Learning Granger Causality from Instance-wise Self-attentive Hawkes Processes

Dongxia Wu†[2] **Tsuyoshi Idé**[*] **Aurélie Lozano**[*] **Georgios Kollias**[*] **Jiří Navrátil**[*]
Naoki Abe[*] **Yi-An Ma**† **Rose Yu**†
†University of California, San Diego, [*]IBM T. J. Watson Research Center
{dowu,yianma,roseyu}@ucsd.edu, {tide,aclozano,gkollias,jiri,nabe}@us.ibm.com

Abstract

We address the problem of learning Granger causality from asynchronous, interdependent, multi-type event sequences. In particular, we are interested in discovering *instance-level* causal structures in an unsupervised manner. Instance-level causality identifies causal relationships among individual events, providing more fine-grained information for decision-making. Existing work in the literature either requires strong assumptions, such as linearity in the intensity function, or heuristically defined model parameters that do not necessarily meet the requirements of Granger causality. We propose Instance-wise Self-Attentive Hawkes Processes (`ISAHP`), a novel deep learning framework that can directly infer the Granger causality at the event instance level. `ISAHP` is the first neural point process model that meets the requirements of Granger causality. It leverages the self-attention mechanism of the transformer to align with the principles of Granger causality. We empirically demonstrate that `ISAHP` is capable of discovering complex instance-level causal structures that cannot be handled by classical models. We also show that `ISAHP` achieves state-of-the-art performance in proxy tasks involving type-level causal discovery and instance-level event type prediction.

1 Introduction

Automated causal discovery from noisy time-series data is a fundamental problem for machine learning in complex domains. Granger causality (Granger, 1969b) is a popular notion of (pseudo) causality in time series data. While extensive prior work exists on Granger causality from regular time series, with the vector autoregressive model being a primary tool (see, e.g., (Shojaie and Fox, 2022) for the latest review), relatively less work has been done on analogous problems on discrete *event* sequences. These event sequences often occur at irregular intervals. There could also be multiple types of events interacting with each other. Together, they pose significant challenges to Granger causality structure learning tasks.

Recently, many have used *point process* model as a vehicle for causal discovery of stochastic temporal events (Xu et al., 2016; Eichler et al., 2017; Zhang et al., 2020b). In particular, Hawkes process (Hawkes, 1971; Hawkes and Chen, 2021) is one of the basic building blocks for Granger-causal analysis of *multi-type event sequences* (i.e., event data that come with a type attribute, indicating what the event is). Thanks to its additive structure over the historical events in the event intensity function, the task of causal discovery can be performed by maximum likelihood estimation of the kernel matrix, which characterizes the causal relationship among different event types.

However, causal structure among event types only provides coarse-grained information that is aggregated over the event sequence. It lacks fine-grained details of causal relationships among individual events. From a practical perspective, extracting instance-level information for causal analysis is critical. In medical diagnosis, for example, we are interested in capturing precursor events that may have caused a specific symptom of patients. These precursors would be valuable for early detection, screening, and treatment as a prevention. Aggregated type-level causality would only reveal information about generic symptoms categories, rather than identifying the exact precursors that progress directly into diseases.

For instance-level causal analysis, there are three lines

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

of research to date. One is based on the classical Hawkes process, typically through the minorization-maximization (MM) framework (Veen and Schoenberg, 2008; Lewis and Mohler, 2011; Li and Zha, 2015; Wang et al., 2017; Idé et al., 2021). But a classical Hawkes model places the linearity assumption in the intensity function, limiting its expressive power. Neural point process models (Xiao et al., 2019; Zuo et al., 2020; Zhang et al., 2020a) are designed to address this limitation but they typically embed event history in the form of a latent state vector losing instance-wise information. As a result, the attention-based score does not necessarily represent the Granger causality. The third one is based on a post-processing step following maximum likelihood (Zhang et al., 2020b), which incurs additional computational costs.

In this paper, we propose a novel deep Hawkes process model, the "instance-wise self-attentive Hawkes process (`ISAHP`)," that achieves better expressiveness while also enabling direct instance-level Granger-causal analysis. From a mathematical perspective, the key design principle is to maintain an *additive structure*, where causal interaction is represented as the summation of individual historical events. To capture complex causal interactions potentially involving multiple events, we leverage the self-attention mechanism of the Transformer model (Vaswani et al., 2017). `ISAHP` can directly capture instance-wise causal relationships with its additive structure. We can also easily obtain type-level causal relationships by simple aggregation. To the best of our knowledge, `ISAHP` is the first deep point process model that allows direct instance-wise causal analysis without post-processing.

We empirically demonstrate that `ISAHP` can discover complex instance-level causal structures that cannot be handled by the classical models and neural point process models without post-processing. Furthermore, our experiments show that `ISAHP` achieves state-of-the-art performance in two proxy tasks, one involving type-level causal discovery and the other involving instance-level event type prediction. It confirms that the instance- and type-level causal inference tasks are coupled, and our proposed framework manages to model them coherently and holistically.

2 Background

2.1 Type-Level Granger Causality

We are given a training data set $\mathcal{D}$ with S event sequences, each of which contains L_s events.

$$\mathcal{D} \triangleq \{(t_i^s, k_i^s) \mid i \in \{1, \dots, L_s\}, s \in \{1, \dots, S\}\}. \quad (1)$$

In the data set, each event is represented by its timestamp of occurrence t and a type attribute k. The timestamps are sorted so that $t_i^s \geq t_j^s$ for $i > j$. We assume that the total number of event types is K and therefore $k_i^s \in \{1, \dots, K\}$.

We formalize the problem of causal discovery from event sequences as a *unsupervised density estimation* task. Given the history of events $\mathcal{H}_t = \{(t_i^s, k_i^s)_{t_i<t}\}$, temporal point processes are generally characterized by a conditional distribution called the intensity function $\lambda_k(t \mid \mathcal{H}_t)$. The intensity function describes the expected rate of occurrence for event type k at a future time point t, and is assumed to have a specific parametric form for causal discovery. For the classical multivariate Hawkes process (MHP), we assume a simple linear form.

$$\lambda_k(t \mid \mathcal{H}_t) = \mu_k + \sum_{i:t_i<t} \alpha_{k,k_i} \phi_{k,k_i}(t - t_i) \quad (2)$$

where μ_k is the background intensity for event type k, $\{\alpha_{k,k_i}\}$ forms a $K \times K$ matrix called the kernel matrix representing the type-level causal influence, and $\phi_{k,k_i}(\cdot)$ is the decay function of the causal influence.

The additive form in Eq. (2) naturally leads to causal interpretation among event types. For example, if $\alpha_{1,2} = 0$, the probability of the next event occurrence of type-1 is not affected at all by the type-2 events in the history. This is indeed the definition of Granger-*non*-causality in point processes (Xu et al., 2016). In general, event A is said to be Granger-non-cause of another event B if A does not affect the occurrence probability of event B.

2.2 Instance-Level Granger Causality

Although MHP in Eq. (2) provides Granger causality interpretations for event sequences, such a causality structure is only at the *type-level*, instead of *instance-level*. To obtain instance-level causality, a direct generalization of MHP for an event i with event type k would be

$$\lambda_{i,k}(t \mid \mathcal{H}_t) = \mu_{i,k} + \sum_{j:t_j<t} \alpha_{i,j,k} \phi_{i,j,k}(t - t_j) \quad (3)$$

where $\mu_{i,k}$ is the background intensity for event i with event type k, $\{\alpha_{i,j,k}\}$ forms a $L \times L \times K$ tensor representing instance-level Granger causality, and $\phi_{i,j,k}(\cdot)$ is the decay function representing time-decay of the causal influence. We assume a maximum sequence length $L = \max(\{L_s\})$ and use padding for varying length sequences.

For instance-level causal analysis, the unsupervised causal discovery problem can be reduced to fitting the

model parameters contained in the intensity function $\lambda_{i,k}(t \mid \mathcal{H}_t)$. However, such an analysis is very challenging. First, capturing long-range causality between events can be difficult due to the intricate nature of temporal dependencies. Second, it requires a substantial number of parameters to adequately parameterize the model, increasing the risk of overfitting. In our model, we use a particular parametric form for $\lambda_{i,k}(t \mid \mathcal{H}_t)$ and regularization terms to mitigate the overfitting issue.

3 Related Work

Granger Causality. Granger causality (Granger, 1969a) was initially developed for causal discovery in multivariate time series. Common approaches in Granger causality use linear models such as Vector Autoregressive Model (VAR) with a group lasso penalty (Arnold et al., 2007; Shojaie and Michailidis, 2010). Neural Granger causality (Tank et al., 2021) relaxed the linearity assumption and introduced sparsity regularization in deep neural networks. Löwe et al. (2022) studied an amortized setting and proposed a deep variational model to handle time series with different underlying graphs. However, most of existing studies focus on time-series sampled at a regular time interval. For event sequence data sampled at irregular time stamps, two categories of prior works are directly relevant in Granger causality: Multivariate Hawkes process (MHP) and Neural Point Processes (NPP).

Multivariate Hawkes process. As suggested by Eq. (2), where the kernel matrix depends only on the event types, maximum likelihood estimation with MHP only leads to type-level causal analysis (Mei and Eisner, 2017; Xu et al., 2016; Achab et al., 2017). One approach to deriving instance-level causality is to leverage the MM algorithm (Veen and Schoenberg, 2008; Lewis and Mohler, 2011; Li and Zha, 2015; Wang et al., 2017; Idé et al., 2021), where the instance-level causal strength is defined through a variational distribution of a lower bound of the log likelihood function. One limitation of this approach is that the instance-level causality is defined only through the first term of Eq. (2). This is reminiscent of Cox's partial likelihood approach (Cox, 1975) for the proportional hazard model, and hence, can be viewed as a heuristic. Another limitation is the linearity assumption in the intensity function. In complex domains, where nonlinear causal effects such as synergistic effects may exist, the restrictive parametric form can lead to a subnormal fit to the data.

Neural Point Processes. NPPs combine point processes with deep neural networks. There are mainly two approaches for NPP. The first approach is based on recurrent neural networks (RNNs), while the other leverages the transformer architecture. In both approaches, the intensity function is typically represented as $\lambda_{k_i}(t_i \mid \boldsymbol{h}_{i-1})$ with $\boldsymbol{h}_{i-1}$ being the embedding vector of the event history $\{(t_j, k_j)\}_{j=1}^{i-1}$ (Xiao et al., 2017). The main advantage of the NPP-based approach is that neural networks, as universal sequence approximators, eliminate the need for carefully choosing a specific parametric form for the intensity function. However, they lose reference to individual event instances because of the embedding of $\mathcal{H}_{t_i}$ into $\boldsymbol{h}_{i-1}$. As a result, retrieving instance-level dependencies generally requires additional model assumptions. One approach is to use the self-attention weights as a proxy of causal strength (Xiao et al., 2019; Zuo et al., 2020; Zhang et al., 2020a), and the other is to introduce an ad-hoc dependency score defined independently of the maximum likelihood framework (Zhang et al., 2020b). In both cases, it is not clear how those scores are related to the notion of Granger causality.

While we also employ the transformer architecture, the key difference from the existing works is that our model is designed to *maintain the additive structure* in the intensity function. This guarantees Granger-causality in computing instance-level causal strengths, unlike the self-attention weights in the existing transformer Hawkes models. To the best of our knowledge, `ISAHP` is the first neural point process model that allows direct instance-level causal analysis.

4 Instance-wise Self-attentive Hawkes Processes

In this section we present our **I**nstance-wise **S**elf-**A**ttentive **H**awkes **P**rocesses (`ISAHP`) model.

Notation. Hereafter, vectors and matrices are denoted in bold italic (such as $\boldsymbol{v}$) and sans serif (such as W_V), respectively. We use $^\top$ to denote the transpose of vectors and matrices. All vectors are column vectors. $\mathbb{R}$ and $\mathbb{R}_+$ denote the set of real numbers and non-negative real numbers, respectively.

4.1 Intensity Function

`ISAHP` has two key features. First, it maintains the additive structure over the historical events in the intensity function, similar to MHP. Therefore, `ISAHP` inherits the interpretability of MHP for Granger causality. Second, `ISAHP` adopts an instance-aware parameterization of the kernel function. Specifically, we associate each event with a latent embedding vector $\boldsymbol{x} = g(t, k)$ and define the embedding function $g(t, k)$ as:

$$g(t, k) = \mathrm{MLP}[t - t_i, \mathrm{MLP}(\mathbf{k})]$$

with the event type (as a K-dimensional one-hot vector $\mathbf{k}$) and the time difference (as $t_i - t_{i-1}$) for $\boldsymbol{x}_i$. We use an MLP layer to embed the one-hot vector for event type and concatenate it with the time difference to form a M-dimensional embedding vector.

We assume the intensity function for an event embedding $\boldsymbol{x}$ in the form of:

$$\lambda(\boldsymbol{x} \mid \mathcal{H}_t) = \mu(\boldsymbol{x} \mid \mathcal{H}_t) + \sum_{j:t_j<t} \alpha(\boldsymbol{x}, \boldsymbol{x}_j)\phi(t - t_j \mid \boldsymbol{x}, \boldsymbol{x}_j), \tag{4}$$

where μ is the background intensity, the function $\alpha(\boldsymbol{x}, \boldsymbol{x}_j) \in \mathbb{R}^+$ is the kernel function. The kernel function characterizes the instance-level causal influence between events, generalizing the vanilla MHP, whose kernel matrix depends only on event types. The decay distribution $\phi(t - t_j \mid \boldsymbol{x}, \boldsymbol{x}_j)$ models the time decay of causal influence. In general, ϕ can be any distribution depending on the statistical nature of the training dataset $\mathcal{D}$. In our experiments, we assume a "neural exponential distribution":

$$\phi(t - t_j \mid \boldsymbol{x}, \boldsymbol{x}_j) = \gamma(\boldsymbol{x}, \boldsymbol{x}_j)\mathrm{e}^{-\gamma(\boldsymbol{x}, \boldsymbol{x}_j)(t-t_j)}, \tag{5}$$

where $\gamma(\boldsymbol{x}, \boldsymbol{x}_j)$ is the decay rate function. We use neural networks to model the functions $\gamma(\boldsymbol{x}, \boldsymbol{x}_j)$, $\mu(\boldsymbol{x} \mid \mathcal{H}_t)$ and $\alpha(\boldsymbol{x}, \boldsymbol{x}_j)$, which will be discussed in later sections.

In the Hawkes-type model, event occurrence probability consists of two components: The spontaneous effect (the μ term) and the causal effects (the α term). While it is true that μ includes the elements of A in average/aggregation, the individual causal effects are best captured by explicitly including the α's. Regularized MLE further resolves this issue near-optimally (c.f. (Eichler et al., 2017)) and, hence, admits the reasonable interpretation that $\alpha_{\boldsymbol{x},\boldsymbol{x}_j} = 0$ indicates Granger non-causality at the instance level.

4.2 Self-attentive Architecture

In multi-type event sequences, there may exist short-term temporal dependencies that the classical linear Hawkes models can capture, or non-trivial long-range temporal dependencies involving multiple event instances. To capture such dependencies, we introduce a neural architecture based on self-attention Vaswani et al. (2017) to parametrize Eq. (4). Fig. 1 illustrates the model architecture. The embedding approach follows the standard key-value-query formalism of the transformer.

Self-Attention. The embedding vectors $\{\boldsymbol{x}_j\}_{j=1}^L$ are linearly transformed to be the "value" vector:

$$\boldsymbol{v}_j = \mathsf{W}_V^\top \boldsymbol{x}_j, \quad \text{or} \quad \mathsf{V} = \mathsf{W}_V^\top \mathsf{X}, \tag{6}$$

where $\mathsf{X} \triangleq [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L] \in \mathbb{R}^{M \times L}$ and $\mathsf{V} \triangleq [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_L] \in \mathbb{R}^{M_V \times L}$. $\mathsf{W}_V \in \mathbb{R}^{M \times M_V}$ is learned from the data.

We capture the dependency among the events through "self-attention" $A(\boldsymbol{x}, \boldsymbol{x}_j) \in \mathbb{R}$, defined by

$$A(\boldsymbol{x}, \boldsymbol{x}_j) = \frac{\exp\left(\boldsymbol{x}^\top \mathsf{K} \boldsymbol{x}_j\right) \mathbb{I}(t > t_j)}{\sum_{l:t_l<t} \exp\left(\boldsymbol{x}^\top \mathsf{K} \boldsymbol{x}_l\right)}, \tag{7}$$

where t is the timestamp associated with $\boldsymbol{x}$, and $\mathbb{I}(t > t_j)$ is the indicator function that assumes the value 1 if the argument is true and 0 otherwise. $\mathsf{K} \in \mathbb{R}^{M \times M}$ is a learnable parameter matrix. Following the notation of the transformer, K corresponds to $\mathsf{W}_Q \mathsf{W}_K^\top$, where W_Q and W_K are the transformation matrix for the queries and keys. As shown in Fig. 1, the self-attention weights are represented as an $L \times L$ matrix $\mathsf{A} = [A_{i,j}]$ during training with $A_{i,j} \triangleq A(\boldsymbol{x}_i, \boldsymbol{x}_j)$. For K types of events, we use multi-head attention with K different heads.

4.3 Kernel matrix, background intensity, and decay rate functions

Now we infer $\gamma(\boldsymbol{x}, \boldsymbol{x}_j)$, $\mu(\boldsymbol{x} \mid \mathcal{H}_t)$ and $\alpha(\boldsymbol{x}, \boldsymbol{x}_j)$ in Eq. (4) and Eq. (5) according to the self-attentive architecture. For the background intensity function, we use the following form:

$$\mu(\boldsymbol{x} \mid \mathcal{H}_t) = \bar{\mu}_k + \sigma\left((\boldsymbol{w}_k^\mu)^\top \sum_{j:t_j<t} A(\boldsymbol{x}, \boldsymbol{x}_j)\boldsymbol{v}_j\right). \tag{8}$$

Here, $\bar{\mu}_k$ is the background intensity, and k is the event type encoded in $\boldsymbol{x}$. The second term represents instance-specific effects in the background intensity, where $\sigma(\cdot)$ denotes an activation function. In our implementation, we used the sigmoid function. This term represents the averaged effect of causal interactions among the events. $\boldsymbol{w}_k^\mu \in \mathbb{R}^{M_V}$ is learned from data.

For the impact and decay rate functions, we adopt the following instance-specific formulation:

$$\begin{aligned} \alpha(\boldsymbol{x}, \boldsymbol{x}_j) &= \sigma_+\left(A(\boldsymbol{x}, \boldsymbol{x}_j)(\boldsymbol{w}_k^\alpha)^\top \boldsymbol{v}_j\right), \\ \gamma(\boldsymbol{x}, \boldsymbol{x}_j) &= \sigma_+\left(A(\boldsymbol{x}, \boldsymbol{x}_j)(\boldsymbol{w}_k^\gamma)^\top \boldsymbol{v}_j + b_k^\gamma\right). \end{aligned} \tag{9}$$

where $\sigma_+(\cdot)$ is the softplus function applied elementwise on the vector argument, and the parameter vectors and matrices $\{\boldsymbol{w}_k^\alpha, \boldsymbol{w}_k^\gamma, b_k^\gamma\}_{k=1}^K$ are learned from the data. The input of these MLPs is $A(\boldsymbol{x}, \boldsymbol{x}_j)\boldsymbol{v}_j$, which can be viewed as a relevant component of $\boldsymbol{v}_j$ in terms of the impact on the target event represented by $\boldsymbol{x}$. To see how this model generalizes the vanilla MHP, imagine that $\sigma_+(\cdot)$ were the identity function. Then we have $\alpha(\boldsymbol{x}, \boldsymbol{x}_j) \to A(\boldsymbol{x}, \boldsymbol{x}_j)\tilde{w}_{k,j}^\alpha$, where $\tilde{w}_{k,j}^\alpha \triangleq (\boldsymbol{w}_k^\alpha)^\top \boldsymbol{v}_j$. By construction, the attention weight $A(\boldsymbol{x}, \boldsymbol{x}_j)$ represents

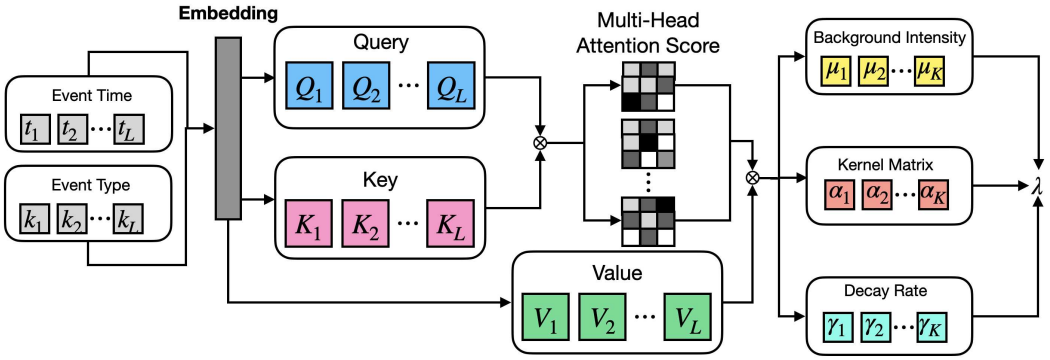

Figure 1: Instance-wise Self-attentive Hawkes Processes (ISAHP) architecture. The input is an event sequence identified by timestamp and event type. the output is an intensity function parameterized by background intensity, kernel matrix, and decay rate.

the similarity between $\boldsymbol{x}$ and $\boldsymbol{x}_j$. On the other hand, $\tilde{w}^{\alpha}_{k,j}$ can be interpreted as the relevance of $\boldsymbol{v}_j$ to type-k events computed through the vector inner product. Compared to MHP's α_{k,k_j}, which depends only on the event types, ISAHP looks at events at a finer granularity by using the embedding vectors.

4.4 Maximum Likelihood Estimation

We learn Eq. (4) based on maximum likelihood estimation. To present the final objective function, we restore the sequence index s hereafter. The main outcome of the unsupervised causal discovery task is $\alpha^s_{i,j} \triangleq \alpha(\boldsymbol{x}^s_i, \boldsymbol{x}^s_j)$, which quantifies the instance-level causal influence of the j-th event on the i-th event. As judged from Eq. (4), $\alpha^s_{i,j} = 0$ meets the definition of Granger-non-causality.

As a side product, the type-level causal dependency, denoted by $\bar{\alpha}_{k,k'}$, can be obtained as the average of instance-level causal influence:

$$\bar{\alpha}_{k,k'} \triangleq \frac{1}{N_{k,k}} \sum_{s=1}^{S} \sum_{i=1}^{L_s} \sum_{j=0}^{i} \delta_{k^s_j,k} \delta_{k^s_i,k} \alpha^s_{i,j}. \qquad (10)$$

where $\delta_{k^s_j,k}$ etc. is Kronecker's delta that is 1 if $k^s_j = k$ and 0 otherwise. $N_{k,k'}$ is the total counts of the event type pair (k, k') in the dataset, defined by $N_{k,k'} \triangleq \sum_{s=1}^{S} \sum_{i=1}^{L_s} \sum_{j=0}^{i} \delta_{k^s_j,k} \delta_{k^s_i,k}$.

The final loss function to be minimized is now given by

$$\mathcal{L} = \sum_{k=1}^{K} \sum_{k'=1}^{K} \left(\omega_1 |\bar{\alpha}_{k,k'}| + \omega_2 \sigma^2_{k,k'}\right) + \sum_{s=1}^{S} \sum_{i=1}^{L_s} \left[\left[\int_{t^s_{i-1}}^{t^s_i} \mathrm{d}t' \lambda(t', \boldsymbol{x}^s_i \mid \mathcal{H}_{t^s_i}) \right] - \ln \lambda(t^s_i, \boldsymbol{x}^s_i \mid \mathcal{H}_{t^s_i}) \right]. \qquad (11)$$

where ω_1, ω_2 are the regularization strengths treated as hyperparameters. In the first term of Eq. (11), we have introduced regularization terms for numerical stability and consistency within the same event type pair. Specifically, the type-level regularization (TLR) term $\omega_1 |\bar{\alpha}_{k,k'}|$ is L_1 regularization on the mean of αs sharing the same event type pair. We also include the variance regularization term (Namkoong and Duchi, 2017; Huang et al., 2020), with $\sigma^2_{k,k'}$ defined as

$$\sigma^2_{k,k'} \triangleq \frac{1}{N_{k,k}} \sum_{s,i,j} \delta_{k^s_j,k} \delta_{k^s_i,k} (\alpha^s_{i,j} - \bar{\alpha}_{k,k'})^2, \qquad (12)$$

to control the variability within the event instances of the same event type pair (k, k'). As we increase the hyperparameter ω_2, ISAHP is encouraged to provide a generative process that is similar to the vanilla MHP for a given decay model. The second term of Eq. (11) corresponds to the negative log-likelihood function, where we used the well-known relationship between the intensity function and the log-likelihood (See, e.g., (Daley et al., 2003)). The integral can be performed analytically for the neural exponential distribution.

5 Experiments

We evaluate Granger causality inference at both the type level and the instance level. We aim to verify that (a) ISAHP outperforms other baselines for the type-level causality discovery task; (b) there is a positive correlation between performances in the type-level Granger causality inference and the instance-level event type prediction, allowing ISAHP to accurately predict the type of next event instance; (c) ISAHP can capture complex synergistic causal effects over multiple event types at the instance level.

5.1 Experimental Set-up

Datasets. For empirical validation, we used two datasets of different sizes: Synergy and Memetracker

(MT). These two datasets were chosen since they contain non-linear causal interactions that are challenging for classical models hence motivating a new solution approach. Out of those benchmark datasets from (Zhang et al., 2020b), they are the only datasets with non-linear causality that require instance-level causality analysis. The details of the data set and statistics are summarized in Appendix 7.2.

Baselines. We compared `ISAHP` with six baselines: Three belong to the category of the classical MHP and three from the NPP family.

- `HExp`: MHP in Eq. (2) with the exponential decay model.
- `HSG`: MHP with a Gaussian mixture decay (Xu et al., 2016), which is known as the state-of-the-art parametric model for Granger causality in classical MHP.
- `CRHG`: A sparse Granger-causal learning framework based on a cardinality-regularized Hawkes process (CRHG) (Idé et al., 2021). Note that CRHG is designed to learn from a single event sequence. To incorporate this baseline into type-level causality analysis, we concatenate sequences from the dataset to form a long sequence.
- `RPPN`: Recurrent Point Process Networks (RPPN) (Xiao et al., 2019). An RNN (recurrent neural network)-based NPP that supports Granger causality inference based on an added attention layer.
- `SAHP`: Self-Attentive Hawkes Process (SAHP) (Zhang et al., 2020a). A transformer-based NPP that enables Granger causality analysis based on the attention mechanism. It directly uses the self-attention from its transformer architecture to aggregate the influence from historical events in determining the intensity function for the next event.
- `CAUSE`: Causality from attributions on sequence of events (CAUSE) (Zhang et al., 2020b). An RNN-based framework for inferring Granger causality. It includes a post-training step to infer the instance-level Granger causality using an attribution method called the integrated gradient. Note that `ISAHP` does not require any post-training step and can directly infer the instance-level Granger causality based on its additive intensity function.

Evaluation Metrics We conducted three different experiments to validate the proposed `ISAHP` in addition to an ablation study to validate TLR. While the main motivation of `ISAHP` is instance-level Granger causal analysis, we include proxy tasks involving type-level inference as well as instance-level event-type prediction, due to the scarcity of ground truth data on instance-level causality.

(1) Type-level Granger causal discovery. We used the area under the curve (AUC) of the true positive vs. false positive curve and Kendall's τ coefficient to measure the accuracy of the inferred Granger causality matrix compared to the ground truth. (2) Next event-type prediction. This can be reduced to a multi-class classification problem given its timestamp. We used the classification accuracy to measure the performance. (3) Instance-level causal discovery. We picked a representative sequence pair involving synergistic causal interactions to highlight qualitative differences from the baselines. We also conducted statistical analysis by measuring the ratio between synergistic and non-synergistic contribution scores. For (1) and (2), we follow the setting of (Zhang et al., 2020b) and report the average results based on five-fold cross-validation.

Implementation Details and Hyperparameter Configurations The `ISAHP` hyperparameter settings for Synergy and MT experiments are shown in Appendix 7.3. These optimal hyperparameter settings were selected based on five-fold cross-validation. We use the Adam optimizer for training. The implementation details for other baselines are in Appendix 7.3.

5.2 Experimental Results

Type-level Causality Analysis We evaluate the performance of type-level Granger causality inference. Table 1 exhibits the accuracy measures for Granger causality inference using AUC and Kendall's τ for `ISAHP` as well as 6 baselines. We see that `ISAHP` generally outperforms all baselines. For AUC, it is the best among all methods. For Kendall's τ, it is the best for the Synergy dataset and is almost tied with the best method (SAHP) for the MT dataset. Note that `ISAHP` always has the smallest variance, indicating that `ISAHP` is the most robust.

For the MHP baselines (HExp, HSG, CRHG), we see that they perform poorly for both Synergy and MT. This is expected to some extent as Synergy involves synergistic effects between multiple causes and MT is based on a real-world dataset including non-linear effects. The underlying data generation mechanisms do not adhere to the linearity assumption of the (type level) intensity functions of these models.

For the two NPP baselines (RPPN, SAHP) that use the self-attention weights for (pseudo) causal attribution, we observe that their performance is quite unstable. SAHP reaches the start-of-the-art performance on MT dataset for Kendall's τ and is the second best on AUC,

Table 1: Results for Granger causality discovery on two datasets with ground-truth causality. For both AUC and Kendall's τ, larger values are better. The result shows that the proposed method, ISAHP, is the most accurate and robust overall.

	Synergy		**MT**	
	AUC	Kendall's τ	AUC	Kendall's τ
HExp	0.885 ± 0.014	0.361 ± 0.013	0.616 ± 0.021	0.061 ± 0.011
HSG	0.306 ± 0.063	-0.182 ± 0.059	0.705 ± 0.015	0.105 ± 0.007
CRHG	0.515 ± 0.033	0.014 ± 0.031	0.611 ± 0.01	0.079 ± 0.006
CAUSE	0.761 ± 0.068	0.244 ± 0.064	0.739 ± 0.042	0.127 ± 0.022
RPPN	0.827 ± 0.017	0.307 ± 0.016	0.437 ± 0.008	-0.031 ± 0.004
SAHP	0.182 ± 0.119	-0.298 ± 0.112	0.832 ± 0.012	**0.251 ± 0.016**
ISAHP	**0.967 ± 0.007**	**0.438 ± 0.006**	**0.835 ± 0.002**	0.247 ± 0.001

Table 2: Prediction accuracy in next-event-type prediction. The higher, the better. ISAHP outperforms all the baselines.

	Synergy	**MT**
HExp	0.349 ± 0.013	0.862 ± 0.004
HSG	0.364 ± 0.009	0.835 ± 0.006
CAUSE	0.37 ± 0.012	0.905 ± 0.004
RPPN	0.364 ± 0.01	0.748 ± 0.056
SAHP	0.343 ± 0.013	0.459 ± 0.033
ISAHP	**0.471 ± 0.008**	**0.974 ± 0.002**

but performs the worst on the Synergy dataset. Similarly, RPPN has a relatively good performance on Synergy but is the worst on the MT dataset. These results indicate that using attention as attribution can be unstable depending on the data characteristics and there is no guarantee on the performance. One key issue is that they do not directly use the self-attention to parameterize the intensity function. Instead, they perform matrix multiplication between the self-attention scores and the value tensor. This step fuses information from the historical events and masks the pairwise causal relationships at the instance level. Similar results have been observed and studied recently (Serrano and Smith, 2019; Jain and Wallace, 2019). This contrasts with our approach that directly uses the attention scores to parameterize the intensity function, which is one of our key contributions.

Note that CAUSE is the second most robust baseline. However, it requires additional computational overhead with the post-training step which makes $O(SK/B)$ invocations of a rather expensive attribution procedure, as we discussed earlier.

Instance-Level Event Type Prediction We consider the event type prediction at the instance level.

In Table 2, we compare all the methods in terms of the classification accuracy score. It is clear that ISAHP performs better than all baselines on both datasets. Specifically, ISAHP reaches 27.3% relative improvement over the second-best method on the Synergy dataset and 7.62% relative improvement on the MT dataset.

Another interesting finding is that although SAHP performs well on the type-level Granger causality discovery for the MT dataset, its event-type prediction accuracy for the same dataset is the worst. A similar phenomenon is observed for RPPN on the Synergy dataset. This is another indication that naively using attention for causal attribution can be unstable.

Instance-level Causality Analysis One of the key advantages of ISAHP is that it can perform accurate *instance-level* causality analysis. Here we present anecdotal evidence of this characteristic, together with statistical analysis. Fig 2 top exhibits two similar event sequences we sampled from the Synergy dataset. Each of them has four events on the timeline. Each event has been assigned a numerical label indicating its event type. The first and second events of the first sequence (on the left) have a synergistic effect on the third event (as indicated by the red square arrow). In contrast, in the second sequence (on the right), the first and second events have causal relationships with the third event, but independently (indicated by the black arrows). To be more precise, the ground truth PGEM model used to generate the data contains a type-level causal relationship $(0 \wedge 1) \to 3$ but not $(0 \wedge 2) \to 3$.

Table 3: Averaged ratio between synergistic and non-synergistic instance-level contribution score on sequences with patterns $'0\#32'$, $'0\#43'$, and $'0\#23'$. A higher ratio means better performance.

	'0#32'	'0#43'	'0#23'
HExp	1	1	1
HSG	1	1	1
SAHP	0.997	0.994	1.025
ISAHP	**1.253**	**1.155**	**1.151**

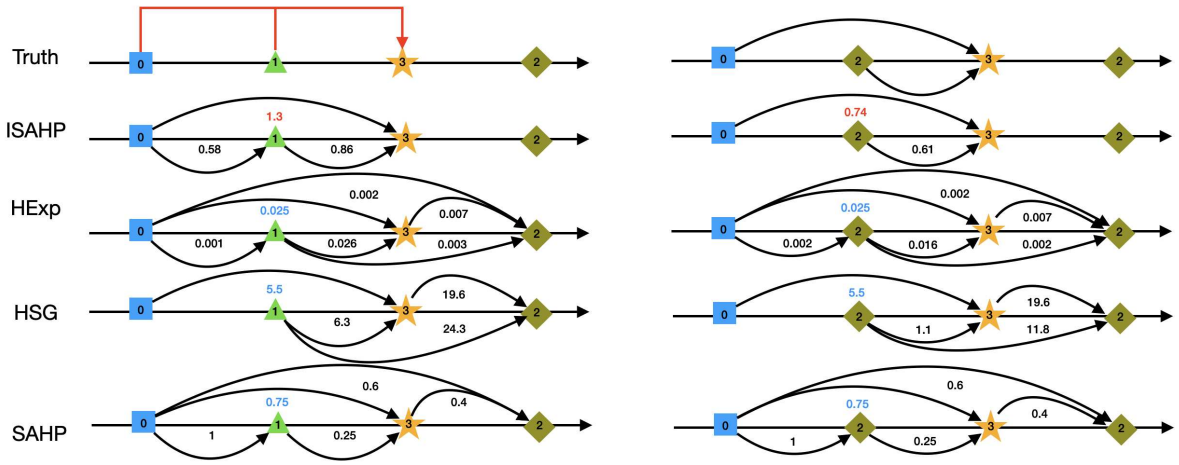

Figure 2: Instance-level causality analysis. The weight of the edge from the first event to the third is what we compare for the synergistic (left) and non-synergistic (right) sequences. Red numbers represent successful cases and blue numbers represent failure cases. `ISAHP` is the only one that successfully captures the synergistic effect at the instance level.

Table 4: Ablation study on type-level regularization.

	Synergy		
TLR	ACC	AUC	Kendall's τ
×	0.344 ± 0.011	0.903 ± 0.127	0.378 ± 0.119
✓	**0.471 ± 0.008**	**0.967 ± 0.007**	**0.438 ± 0.006**
	MT		
×	0.95 ± 0.007	0.814±0.005	0.232±0.004
✓	**0.974 ± 0.002**	**0.835 ± 0.002**	**0.247 ± 0.001**

We would expect an effective causal attribution method to differentiate between the contribution from the first event to the third event under the synergistic and non-synergistic contexts. Fig 2 shows that `ISAHP` successfully assigns larger contribution scores in the synergistic case: The weight of the edge (from the first event to the third) is 1.3 for the synergistic case on the left, while the edge weight for the non-synergistic case on the right is 0.74. On the other hand, all 3 baselines, Hexp, HSG, and SAHP fail. Hexp and HSG are not able to capture the synergistic effects (they assign identical weights in both cases) because they are parameterized to infer Granger causality at the type level. SAHP is intended to infer Granger causality at the instance level, but assigns essentially identical weights in the two cases.

To further verify the superior performance of `ISAHP`, we perform statistical analysis by traversing the dataset to identify all sub-sequences that match the patterns $'0\#32'$, $'0\#43'$, and $'0\#23'$, where event type $\# \in \{0, 1, 2, 3, 4\}$. The synergistic effect occurs only when $\# = 1$. Table 3 shows the ratio of the average inferred instance-level contribution of events with type 0 to events with type 3 in the presence and absence of a synergistic effect. The results show that `ISAHP` always has the optimal performance compared with the baselines. `ISAHP` is able to achieve that because of the tight coupling between the type level and instance level causal learning. In effect, `ISAHP` captures synergistic effects at the type level using its additive structure at the instance level.

Ablation Study Finally, we conducted an ablation study on type-level regularization (TLR) to verify that including TLR does improve ISHAP's performance. For each dataset, we compared TLR and non-TLR cases for both type-level causality analysis and type prediction, based on AUC, Kendall's τ, and accuracy (ACC). The results in Table 4 show that including TLR does improve the model performance for both datasets.

6 Concluding Remarks

We address the problem of discovering instance-level causal structures from asynchronous, interdependent, multi-type event sequences in an unsupervised manner. We proposed a novel self-attentive deep Hawkes model, which enjoys the ability of instance-level causal discovery and the high expressiveness of deep Hawkes

models. It is the first neural point process model that can be directly used for instance-level causal discovery to the best of our knowledge. Our empirical evaluation showed that the proposed instance-aware model significantly improved the performance of type-level tasks as well, suggesting that instance- and type-level causal inference tasks are tightly coupled. One of the important future research topics is to conduct further empirical validation of its performance in instance-level causal analysis, while addressing the challenge due to the scarcity of instance-level ground truth data.

Acknowledgments

This work was conducted as an internship project of Dongxia Wu at IBM Research. Yi-An Ma and Rose Yu are partially supported by the U.S. Army Research Office under Army-ECASE award W911NF-07-R-0003-03, the U.S. Department Of Energy, Office of Science, IARPA HAYSTAC Program, CDC-RFA-FT-23-0069, NSF Grants #2205093, #2146343, #2134274, SCALE MoDL-2134209, and CCF-2112665 (TILOS).

References

Achab, M., Bacry, E., Gaïffas, S., Mastromatteo, I., and Muzy, J.-F. (2017). Uncovering causality from multivariate Hawkes integrated cumulants. *The Journal of Machine Learning Research*, 18(1):6998–7025.

Arnold, A., Liu, Y., and Abe, N. (2007). Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 66–75.

Bacry, E., Bompaire, M., Gaïffas, S., and Poulsen, S. (2017). Tick: a python library for statistical learning, with a particular emphasis on time-dependent modelling. *arXiv preprint arXiv:1707.03003*.

Bhattacharjya, D., Subramanian, D., and Gao, T. (2018). Proximal graphical event models. *Advances in Neural Information Processing Systems*, 31.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Daley, D. J., Vere-Jones, D., et al. (2003). *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer.

Eichler, M., Dahlhaus, R., and Dueck, J. (2017). Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242.

Granger, C. W. (1969a). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

Granger, C. W. J. (1969b). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Hawkes, A. and Chen, J. (2021). A personal history of hawkes process. *Proceedings of the Institute of Statistical Mathematics*, 69(1).

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

Huang, R., Sun, H., Liu, J., Tian, L., Wang, L., Shan, Y., and Wang, Y. (2020). Feature variance regularization: A simple way to improve the generalizability of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4190–4197.

Idé, T., Kollias, G., Phan, D., and Abe, N. (2021). Cardinality-regularized Hawkes-Granger model. *Advances in Neural Information Processing Systems*, 34:2682–2694.

Jain, S. and Wallace, B. C. (2019). Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Lewis, E. and Mohler, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, 1(1):1–20.

Li, L. and Zha, H. (2015). Energy usage behavior modeling in energy disaggregation via marked Hawkes process. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Löwe, S., Madras, D., Zemel, R., and Welling, M. (2022). Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, pages 509–525. PMLR.

Mei, H. and Eisner, J. M. (2017). The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6754–6764.

Namkoong, H. and Duchi, J. C. (2017). Variance-based regularization with convex objectives. *Advances in neural information processing systems*, 30.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Shojaie, A. and Fox, E. B. (2022). Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9:289–319.

Shojaie, A. and Michailidis, G. (2010). Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523.

Tank, A., Covert, I., Foti, N., Shojaie, A., and Fox, E. B. (2021). Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Veen, A. and Schoenberg, F. P. (2008). Estimation of space–time branching process models in seismology using an EM–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624.

Wang, P., Fu, Y., Liu, G., Hu, W., and Aggarwal, C. (2017). Human mobility synchronization and trip purpose detection with mixture of Hawkes processes. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 495–503.

Xiao, S., Yan, J., Farajtabar, M., Song, L., Yang, X., and Zha, H. (2019). Learning time series associated event sequences with recurrent point process networks. *IEEE transactions on neural networks and learning systems*, 30(10):3124–3136.

Xiao, S., Yan, J., Yang, X., Zha, H., and Chu, S. M. (2017). Modeling the intensity function of point process via recurrent neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 1597–1603.

Xu, H., Farajtabar, M., and Zha, H. (2016). Learning Granger causality for Hawkes processes. In *International conference on machine learning*, pages 1717–1726.

Zhang, Q., Lipani, A., Kirnap, O., and Yilmaz, E. (2020a). Self-attentive Hawkes process. In *International Conference on Machine Learning*, pages 11183–11193. PMLR.

Zhang, W., Panum, T., Jha, S., Chalasani, P., and Page, D. (2020b). Cause: Learning Granger causality from event sequences using attribution methods. In *International Conference on Machine Learning*, pages 11235–11245. PMLR.

Zuo, S., Jiang, H., Li, Z., Zhao, T., and Zha, H. (2020). Transformer Hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, volume PMLR 119, pages 11692–11702.

7 SUPPLEMENTARY MATERIAL

7.1 Time Complexity Analysis

ISAHP can directly infer Granger causality at the event instance level through a forward pass during the evaluation phases. Compared with NPP benchmark CAUSE, ISAHP no longer requires a post-training process, which would take $O(SK/B)$ invocations of the (integrated gradient) attribution process, where S is the number of sequences, K is the number of event types, and B is the batch size. ISAHP avoids this computation which can be taxing when the number of sequences is large.

7.2 Statistics of datasets

`Synergy`: A synthetic dataset with complex type-level interaction including a synergistic causal dependency triggered by a pair of event types on the third event type. We used it to test whether ISAHP can reproduce such complex type-level interactions through its pairwise kernel function. The data is generated by a proximal graphical event model (PGEM) (Bhattacharjya et al., 2018). The ground-truth causality matrix is binary and based on the dependency graph of the PGEM simulator.

`MemeTracker (MT)`: A larger-scale real-world dataset, where each sequence represents how a phrase or quote appeared on various online websites over time. We chose the top 25 websites as our event types from August 2008. The ground-truth causality matrix is weighted and approximated by whether one site contains any URL links to another site (Achab et al., 2017; Xiao et al., 2019).

The dataset statistics are summarized in Table 5, where S is the number of sequences, K is the number of event types, L_s is the sequence length for each sequence s.

Table 5: Statistics of datasets used.

Dataset	S	K	$\sum_{s=1}^{S} L_s$	Ground Truth
Synergy	1,000	5	16,101	Binary
MT	8,703	25	90,787	Binary

7.3 Implementation Details and Hyperparameter Configurations

For MHP baselines (HExp, HSG), we use the implementations provided by the tick package (Bacry et al., 2017). For RPPN and CAUSE, the implementation is from (Zhang et al., 2020b). For SAHP, we use the implementation by (Zhang et al., 2020a). The hyperparameters for these baselines were tuned by cross-validation. Specifically, we tuned the learning rate, batch size, and hidden size. The tuned hyperparameter configurations for `ISAHP` are shown in Table 6.

Table 6: Hyperparameter configurations for ISAHP.

Hyperparameter	**Synergy**	**MT**
Learning rate	0.001	0.001
Batch Size	8	16
Hidden Size	10	50
Number of Attention Heads	2	2
ω_1	0.025	0.
ω_2	0.25	5.