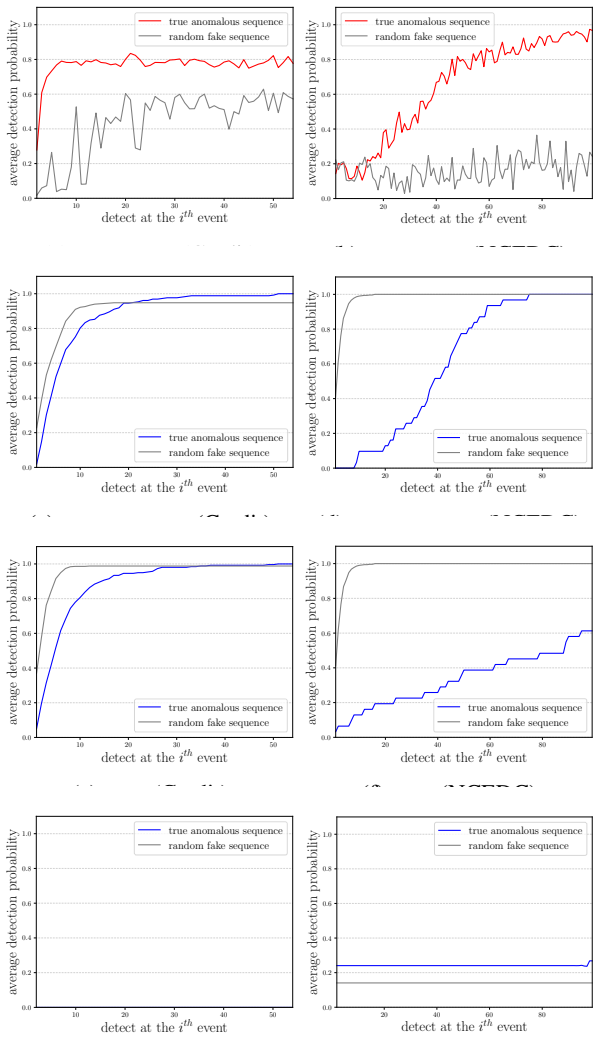
Earthquake event data. The Northern California Earthquake Data Center (NCEDC) provides data [17] which comes from broadband, short period, strong motion seismic sensors, GPS, and other geophysical sensors. This dataset contains 16,401 major earthquake records with magnitudes larger than 3 from 1978 to 2018 in the Northern California area.

Credit card fraud transaction data. The records of identified fraudulent credit card transactions contain date and location of transaction, transaction amount, and loss type. Each case is associated with a sequence of frauds on a specific card. There are in total 534 different locations, 641,071 fraud transactions observed, and 30,078 credit cards involved.

Comparing the proposed approach (`AdvLSTM`) with three other baseline methods, its efficacy can be understood. These baselines are briefly introduced as follow: (1) `PCA`. Principle component analysis is carried out and the principle feature components are obtained. Then the event sequences of various lengths are compared with the features. If a specific testing sequence corresponds to an outlier from the principle features, then it is determined that such sequence should not be considered as an anomaly. PCA does not take the order of events in the sequence into account; (2) `PCA + CUSUM`. PCA-based detection with CUSUM has a similar structure to the standard PCA method. The major difference is the introduction of CUSUM statistics to the method as the trigger for detection. This takes the sequence of events into consideration. (3) `MLE-LSTM`. Using the LSTM framework only to find representations of events by Maximum Likelihood Estimation as shown in Appendix B. Detection is declared if the event is sufficiently dissimilar to these representations.

All approaches are applied to both the credit card fraudulent transactions and NCDEC earthquake event data. The desired performance is that anomalies can be correctly detected as quickly as possible at a considerable level of accuracy. On the other hand, if random noise data is fed towards the detection metric, ideally they should not be picked up by the detector as anomalies.

Following the experiments, the performance of these methods can be obtained in terms of the accuracy of detecting anomaly and erroneously identifying noise data as anomalies. It is observed in Fig.4 that for the method proposed in our work, the detection probability of correctly detecting the anomaly data in both cases is high. For the credit card fraud, the probability of detection reaches 80% within the first 5 events of a sequence, whilst for the earthquake case, the same level of detection is reached within the first 40 events. Simultaneously the probabilities of identifying random noise as anomalies in both cases are significantly lower. This indicates the proposed method is able to detect the anomalies while being resistant to random noise, indicating the power of adversarial learning. On the contrary, the baselines are not performing as well. They have generally failed to differentiate between the noise data and genuine anomaly data. Therefore they are not as accurate in anomaly detection as the devised method in this work.



(g) `MLE-LSTM` (Credit) (h) `MLE-LSTM` (NCEDC)

Fig. 4: Comparisons between our method and baselines.

5. CONCLUSIONS

This paper has proposed a novel anomaly detection approach leveraging the power of adversarial learning. The anomaly data is modeled using marked spatio-temporal point processes framework where historical information is specified as the final hidden embedding of an LSTM. Using real credit card frauds and earthquake records, it is shown that this approach outperforms other baseline methods in terms of prediction accuracy.

6. REFERENCES

[1] Chris Dyer, "Notes on noise contrastive estimation and negative sampling," 2014.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 2672–2680. Curran Associates, Inc., 2014.

[3] Ian Goodfellow, "On distinguishability criteria for estimating generative models," 2014.

[4] Elizabeth Hou, Yasin Yılmaz, and Alfred O Hero, "Anomaly detection in partially observed traffic networks," *IEEE Transactions on Signal Processing*, vol. 67, no. 6, pp. 1461–1476, 2019.

[5] Kaan Gokcesu, Mohammadreza Mohaghegh Neyshabouri, Hakan Gokcesu, and Suleyman Serdar Kozat, "Sequential outlier detection based on incremental decision trees," *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 993–1005, 2018.

[6] Panos P Markopoulos, Mayur Dhanaraj, and Andreas Savakis, "Adaptive l1-norm principal-component analysis with online outlier rejection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 6, pp. 1131–1143, 2018.

[7] Federico Di Mattia, Paolo Galeone, Michele De Simoni, and Emanuele Ghelfi, "A survey on gans for anomaly detection," 2019.

[8] Dan Li, Dacheng Chen, Jonathan Goh, and See kiong Ng, "Anomaly detection with generative adversarial networks for multivariate time series," 2018.

[9] Thomas Schlegl, Philipp Seebck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical Image Analysis*, vol. 54, pp. 30 – 44, 2019.

[10] Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft, "Image anomaly detection with generative adversarial networks," in *Machine Learning and Knowledge Discovery in Databases*, Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Ifrim, Eds., Cham, 2019, pp. 3–17, Springer International Publishing.

[11] Hongyuan Mei and Jason Eisner, "The neural hawkes process: A neurally self-modulating multivariate point process," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, USA, 2017, NIPS'17, pp. 6757–6767, Curran Associates Inc.

[12] 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*, New York, NY, USA, 2016, KDD '16, pp. 1555–1564, ACM.

[13] Junchi Yan, Xin Liu, Liangliang Shi, Changsheng Li, and Hongyuan Zha, "Improving maximum likelihood estimation of temporal point process via discriminative and adversarial learning," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. 7 2018, pp. 2948–2954, International Joint Conferences on Artificial Intelligence Organization.

[14] Quan Kong, Bin Tong, Martin Klinkigt, Yuki Watanabe, Naoto Akira, and Tomokazu Murakami, "Active generative adversarial network for image classification," 2019.

[15] Sepp Hochreiter and Júrgen Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[16] Alex Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, vol. 385 of *Studies in Computational Intelligence*, Springer, 2012.

[17] Northern California Earthquake Data Center. UC Berkeley Seismological Laboratory, "NCEDC," 2014.

[18] D.J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*, Probability and Its Applications. Springer New York, 2007.