THP: Topological Hawkes Processes for Learning Granger Causality on Event Sequences

Ruichu Cai, Member, IEEE, Siyu Wu, Jie Qiao, Zhifeng Hao, Keli Zhang, Xi Zhang

Abstract—Learning Granger causality among event types on multi-type event sequences is an important but challenging task. Existing methods, such as the Multivariate Hawkes processes, mostly assumed that each sequence is independent and identically distributed. However, in many real-world applications, it is commonplace to encounter a topological network behind the event sequences such that an event is excited or inhibited not only by its history but also by its topological neighbors. Consequently, the failure in describing the topological dependency among the event sequences leads to the error detection of the causal structure. By considering the Hawkes processes from the view of temporal convolution, we propose a Topological Hawkes processes (THP) to draw a connection between the graph convolution in topology domain and the temporal convolution in time domains. We further propose a Granger causality learning method on THP in a likelihood framework. The proposed method is featured with the graph convolution-based likelihood function of THP and a sparse optimization scheme with an Expectation-Maximization of the likelihood function. Theoretical analysis and experiments on both synthetic and real-world data demonstrate the effectiveness of the proposed method.

Index Terms—Topological Hawkes Processes, Granger Causality, Event Sequences.

I. INTRODUCTION

Learning the Granger causality among event types on multi-type event sequences is an important task in many real-world applications. For example, the social scientist interest in the study of the causal relationships among the social events [1], the economist interest in analyzing the causality among economic time series [2], and the network operation and maintenance engineers try to locate the root cause of the alarm events [3]. Various methods have been proposed to discover the Granger causality on multi-type event sequences. One line of research is the constraint-based methods, focusing on exploring the independence within the causal variables. The typical methods include PCMCI [4], and the transfer entropy based methods [5]. Another line is the point process based methods, focusing on modeling the generation process of the events. The typical methods include the Multivariate Hawkes processes based methods [6, 7, 8]. Our work belongs to the Multivariate Hawkes processes based methods.

One common assumption of these existing methods is that the event sequences are independent and identically distributed (i.i.d.). However, in many real-world scenarios, the event sequences are usually generated by nodes in a topological network such that an event will not only be excited or inhibited by the event that happens in the sequence itself but also by the event in its topological neighbors. For example, considering the mobile network given in Fig. 1(a), there are three types of alarms $v_1, v_2, v_3$ spreading across the network stations $G_N$ with the causal structure $G_Y$. In such a case, the event sequences in nodes $n_1, n_2, n_3$ are no longer independent, e.g., a cascading failure occurs from one node to another node that nearby. That is, the causal relationship of the failure alarm types $v_1 \rightarrow v_2$ will also appear among the neighbors (the dashed line). Then, if we ignore the topological structure behind the observed event sequences treating the event sequences of different nodes independently, the existing methods fail to detect the causal relationship cross nodes, then may introduce unobservable confounder and return vulnerable and unstable results. As illustrated in Fig. 1(b), if we ignore the topology information and falsely treating the sequences independently, it turns out that a falsely discovered causal relationship $v_2 \rightarrow v_3$ will be made. Specifically, under the i.i.d. assumption, for $v_2$ and $v_3$ in $n_3$, $v_1$ in $n_2$ turns to be an unobservable confounder leading to the false detect of strong dependence between $v_2$ and $v_3$.

Thus, it is crucial to consider the topological structure behind the data for learning Granger causality. However, since the causal influence from an event can complexly propagate through the topological network, modeling the topological information is a non-trivial task. To build this gap, we first reveal the connection between the convolution operation and the Hawkes process, by showing that the intensity of Hawkes processes can be viewed as the convolution in time domain; we further extend the above time domain convolution to the time-graph domain to tackle the topological dependency and finally propose the Topological Hawkes processes (THP). In such a process, the events’ generation process is constrained by the topological structure $G_N$ and the causal structure $G_Y$, as illustrated in Fig. 1(c). We further derive the likelihood function of THP using both the time convolution and the graph convolution. Such a likelihood model is optimized by an Expectation-Maximization based optimization scheme.

As a summary, our contributions of this paper include, 1) proposing a Topological Hawkes Processes (THP) for the event sequences generated by the topological networks, 2) deriving the likelihood function of THP using the joint convolution on the graph-time domain, 3) developing an effective sparse optimization schema for the above likelihood function, and 4) conducting theoretical analysis and extensive experiments on the proposed model.
II. RELATED WORK

Due to space limitation, more related works are given in Appendix A. Most related works are based on the Hawkes process [9], and the extensions mainly focus on the design of regularization and the intensity function. For example, ADM4 [11] uses a nuclear and $l_1$ norm as the sparse regularization term, MLE-SGL [6] further considers the temporal sparsity and pairwise similarity regularization, and NPHC [10] proposes a norm as the sparse regularization term, and the intensity function. For example, ADM4 [11] introduces the attention mechanism to preserve the quasi-causality of events, and CAUSE [12] uses the attribution method combined with a deep neural network to learn the Granger causality. While GHP [13] uses the graph convolution network to model the relation among the parameters of self-exciting Hawkes processes but not model the dependency among the multitask processes. Our work is also related to the PGEM [2] which uses a greedy algorithm to search the causal graph but assumes the event types depend only on the most recent history with a fixed window size.

III. PRELIMINARY

A. Multivariate Point Processes

A multivariate point process is a random process whose realization can be presented by $\mathcal{E} = \{v_i, t_i\}_{i=1}^m$, where $v_i \in \mathbf{V}$ and $t_i \in \mathbf{T}$ are the type and the occurrence time of the $i$-th event, $m$ is the number of events. It can be equivalently represented by a $|\mathbf{V}|$-dimensional counting process $C = \{C_v(t) \mid t \in \mathbf{T}, v \in \mathbf{V}\}$ where $C_v(t) \in \mathbb{N}$ records the number of events that occur before time $t$ in the event type $v$. One way to model the above counting process is to use the following conditional intensity function to characterize the probability of the count of event occurrence given the history:

$$\lambda_v(t) dt = \lambda_v(t) \mathcal{H}_v^V dt = \mathbb{E}[dC_v(t) | \mathcal{F}(t)],$$

where $\mathcal{H}_v^V = \{t_i, v_i \mid t_i < t, v_i \in \mathbf{V}\}$ collects the historical events of all event types before time $t$ and $\mathcal{F}$ is the canonical filtration of all counting processes.

A typical class of the multivariate point processes is Hawkes processes. It considers a particular type of intensity function which measures the excitation from the past event:

$$\lambda_v(t) = \mu_v + \sum_{v' \in \mathbf{V}} \int_{t' \in T_{v'}} \phi_{v',v}(t-t') dC_{v'}(t'),$$

where $T_{v'} = \{t' \mid t' \in \mathbf{T}, t' < t\}$, and the intensity $\lambda_v(t)$ is a summation over the base intensity $\mu_v$ and the endogenous intensity $\sum_{v' \in \mathbf{V}} \int_{t' \in T_{v'}} \phi_{v',v}(t-t') dC_{v'}(t')$ aiming to capture the peer influence occurring near time $t$ [14]. $\phi_{v',v}(t)$ is an impact function characterizing the time-decay of the causal influence.

B. Graph Convolution

Formally, we use an undirected graph $\mathcal{G}_N = (\mathbf{N}, \mathbf{E}_N)$ to represent the topological structure over a set of nodes $\mathbf{N}$ and edges $\mathbf{E}_N$. Let $A$ denotes the adjacency matrix of $\mathcal{G}_N$ in which $A_{ij} = 1$ if $(i, j) \in \mathbf{E}_N$. The normalized Laplacian matrix can be defined as $L = I - D^{-1/2}AD^{-1/2}$, where $D \in \mathbb{R}^{N \times N}$ is the diagonal degree matrix with $D_{ii} = \sum_j A_{ij}$ and $I$ is the identity matrix. As $L$ is a real symmetric positive semidefinite matrix, it has a set of orthonormal eigenvectors $\{u_{(i)}\}_{i=1}^N$ with the orthonormal eigenvectors matrix $U \in \mathbb{R}^{N \times N}$, and the associated eigenvalues $\{\gamma_i\}_{i=0}^{N-1}$ with the diagonalization eigenvalues matrix $\Gamma = \text{Diag}([\gamma_0, \ldots, \gamma_{N-1}])$.

The graph convolution is built upon the spectral convolutions using the graph Fourier transform [15, 16] such that the graph Fourier transform of a signal $s \in \mathbb{R}^N$ can be defined as $\tilde{s} = U^T s$ and its inverse as $s = U \tilde{s}$. Then, graph convolution over $\mathcal{G}_N$ is $(s_1 \ast _2 s_2)_{\mathcal{G}_N} = U((U^T s_1) \odot (U^T s_2))$, where $\odot$ is the element-wise Hadamard product and $s_1, s_2$ are the signal vectors. It follows that a signal $s$ is filtered by $g_\theta$ as:

$$y = g_\theta(L)s = g_\theta(UU^T s) = Ug_\theta(I)U^T s,$$

where $g_\theta$ is the graph convolution kernel.

IV. TOPOLOGICAL HAWKES PROCESS

A. Problem Formalization

In this work, we consider the real-world scenarios that the event sequences are usually generated by nodes in a topological
network such that an event will not only be excited or inhibited by the cause event that happens in the sequence itself but also by the cause event that is in its topological neighbors. To formalize the above scenarios, we use an undirected graph \( G_N = (N, E_N) \) to represent the topological graph among the nodes \( N \) and a directed graph \( G_V = (V, E_V) \) to represent the causal structure among event types \( V \), where \( E_N \) and \( E_V \) are the sets of edges in the topological graph and the causal graph respectively. Based on the above notations, we further extend the traditional event sequences \( E = \{v_i, t_i\}_{i=1}^{m} \) to \( E = \{n_i, v_i, t_i\}_{i=1}^{m} \), by using \( n_i \in N \) to denote the corresponding node in the topological graph. Then, we formalize our problem as follows:

**Definition 1** (Learning Granger causality on topological event sequences). *Given a set of observed event sequences \( E = \{n_i, v_i, t_i\}_{i=1}^{m} \) in \( N, v_i \in V, t_i \in T \) and their corresponding topological graph \( G_N \), the goal of this work is to discover the causal structure \( G_V \) among the event types \( V \).*

### B. Formalization of Topological Hawkes Process

It is easy to see that the existing multivariate Hawkes processes are limited for the above problem. Because the existing multivariate Hawkes processes assume the set of sequences are i.i.d. and ignores the underlying topological structure behind the data generating process, which may lead to false detection as mentioned in Fig. [b]. Thus, by introducing the topological graph into the Hawkes process, we propose the topological Hawkes processes to handle the topological event sequences. However, introducing the topological graph into the Hawkes processes is a non-trivial task, as an event could be excited by its cause event from its topological neighbor through different paths. Such observation implies that the intensity of an event type can be viewed as the summation of the cause event type intensity over different paths, which inspires a way to model it using the graph convolution.

To introduce the graph convolution into the multivariate Hawkes process, we begin with the temporal convolution view of the traditional Hawkes process. From the perspective of temporal convolution, the intensity function of in Eq. [1] can be equivalently formalized as follows:

\[
\lambda_v(t) = \mu_v + \sum_{v' \in V} (\psi_{v', v} \ast dC_v)_{T}(t). \tag{3}
\]

The above formalization reveals that the intensity function of Hawkes processes is essentially a convolution operation in time domain. By extending the convolution in time domain to the join convolution in the graph-time domain \( G_N \times T \), the intensity of the event type \( v \) at node \( n \) in time \( t \) is derived as follows:

\[
\lambda_v(n, t) = \mu_v + \sum_{v' \in V} (dC_{v'} \ast \psi_{v', v})_{G_N \times T}(n, t). \tag{4}
\]

where \( dC_{v'} \) is a function defined on \( G_N \times T \) such that \( dC_{v'}(n, t) \) denotes the count of occurrence events of type \( v' \) event in node \( n \) in time interval \([t - dt, t)\), \( \psi_{v', v} \) is the convolution kernel which measures the causal impact from \( v' \) in both the past and neighbors, and \((\psi_{v', v} \ast dC_v)_{G_N \times T}(n, t)\) depicts the intensity of type \( v \) excited by type \( v' \) events from both graph and time domain.

Practically, by assuming the topological graph invariant across time, such joint convolution operation can be decomposed into two folds:

\[
s_{v', v, t} = (\phi_{v', v} \ast dC_v)_{T}(t), \tag{5}
\]

\[
\lambda_v(n, t) = \mu_v + \sum_{v' \in V} (g_0 \ast s_{v', v, t})_{G_N}(n), \tag{6}
\]

where \( g_0 \) and \( \phi_{v', v} \) are the convolution kernel on the graph domain \( G_N \) and the time domain \( T \), respectively. First, in Eq. (5) \( s_{v', v, t} \) can be viewed as the summation of the impact from the past with the impact function \( \phi_{v', v}(t) \), which is the vanilla endogenous intensity of multivariate Hawkes processes (see Sec. III-A). Second, \( s_{v', v, t} \) will be considered as the signal function in graph convolution. Then, after considering all the impact from cause event types, we have the overall intensity \( \lambda_v(n, t) \) in Eq. (6).

An important question is how to design the convolution kernel. As for the temporal convolution kernel, referring to other works, we set it to \( \phi_{v', v}(t) = \left\{ \begin{array}{ll} \beta_{v', v} \kappa(t) & t > 0 \\ 0 & t \leq 0 \end{array} \right. \) where \( \beta_{v', v} \) is the Granger causal strength of edge \( v' \rightarrow v \) and \( \kappa(t) \) characterize the time-decay of the causal influence usually using the exponential form \( \kappa(t) = \exp(\delta t) \) with the hyper-parameter \( \delta \). As for the graph convolution kernel, recall Eq. (2), the graph convolution in Eq. (6) can be rewritten as \( \lambda_v(n, t) = \mu_v + \sum_{v' \in V} (g_0(\Gamma) \ast (s_{v', v, t}))(n) \). In order to characterize the signal propagate at different topological distances in the graph and reduce the learning complexity, according to [16], we use the following polynomial graph kernel \( g_0(\Gamma) = \sum_{k=0}^{K} \theta_k (I - \Gamma)^k \).

It follows that the intensity function is formalized as:

\[
\lambda_v(n, t) = \mu_v + \sum_{v' \in V} \sum_{n' \in N} \sum_{k=0}^{K} \alpha_{v', v, n', n} \delta_{n', n} \int_{\Gamma}
\kappa(t - \tau) dC_v(n', t'). \tag{7}
\]

where \( \hat{A} = D^{-1/2} A D^{-1/2} \) is the normalized adjacency matrix, and \( \hat{A}_{n, n'} \) is the \( n, n' \) entry of matrix \( \hat{A} \). We further let \( \alpha_{v', v, n} = \beta_{v', v} \theta_k \) represent the causal influence between \( v' \rightarrow v \) from the \( k \)-hop paths and we can directly optimize the parameter \( \alpha \) equivalently. The details of the derivation of the intensity function will be given in Appendix B. Intuitively, \( \hat{A} \) measure the accumulate effect from all length-k closed walks in graph \( G_N \). Therefore, the intensity of type \( v \) events at node \( n' \), time \( t \) in Eq. (7) is the summation over all cause event types from history through all paths in different lengths.

To further illustrate the property of THP, we provide theoretical analyses based on work in [17].

**Theorem 1.** Given the Granger causality graph \( G_V(V, E_V) \) and \( G_N(N, E_N) \). Let \( C \) be the topological Hawkes processes with the intensity function defined in Eq. (7). Then \( v \rightarrow v' \in E_V \) if and only if \( \alpha_{v', v, n} \hat{A}_{n, n'} = 0 \) for all \( k \in \{0, \ldots, K\} \), \( n, n' \in N \), and \( t \in [0, \infty) \).

The complete proof is given in Appendix C. The sketch of the proof is that for the counting process \( C_v(n, t) \) at node \( n \), the
conditional intensity function \( \lambda_v(n, t) \) is \( \mathcal{F}_{v^{-}}(t-^-) \)-measurable if and only if \( \alpha_{v,v',k} A_{v,n}^k = 0 \) for any \( k \in \{0, \ldots, K\}, n', n \in \mathbb{N} \). Thus, Theorem 1 inspires a way to identify the causal relationship by detecting whether the impact \( \alpha_{v,v',k} \) is zero or not.

C. Likelihood of the Topological Hawkes Process

Based on the above theoretical analysis, we know that it is crucial to constraint the sparsity to obtain a reasonable causal structure. Hence, we devise the objective function with sparsity constraint using BIC score under the likelihood framework.

Due to the event sequences are often collected within a period of time in many real-world application, we mainly focus on discrete time scenario in this work. In the discrete time scenario, the minimal time interval is \( \Delta t \). \( \mathbf{T} \) is reduced to the discrete set \( \mathbf{T} = \{0, \Delta t, 2\Delta t, \ldots \} \). \( dC_v(n, t) \) is reduced to the count of occurrence events of type \( v \) event at node \( n \) in time interval \([t- \Delta t, t]\). It follows that we can let \( \mathbf{X} = \{X_{n,v,t} | n \in \mathbb{N}, v \in \mathbf{V}, t \in \mathbf{T} \} \) denote a set of observations of \( dC_v(n, t) \). Given the set of the observations, we can estimate the intensity function as follows:

\[
\lambda_v(n, t) = \mu_v + \sum_{v' \in \text{PA}_v} \sum_{v'' \in \mathbb{N}} \sum_{k=0}^{K} \alpha_{v,v',k} \sum_{t \in \mathbf{T}_v} \hat{A}_{v,v''}^k n \kappa(t - \tau) X_{v',v'',t},
\]

where \( \text{PA}_v \) denotes the set of Granger cause event types of \( v \).

Given Eq. (8), the likelihood can be expressed as the log-likelihood function of \( \Theta \) and the topological structure \( \mathcal{G}_N \) where \( \Theta = (\mu_v, \alpha_{v,v',k})_{v,v',k=0,\ldots,K} \) contains all the parameters of the model. The log-likelihood is given in Eq. (9) and the detail of the derivation will be provided in Appendix D.

\[
L(\mathcal{G}_V, \Theta; \mathbf{X}, \mathcal{G}_N) = \sum_{v \in \mathbf{V}} \sum_{t \in \mathbf{T}} \sum_{n \in \mathbb{N}} \mathbb{P}(X_{n,v,t} \mid H^\text{PA}_v) \left( -\lambda_v(n, t) \Delta t + X_{n,v,t} \log(\lambda_v(n, t)) \right) + \text{Const}
\]

where \( H^\text{PA}_v \) denotes the set of historical events of the cause of the event type \( v \), which are occurred before time \( t \). Note that the log-likelihood function will tend to produce excessive redundant causal edges. Thus, to enforce the sparsity, we introduce the Bayesian Information Criterion (BIC) penalty \( p \log(m) \) into \( L(\mathcal{G}_V, \Theta; \mathbf{X}, \mathcal{G}_N) \), where \( p \) is the number of parameters, and \( m \) is the total number of events that have occurred in \( \mathbf{T} \), according to [18]. The new objective with BIC penalty is given as follows:

\[
L_B(\mathcal{G}_V, \Theta; \mathbf{X}, \mathcal{G}_N) = L(\mathcal{G}_V, \Theta; \mathbf{X}, \mathcal{G}_N) - \frac{p \log(m)}{2},
\]

where the BIC penalty is equivalent to the \( \ell_0 \)-norm of parameters.

D. Sparse Optimization of the Likelihood

In this section, we show how to optimize the above objective function with the \( \ell_0 \) sparsity constraints. Though various methods have been proposed for the sparse problem, e.g., the ADMM method [1] uses a low-rank constraint on the causal graph, the MLE-SGL method [6] further introduces the sparse-group-lasso regularization, their performance still highly depends on the selection of the strength of sparsity regularization or the thresholds for pruning the structure which is not previously known.

To address this problem, we propose the sparse optimization scheme with two steps optimization. Such that the estimation and optimization steps are iteratively conducted to force to learn a sparse causal structure, i.e., estimation step: \( \sup_{\mathcal{G}_V} L_B(\mathcal{G}_V, \Theta; \mathbf{X}, \mathcal{G}_N) \) and searching step: \( \max_{\mathcal{G}_V} \sup_{\Theta} L_B(\mathcal{G}_V, \Theta; \mathbf{X}, \mathcal{G}_N) \). Briefly, in the estimation step, we estimate the BIC score given causal structure \( \mathcal{G}_V \), and following [19], [20], we use an EM-based algorithm to estimate the parameters by optimizing \( \sup_{\Theta} L_B(\mathcal{G}_V, \Theta; \mathbf{X}, \mathcal{G}_N) \). In the searching step, we aim to search the best causal graph \( \mathcal{G}_V \) with the highest score. According to the work [21], such sparse optimization is able to find the truth causal graph due to the consistency property of BIC score.

In detail, in the estimation step, we perform the EM-based algorithm to update the parameters iteratively. The details of the derivation of EM algorithm is given in Appendix E. Specifically, in the \( i \)-th step, the parameters are updated as follows:

\[
\begin{align*}
\lambda^{(i)}_{v,n,t} &= \frac{\sum_{n \in \mathbb{N}} \sum_{t \in \mathbf{T}} q^{(i)}_{n,v,t} X_{n,v,t}}{|\mathbf{N}| |\mathbf{T}| \Delta t}, \\
\mu^{(i)}_v &= \frac{\sum_{n \in \mathbb{N}} \sum_{t \in \mathbf{T}} q^{(i)}_{n,v,t} X_{n,v,t}}{|\mathbf{N}| |\mathbf{T}| \Delta t}, \\
\mathcal{G}^{(i)}_{V} &= \hat{\Theta}^{(i)}
\end{align*}
\]

where \( q^{(i)}_{n,v,t} = \frac{\mu^{(i-1)}_v}{\lambda^{(i-1)}_{v,n,t}} \), and

\[
\alpha^{(i)}_{v,v',k} = \frac{\sum_{n \in \mathbb{N}} \sum_{t \in \mathbf{T}} \left[ \sum_{t' \in \mathbf{T}} \sum_{t'' \in \mathbf{T}} q^{(i)}_{n,v',t'} X_{n,v',t'} \chi_{v''} \hat{A}_{v,v,k}^k \kappa(t - \tau) X_{v',v'',t''} \right] \Delta t}{\sum_{n \in \mathbb{N}} \sum_{t \in \mathbf{T}} \left[ \sum_{t' \in \mathbf{T}} \sum_{t'' \in \mathbf{T}} \hat{A}_{v,v,k}^k \kappa(t - \tau) X_{v',v'',t''} \right] \Delta t},
\]

where

\[
\begin{align*}
\lambda^{(i-1)}_{v,n,t} &= \mu^{(i-1)}_v + \sum_{v' \in \text{PA}_v} \sum_{v'' \in \mathbb{N}} \sum_{k=0}^{K} \alpha^{(i-1)}_{v,v',k} \sum_{t' \in \mathbf{T}_v} \kappa(t - \tau) X_{v',v'',t''}.
\end{align*}
\]

To learn the best Granger causal graph, a Hill-Climbing algorithm is proposed for the best \( \mathcal{G}_V \) with the highest \( \sup_{\Theta} L_B(\mathcal{G}_V, \Theta; \mathbf{X}, \mathcal{G}_N) \) that optimized by the above EM algorithm (Line 7). Note that in case EM might not converge to the global optimum, we can simply try different initial points and choose the best one. In detail, as shown in Algorithm 1, let \( \mathcal{V}(\mathcal{G}_V) \) denote the vicinity of the current causal graph structure, which is obtained by performing a one time operation on \( \mathcal{G}_V \) with edge added, deleted, or reversed operation. Then, at each iteration we evaluate the graphs in \( \mathcal{V}(\mathcal{G}_V) \) to find optimum graph \( \mathcal{G}_V^* \) with the highest score (Line 5-9) and finally, to start over, set \( \mathcal{G}_V = \mathcal{G}_V^* \) (Line 4), repeat until the score no longer increase. In addition, due to the modularity of causal structure, we can locally update the score and the relevant parameters of \( v \) if only \( v \)'s parents has changed.
Algorithm 1 Learning Granger causality using THP with sparse optimization

**Input:** Observation X

**Output:** \( \mathcal{G}_V, \Theta \)

1. \( \mathcal{G}_V^* \leftarrow \text{empty graph}, \ L_B^* \leftarrow -\infty, \ L_B \leftarrow L_B - 1 \)
2. Initialize \( \Theta^* \) randomly
3. while \( L_B < L_B^* \)
4. \( \langle \mathcal{G}_V, \Theta, L_B \rangle \leftarrow \langle \mathcal{G}_V^*, \Theta^*, L_B^* \rangle \)
5. for every \( \mathcal{G}_V' \in \mathcal{V}(\mathcal{G}_V) \) do
6. Initialize \( \Theta_1^* \) randomly;
7. Repeat Update \( L_B' \) and \( \Theta_1^* \) via (11) and (12) Until convergence
8. end for
9. \( \langle \mathcal{G}_V, \Theta_1^*, L_B^* \rangle \leftarrow \langle \mathcal{G}_V', \Theta_1^*, L_B' \rangle \) with largest \( L_B' \)
10. end while
11. return \( \langle \mathcal{G}_V, \Theta \rangle \)

The time complexity of Algorithm 1 is listed below. The worst-case computational complexity of our proposed method is \( O \left( Km^5 |E|^{1/2} |V| + 1 \right) \) and the best-case computational complexity is \( O \left( Km^3 |V|^{2 |E|/|V|} \right) \). The details of the complexity analysis are given in Appendix F.

**V. Experiments**

In this section, we test the proposed THP and the baselines on both the synthetic data and the real-world metropolitan cellular network alarm data. The baselines include PCMCI [4], NPHC [10], MLE-SGL [3], and ADM4 [11]. To study the performance of each component, we also develop two variants of THP, namely THP-NT and THP-S. THP-NT is a variant of THP by removing the topology information. THP-S is another variant of THP by removing the sparse optimization schema but use the causal strength threshold to determine the causal structure following the work of ADM4 and MLE-SGL. The details of these baselines are provided in Appendix G. In all the following experiments, Recall, Precision, and F1 are used as the evaluation metrics.

**A. Synthetic Data**

a) Data Generation: We use the following four steps to generate the multi-type event sequences with topological graph: 1) randomly generate a directed causal graph \( \mathcal{G}_V \) and the undirected topological graph \( \mathcal{G}_N \); 2) generate the events in the root of types (i.e., there are no parents for these root types) using the Poisson process; 3) generate the other type of events according to the causal structure and the topological graph with randomly generated parameters \( \alpha, \mu \). \( \alpha \) denotes the causal strength in the intensity function and \( \mu \) is the base intensity. Specifically, the generated parameters are fixed with default settings and traversed one by one as shown in Fig. 2. The default settings are listed below, the number of nodes = 40, the average degree = 1.5, the number of event types = 20, the sample size = 20000, the range of \( \mu = [0.00005, 0.0001] \), and the range of \( \alpha_{v,v'} = [0.03, 0.05] \). Note that the range of these parameters is selected based on real-world data. All six experiments are given in Fig. 2. The figures of F1, Recall, Precision with variation are given in Appendix H.

b) Comparison with the Baselines: In all six experiments, THP achieves the best results compared with other baselines. Among the baselines, ADM4 and MLE-SGL are relatively better compared with NPHC and PCMCI. The reason is that ADM4 and MLE-SGL have the sparse constraint but others not, which shows the importance of the sparse constraint.

c) Ablation Study: Overall, THP is generally better than both THP_S and THP_NT, which shows the necessity of introducing the topological structure and the sparse optimization. Though THN_S does not use the sparse optimization, it still better than other baselines, which shows the effectiveness of THP.

d) Sensitivity Analysis: For the sensitivity of sample size, Fig. 2(c) shows that a large enough sample size is needed for all methods. In THP, 6000 sample size is enough to achieve relatively good performance while other baselines require at least 15000 or more. For the rest parameters, THP is relatively not sensitive compared with the baselines, which shows the robustness of our method. On the contrary, the baselines are sensitive on either the number of event types or the range of \( \alpha \) and \( \mu \). In particular, THP_S is not sensitive in most cases, which shows the necessity of the topological information.

**B. Real World Data**

a) Data Description and Processing: In this part, we test our algorithms on a very challenging dataset from a real metropolitan cellular network, to find causal structure among 18 types of alarms. It has 3087 network elements, the minimal occurrence times of the alarms are 2000, and there are total 228030 alarms in the data set. In this dataset, we regard the network element as node in the topological graph. The description of the real-world data is provided in Appendix I.

b) Comparison with Baselines: As shown in Table I, our method outperforms the compared methods. However, in terms of recall, the performances of all methods are not well enough and the possible reason is that the time span of the data is only one week and some causal relationships may have relatively weak causal strength which will not be detected. In fact, how to deal with the weak causal strength is still an open problem and we will leave it as future work. As for precision, THP shows a relatively high precision which is important for real-world applications.

### Table I: Results on Real World Data

| Method   | Recall  | Precision | F1     |
|----------|---------|-----------|--------|
| PCMCI    | 0.29787 | 0.32558   | 0.31111|
| NPHC     | 0.34042 | 0.38554   | 0.36158|
| MLE-SGL  | 0.21276 | 0.47619   | 0.29411|
| ADM4     | 0.34042 | 0.49231   | 0.39241|
| THP-NT   | 0.34286 | 0.55385   | 0.42352|
| THP-S    | 0.39362 | 0.49333   | 0.43787|
| THP      | 0.39362 | 0.72549   | 0.51034|
c) Case Study:: To study the necessity of the topological structure in THP, we visualize the causal strength at different topology distance in Fig. 3. As shown in the figure, the causal strength decreases as the distance $k$ increases, which is consistent with our common sense that an event is usually influenced by its neighbors. Another important finding is that some edges only have strong causal strength at $k = 1$ (while not at $k = 0$), such as NCB→NNL and ELD→SSC. It implies that these kind of event types are only triggered by their neighbors. This kind of causality will be missed, if the topological structure is ignored. The operation and maintenance manual also supports our finding. For example, consider ELD→SSC, where SSC indicates that synchronous is changed in S1 mode and ELD indicates that the link connected to an Ethernet port is down. According to the operation and maintenance manual, SSC may be caused by faulty of fiber connection of the alarm, such as ELD, generated by the upstream network elements which are the neighbors.

VI. CONCLUSION

In this work, we propose a topological Hawkes processes for learning Granger causality on event sequences. By extending the convolution in time domain to the joint convolution in graph-time domain, the model takes both the topological constraint and the causal constraint in a unified likelihood framework, and successfully recovers the causal structure among the event types. To the best of our knowledge, this is the first Granger causality learning method for the event sequences with topological constraint. The success of THP not only provides an effective solution for the real-world event sequences, but also shows a promising direction for the causal discovery on the non-i.i.d. samples. In the future, we plan to extend our work to a general point process.

REFERENCES

[1] K. Zhou, H. Zha, and L. Song, “Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes,” in Artificial Intelligence and Statistics, 2013, pp. 641–649.
[2] E. Bacry, I. Mastromatteo, and J.-F. Muzy, “Hawkes processes in finance,” Market Microstructure and Liquidity, vol. 1, no. 01, p. 1550005, 2015.
[3] S.-H. Chang, M. Merabti, and H. Mokhtar, “A causal model method for fault diagnosis in wireless sensor networks,” in 2010 10th IEEE International Conference on Computer and Information Technology. IEEE, 2010, pp. 155–162.
[4] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Jejjilovic, “Detecting and quantifying causal associations in large nonlinear time series datasets,” Science Advances, vol. 5, no. 11, p. eaau4996, 2019.
[5] L. Novelli, P. Wollstadt, P. Mediano, M. Wibral, and J. T. Lizier, “Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing,” Network Neuroscience, vol. 3, no. 3, pp. 827–847, 2019.
[6] H. Xu, M. Farahat, and H. Zha, “Learning granger causality for hawkes processes,” in International Conference on Machine Learning, 2016, pp. 1717–1726.
D. Bhattacharjya, D. Subramanian, and T. Gao, “Proximal graphical event models,” in Advances in Neural Information Processing Systems, 2018, pp. 8136–8145.

F. Salehi, W. Trouleau, M. Grossglauser, and P. Thiran, “Learning hawkes processes from a handful of events,” in Advances in Neural Information Processing Systems, 2019, pp. 12694–12704.

A. G. Hawkes, “Spectra of some self-exciting and mutually exciting point processes,” Biometrika, vol. 58, no. 1, pp. 83–90, 1971.

M. Achab, E. Bacry, S. Gaïffas, I. Mastromatteo, and J.-F. Muzy, “Uncovering causality from multivariate hawkes integrated cumulants,” The Journal of Machine Learning Research, vol. 18, no. 1, pp. 6998–7025, 2017.

S. Xiao, J. Yan, M. Farajtabar, L. Song, X. Yang, and H. Zha, “Learning time series associated event sequences with recurrent point process networks,” IEEE transactions on neural networks and learning systems, vol. 30, no. 10, pp. 3124–3136, 2019.

W. Zhang, T. Panum, S. Jha, P. Chalasani, and D. Page, “Cause: Learning granger causality from event sequences using attribution methods,” in Proceedings of the 37th International Conference on Machine Learning, 2020.

J. Shang and M. Sun, “Geometric hawkes processes with graph convolutional recurrent neural networks,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 4878–4885.

M. Farajtabar, N. Du, M. G. Rodriguez, I. Valera, H. Zha, and L. Song, “Shaping social activity by incentivizing users,” in Advances in neural information processing systems, 2014, pp. 2474–2482.

A. Sandryhaila and J. M. Moura, “Discrete signal processing on graphs,” IEEE transactions on signal processing, vol. 61, no. 7, pp. 1644–1656, 2013.

M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” Advances in neural information processing systems, vol. 29, pp. 3844–3852, 2016.

M. Eichler, R. Dahlhaus, and J. Dueck, “Graphical modeling for multivariate hawkes processes with nonparametric link functions,” Journal of Time Series Analysis, vol. 38, no. 2, pp. 225–242, 2017.

J. Chen, A. Hawkes, E. Scalas, and M. Trinh, “Performance of information criteria used for model selection of hawkes process models of financial data,” SSRN Electronic Journal, 02 2017.

E. Lewis and G. Mohler, “A nonparametric em algorithm for multiscale hawkes processes,” Journal of Nonparametric Statistics, vol. 1, no. 1, pp. 1–20, 2011.

K. Zhou, H. Zha, and L. Song, “Learning triggering kernels for multi-dimensional hawkes processes,” in International Conference on Machine Learning, 2013, pp. 1301–1309.

D. M. Chickering, “Optimal structure identification with greedy search,” Journal of machine learning research, vol. 3, Nov, pp. 507–554, 2002.

Q. Zhao, M. A. Erdogan, H. Y. He, A. Rajaraman, and J. Leskovec, “Seismic: A self-exciting point process model for predicting tweet popularity,” in Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 1513–1522.

R. Zhang, C. Walder, M. A. Rizoiu, and L. Xie, “Efficient non-parametric bayesian hawkes processes,” the 28th International Joint Conference on Artificial Intelligence, 2019.

D. Luo, H. Xu, Y. Zhen, X. Ning, H. Zha, X. Yang, and W. Zhang, “Multi-task multi-dimensional hawkes processes for modeling event sequences,” in Proceedings of the 24th International Conference on Artificial Intelligence, ser. IJCAI’15, 2015, p. 3685–3691.

N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song, “Recurrent marked temporal point processes: Embedding event history to vectors,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1555–1564.

H. Mei and J. M. Eisner, “The neural hawkes process: A neurally self-modulating multivariate point process,” in Advances in Neural Information Processing Systems, 2017, pp. 6754–6764.

APPENDIX A
COMPLEMENTARY TO RELATED WORK

a) Point processes.: There are mainly two types of point processes for the complicated event sequences in recent researches. The first type is the Hawkes processes [9] which assumes that the historical events have influences on the future ones. It uses a parametric or non-parametric intensity function to model the self-exciting and mutually-exciting mechanism of the point process in 1-dimensional [19], [22], [23] and multi-dimensional case [14], [20], [24]. The second type of point processes is the deep-learning-based point process methods [25], [26], [13], which use a learnable intensity function to capture the nonlinear impact from past to future. In THP, graph convolution is introduced to capture the signals from different historical events sequences in a topological graph.

APPENDIX B
DERIVATION OF INTENSITY FUNCTION

Recall Eq. (5) and Eq. (6) in the main article, the intensity function of THP is written as follows:

\[ \lambda_v(n, t) = \mu_v + \sum_{\nu \in V} (g_{\nu \nu} \ast s_{\nu, \nu, t}) \gamma_n(n). \]  

(B.2)

By introducing the graph convolution in Section 3.2, Eq. (B.2) can be rewritten as:

\[ \lambda_v(n, t) = \mu_v + \sum_{\nu \in V} \left( U g_{\nu \nu}(\Gamma) U^T s_{\nu, \nu, t} \right) \gamma_n(n). \]  

(B.3)

Let the graph convolution kernel be \( g_{\nu \nu}(\Gamma) = \sum_{k=0}^{K} \theta_k (I - \Gamma)^k \) and temporal convolution be \( \phi_{\nu, \nu}(t) = \beta_{\nu, \nu} \kappa(t) \), the intensity function can be reformulated as

\[ \lambda_v(n, t) = \mu_v + \sum_{\nu \in V} \left( U \sum_{k=0}^{K} \theta_k (I - \Gamma)^k U^T s_{\nu, \nu, t} \right) \gamma_n(n). \]  

Let the graph convolution kernel be \( g_{\nu \nu}(\Gamma) = \sum_{k=0}^{K} \theta_k (I - \Gamma)^k \) and temporal convolution be \( \phi_{\nu, \nu}(t) = \beta_{\nu, \nu} \kappa(t) \), the intensity function can be reformulated as

\[ \lambda_v(n, t) = \mu_v + \sum_{\nu \in V} \left( U \sum_{k=0}^{K} \theta_k (I - \Gamma)^k U^T s_{\nu, \nu, t} \right) \gamma_n(n). \]  

(B.3)

APPENDIX C
PROOF

The theoretical analyses is based on the work [17]. We start with a recapitulation of the multivariate point processes version of the Granger non-causality.

**Definition 2** (Granger non-causality of multivariate point processes). Let \( C \) be a stationary multivariate point processes
with canonical filtration $\mathcal{F}$. Then we say $C_v$ does not Granger-cause $C_{v'}$ if conditional intensity function $\lambda_v(t)$ is $\mathcal{F}_{v'}(t)$-measurable, where $\mathcal{F}_{v'}(t)$ is the sub-$\sigma$-algebra of the all event types except $v$ that up to but excluding time $t$ corresponding to the sub-processes $C_{v'}$ that excluding $v$.

However, the above definition is only applicable to the multivariate point processes. Thus we first generalize it to the topological Hawkes processes as follows:

**Proposition 1** (Granger non-causality of topological Hawkes process). Let $C$ be a topological Hawkes processes following the causal structure graph $\mathcal{G}_V$ and the topological structure $\mathcal{G}_N$ with canonical filtration $\mathcal{F}$. Then we say $v$ does not Granger cause $v'$ if conditional intensity function $\lambda_v(n, t)$ is $\mathcal{F}_{v'}(t)$-measurable for all $n \in \mathbb{N}, t \in [0, \infty)$, where $\mathcal{F}_{v'}(t)$ is the sub-$\sigma$-algebra of the all event types except $v$ that up to but excluding time $t$ corresponding to all the sub-processes $(C_{v'}(n', t))_{n' \in \mathbb{N}}$ that excluding $v$.

**Proof.** If $v$ does not Granger cause $v'$, then based on Def. 2 for all $n, n' \in \mathbb{N}, C_v(n, t)$ does not Granger cause $C_{v'}(n', t)$. That is, $\lambda_v(n, t)$ is $\mathcal{F}_{v'}(t)$-measurable for all $n \in \mathbb{N}, t \in [0, \infty)$. Here the filtration is extended as the sub-$\sigma$-algebra corresponding to all the sub-processes $(C_{v'}(n', t))_{n' \in \mathbb{N}}$. □

Proposition 1 illustrates the fact that for the process $C_v(n, t)$, the conditional intensity function $\lambda_v(n, t)$ is $\mathcal{F}_{v'}(t)$-measurable if and only if $\alpha_{v', v, k}A^k_{n, n'} = 0$ for any $k \in \{0, \ldots, K\}$, $n, n' \in \mathbb{N}$. Thus the following theorem is straightforward.

**Theorem 1.** Given the Granger causality graph $\mathcal{G}_V(V, E_V)$ and $\mathcal{G}_N(N, E_n)$. Let $C$ be the topological Hawkes processes with the intensity function defined in Eq. (9). Then $v \rightarrow v' \notin E_V$ if and only if $\alpha_{v', v, k}A^k_{n, n'} = 0$ for all $k \in \{0, \ldots, K\}$, $n, n' \in \mathbb{N}$ and $t \in [0, \infty)$.

**Proof.** $\implies$: Note that, if $v \rightarrow v' \notin E_V$, then based on Prop. 1 we have for all node $n$, $\lambda_v(n, t)$ is $\mathcal{F}_{v'}(t)$-measurable. Suppose that there exist one $\alpha_{v', v, k}A^k_{n, n'} \neq 0$ then the corresponding $\lambda_v(n, t)$ must not $\mathcal{F}_{v'}(t)$-measurable, which is contradiction.

$\impliedby$: Suppose that $v \rightarrow v' \notin E_V$ then the intensity $\lambda_v(n, t)$ is $\mathcal{F}_{PA_v}(v(t))$-measurable but not $\mathcal{F}_{PA_{v'}}(v'(t))$-measurable, then there must exist $k \in \{0, \ldots, K\}$, $n, n' \in \mathbb{N}$ such that $\alpha_{v', v, k}A^k_{n, n'} \neq 0$, which is contradiction. □

**APPENDIX D**

**DERIVATION OF LIKELIHOOD FUNCTION**

In the discrete time domain, assuming that the observed event sequences $X$, the likelihood function can be derived as follows:

$$L(G_V, \Theta; X, G_N) = \sum_{v \in V} \sum_{t \in T} \sum_{n \in \mathbb{N}} P(X_{n, v, t} | H_t^{PA_v})$$

$$= \sum_{v \in V} \sum_{t \in T} \sum_{n \in \mathbb{N}} \log \left( \frac{e^{-\lambda_v(n, t)\Delta t}}{N_{n, v, t}!} \lambda_v(n, t) \Delta t \right)^{X_{n, v, t}}$$

$$+ \sum_{v \in V} \sum_{t \in T} \sum_{n \in \mathbb{N}} \left[ -\lambda_v(n, t) \Delta t + X_{n, v, t} \log(\lambda_v(n, t)) \right]$$

$$\lambda_v(n, t) = \mu_v + \sum_{v' \in PA_v} \sum_{n' \in \mathbb{N}} \alpha_{v', v, k}A^k_{n, n'} \sum_{t' \in T_{v'}} \kappa(t - t') X_{n', v', t'}$$

(D.1)

**APPENDIX E**

**DERIVATION OF EM ALGORITHM**

Following the work [19, 20], we utilize the EM algorithm to search the optimal parameters $\Theta$. We first introduce a hidden variables $z_{n, v, t} \in \mathbb{N} \times V \times T \times \{0, \ldots, K\}$, and let $z_{n, v, t} = (n', v', t', k)$ indicate that whether an event with type $v$ in node $n$ at time $t$ is triggered by type $v'$ events in node $n'$ at $t'$, through a path of length $k$. The conditional distribution of $z_{n, v, t}$ given $X_{n, v, t}, H_t^{PA_v}, \Theta^{(i-1)}$ is $p(z_{n, v, t} | X_{n, v, t}, H_t^{PA_v}, \Theta^{(i-1)})$. By modeling such latent variable, the EM algorithm for $\Theta$ consists of two steps. First is the expectation step, which construct the lower bound $Q(\Theta, \Theta^{(i-1)})$ by given the last estimation parameters $\Theta^{(i-1)}$. Second is the maximization step, which maximum $Q$ to obtain the best estimation in $i$ step $\Theta^{(i)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(i-1)})$.

Generally, we first randomly initialize the parameters $\Theta^{(0)}$, then perform the following E-M steps iteratively until $Q$ converges. The derivation of the updated rule in E-step and M-step are given as follows:
Expectation step:
In the E step, we construct the lower bound for the last parameter estimation $\Theta^{(i-1)}$ as follows:

$$Q(\Theta, \Theta^{(i-1)}) = \sum_{n \in \mathcal{N}} \sum_{v \in \mathcal{V}} \sum_{t \in \mathcal{T}} \left[ \sum_{n_0 \in \mathcal{N}_n} \sum_{v_0 \in \mathcal{V}_n} \sum_{t_0 \in \mathcal{T}_n} \sum_{k=0}^{K} q_{n,v,t}(n', v', t', k) \right. $$

$$ \times \log p(X_{n,v,t}, z_{n,v,t} = (n', t', v, t', k), H_t^{PA_V}, \Theta) $$

$$ + \sum_{n \in \mathcal{N}} \sum_{v \in \mathcal{V}} \sum_{t \in \mathcal{T}} \left[ q_{n,v,t}^\mu \log p(X_{n,v,t}, z_{n,v,t} = (n, v, t, 0), H_t^{PA_V}, \Theta) \right] $$

$$ = \sum_{n \in \mathcal{N}} \sum_{v \in \mathcal{V}} \sum_{t \in \mathcal{T}} \left[ \sum_{n_0 \in \mathcal{N}_n} \sum_{v_0 \in \mathcal{V}_n} \sum_{t_0 \in \mathcal{T}_n} \sum_{k=0}^{K} q_{n,v,t}^\mu \log p(X_{n,v,t}, z_{n,v,t} = (n, v, t, 0), H_t^{PA_V}, \Theta) \right] $$

$$ \times X_{n,v,t} \log(\alpha_{v',v, t, t'} \hat{A}_{n,v,t}^\mu \kappa(t-t')X_{n', v', t'}) $$

$$ + \sum_{n \in \mathcal{N}} \sum_{v \in \mathcal{V}} \sum_{t \in \mathcal{T}} q_{n,v,t}^\mu \log \mu_v $$

$$ - \sum_{n \in \mathcal{N}} \sum_{v \in \mathcal{V}} \sum_{t \in \mathcal{T}} \lambda_v(n, t) \Delta t + \text{Const},$$

(E.1)

where $q_{n,v,t}^\mu = p(z_{n,v,t} = (n, v, t, 0) | X_{n,v,t}, H_t^{PA_V}, \Theta^{(i-1)})$ indicates the probability of event $v$ occurred spontaneously in node $n$ at $t$, i.e.,

$$q_{n,v,t}^\mu = p(z_{n,v,t} = (n, v, t, 0) | X_{n,v,t}, H_t^{PA_V}, \Theta^{(i-1)}) = \frac{\mu_v^{(i-1)}}{\lambda_v^{(i-1)}(n, t)},$$

(E.2)

Maximization step:
In M step, we aim to find a $\Theta^{(i)}$ that maximizes $Q(\Theta^{(i)}, \Theta^{(i-1)})$. By using the KKT condition:

$$\begin{align*}
\frac{\partial Q(\hat{\Theta}^{(i)}, \hat{\Theta}^{(i-1)})}{\partial \mu_v^{(i)}} &= 0 \\
\frac{\partial Q(\hat{\Theta}^{(i)}, \hat{\Theta}^{(i-1)})}{\partial \alpha_{v',v,k}^{(i-1)}} &= 0
\end{align*}$$

(E.5)

we obtain the close form solution of the maximum values:

$$\mu_v^{(i)} = \frac{\sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} q_{n,v,t}^\mu X_{n,v,t}}{|\mathcal{N}| |\mathcal{T}| \Delta t},$$

(E.6)

$$\alpha_{v',v,k}^{(i)} = \frac{\sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} q_{n,v,t}^\mu \alpha_{v',v,k}^{n,v,t} X_{n,v,t}}{\sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \sum_{n_0 \in \mathcal{N}_n} \sum_{v_0 \in \mathcal{V}_n} \sum_{t_0 \in \mathcal{T}_n} \hat{A}_{n,n}^{n,v,t}(t-t')X_{n', v', t'}} \Delta t$$

(E.7)

APPENDIX F

COMPUTATIONAL COMPLEXITY OF TWO STEPS

OPTIMIZATION ALGORITHM

In this section, we will analyze the complexity of Algorithm 1. The analysis will be conducted on estimation step and the maximization step respectively. In the estimation step, the computational complexity mainly depends on the EM algorithm. By using the constant iterative step of EM, given a causal graph $G_V$, the complexity is $O(Km^3c)$, where $K$ represents the farthest topological distance considered in the intensity function, $m$ is the number of occurring events and $c$ denotes the number of edges in $G_V$. Note that if $\kappa$ satisfies $\kappa(t_1 + t_2) = \kappa(t_1)\kappa(t_2)$, the complexity of EM decreases to $O(Km^2c)$ because the intensity at each occurred event can be computed using only the intensity from the previous occurred event. In fact, in our implementation, the complexity is $O(Km^2c)$ because of the exponential kernel. For general propose, the following analysis does not include the special kernel.

In the searching step, because the initial graph is empty. That is the number of edges will start at $e = 1$, then the time complexity highly depends on the number of parents at each event type. With the assumption of consistency score, the total steps can be considered as the number of edges that need to be added, which is denoted by $E$. Then, the worst time complexity is the case that all edge are concentrated in certain event types. Note that each step in the searching will need to apply the estimation step, thus we have $O\left(\frac{E}{|V|} \sum_{e=1}^{E} \sum_{V} |V||Km^3c| \right) = O\left(Km^3E|V||V|^{4} \right)$. Note that the complexity of adding one edge in the worst case is $O(Km^3|V||V|c)$ because we only need to update the score on the type that parents changed and all the scores on other types will be cached. Similarly, for the best case, each causal relationship are evenly distributed to each event type such that the number of parent of each type is $\lceil \frac{E}{|V|} \rceil$, and we have the best case time complexity $O\left(\frac{E}{|V|} \sum_{e=1}^{E} \sum_{V} |V||Km^3c| \right) = O\left(Ke |V|^2 \right)$.

APPENDIX G

DETAILS OF BASELINE METHODS

In this section, we briefly introduce each baseline method and its corresponding settings.

A. THP-NT & THP-S

THP-NT and THP-S are two variations of THP. THP-NT is a version of THP that do not consider the topological structure of nodes. The intensity function of THP-NT is given in Eq.
where the impact function does not depend on nodes. The same with THP, THP-NT use two steps based searching algorithm to learn the causal graph.

\[
\lambda_v(n, t) = \mu_v + \sum_{v' \in V} \sum_{t' \in T_v} \alpha_{v', v}(t - t') X_{n, v', t'}, \quad (G.1)
\]

THP-S takes topology information into intensity function likes THP but does not use the sparse optimization scheme. Similar to ADM4, THP-S first initialize causal structure as a complete graph. Then an EM algorithm is applied to optimize the parameters with a \( \ell_1 \) norm regularization on \( \alpha_{v', v, k} \).

\[
\lambda_v(n, t) = \mu_v + \sum_{v' \in V} \sum_{t' \in T_v} \sum_{k=0}^{K} \alpha_{v', v, k}(t - t') \tilde{X}_{n, v', t'} \quad \text{(G.2)}
\]

In all THP and its variations methods, the decay kernel function \( \kappa(t) \) is set to the exponential form \( \kappa(t) = \exp(-\delta(t)) \). In real-world data experiments, according to the alarms table, multiple alarms usually occur within a short time period. Thus, to focus on the short term effects, the decay parameters \( \delta \) is set to 0.11.

B. PCMCI

PCMCI [4] is a causal discovery algorithm for time series data based on conditional independence test. It consists of two stages: 1) PC₁ conditional selection to remove the irrelevant conditions for each variables by iterative independence testing and 2) use the momentary conditional independence (MCI) test to test whether the causal direction holds. Using different types of conditional independence test, this method can be applied to different types of time series data. Here, PCMCI is applied to discover the causal relationship in event sequences. In our experiments, we use the mutual information as the independence test for the discrete event type data, and set the time interval to 5 seconds with max lag up to 2.

C. ADM4

As in the intensity function, ADM4 [1] uses \( \alpha_{v', v, \beta} \exp(-\beta(t - t')) \) to represent the impact function \( \phi_{v', v} \) capturing the causal strength \( \alpha_{v', v} \) from \( v' \) to \( v \). To further constraint the low-rank structure of the causal graph, the Nuclear and \( \ell_1 \)-norm regularization is used. In detail, the nuclear norm \( \|A\|_1 \) is the sum of the singular value \( \sum_{i=1}^{\|V\|} \sigma_i \), where \( A \in \mathbb{R}^{\|V\| \times \|V\|} \) is the matrix of \( \alpha_{v', v} \), and \( \ell_1 \) norm of \( A \) is \( ||A||_1 = \sum_{v' \in V} ||\alpha_{v', v}||_1 \). In our experiments, following the original work, the coefficients of \( \ell_1 \) and nuclear regularization term are both set to 500 and the parameter \( \beta \) of impact function is 0.1.

\[
\lambda_v(t) = \mu_v + \sum_{v' \in V} \int_{t_0}^{t} \phi_{v', v}(t - t') dC_{v'}(t') 
= \mu_v + \sum_{v' \in V} \int_{t_0}^{t} \alpha_{v', v, \beta} \exp(-\beta(t - t')) dC_{v'}(t') \quad \text{(G.3)}
\]

D. NPHC

NPHC [10] a nonparametric method that can estimate the matrix of integrated kernels of a multivariate Hawkes process. This method relies on the matching of the integrated order 2 and order 3 empirical cumulants to recover the Granger causality matrix \( G_{V} \).

E. MLE-SGL

In MLE-SGL [6], the impact function \( \phi_{v', v} \) in the intensity is represented via a linear combination of basis functions \( \kappa_{m} \) (sinc or Gaussian function), which is shown in \( \text{(G.4)} \).

\[
\lambda_v(t) = \mu_v + \sum_{v' \in V} \int_{t_0}^{t} \phi_{v', v}(t - t') dC_{v'}(t') = \mu_v + \sum_{v' \in V} \int_{t_0}^{t} \sum_{m=1}^{M} \alpha_{v', v, m}^m \kappa_{m}(t - t') dC_{v'}(t') \quad \text{(G.4)}
\]

APPENDIX H

DETAILED OF THE SENSITIVITY EXPERIMENTS

The main paper has shown the F1 scores in all six experiments of synthetic data. Here, we further provide F1, Precision and Recall in these experiments with the variants at each experimental setting. The results are given in Fig. 1, Fig. 2 and Fig. 3. All experiments are conducted on Intel(R) Xeon(R) CPU ES-2620 v4 @ 2.10GHz with 64G RAM.

APPENDIX I

REAL-WORLD DATA DESCRIPTION

The real-world data used in this paper are the collection of the alarms records that occurred in a metropolitan cellular network within a week. This data set includes an alarm table and a topological table of network elements. We will release the data set after this work is accepted.

A. Alarm Table

The alarm table consists of 4655592 records, containing 5 fields including 'Alarm Name', 'First Occurrence', 'Cleared On', 'Last Occurrence', and 'Alarm Source'. The meaning of the fields is as follows:

- 'Alarm Name': The name of alarm type;
- 'First Occurrence': The time when the alarm first occurred;
- 'Last Occurrence': The time when the alarm last occurred;
- 'Cleared On': The time when the alarm is cleared;
- 'Alarm Source': The id of the network element where the alarm occurred.
There are 172 types of alarms in the data set. Some typical alarms are as follows: 'ETH_LOS' alarm indicates the loss of connection on an optical Ethernet port, 'MW_LOF' alarm indicates loss of microwave frames and 'HARD_BAD' alarm indicates that the hardware is faulty.

B. Topological Table

The topological table contains 4 fields including 'Path ID', 'NE_NAME', 'NE_TYPE', and 'PATH_HOP'. The specific meanings of the fields are as follows:

- 'Path ID': The id of the path in the metropolitan cellular network;
- 'NE_NAME': The name of the network element, which corresponds to the 'Alarm Source' in the alarm table;
- 'NE_TYPE': The type of network element in the path, including ROUTER, MICROWAVE, and NODEB;
- 'PATH_HOP': The relative position of the network element in the path.

Besides, there are 41143 network elements involved in the topological table.

C. Ground Truth

The ground truth of the causal relationship among alarm types are provided by domain experts. Thus, in our experiment, we only select the sample whose alarm type is labeled by expert. Moreover, since there are large number of network elements involved in this data set but the time span is small, some of the alarm types might suffer from data insufficiency which may lead to unreliable results. To avoid this problems, we further filter the alarm types with more than 2000 occurrences, such that 18 types of alarms and 3087 network elements are involved. The ground truth of the causal relationships among the alarm types are given in Table III.
Table II: Ground Truth

| Cause                        | Effect                        | Cause                        | Effect                        |
|------------------------------|-------------------------------|------------------------------|-------------------------------|
| MW_RDI                       | LTI                           | MW_BER_SD                    | LTI                           |
| MW_RDI                       | CLK_NO_TRACE_MODE             | MW_BER_SD                    | S1_SYN_CHANGE                 |
| MW_RDI                       | S1_SYN_CHANGE                 | MW_BER_SD                    | PLA_MEMBER_DOWN               |
| MW_RDI                       | LAG_MEMBER_DOWN               | MW_BER_SD                    | MW_RDI                        |
| MW_RDI                       | PLA_MEMBER_DOWN               | MW_BER_SD                    | MW_LOF                        |
| MW_RDI                       | ETH_LOS                       | MW_BER_SD                    | NE_COMMU_BREAK                |
| MW_RDI                       | ETH_LINK_DOWN                 | MW_BER_SD                    | R_LOF                         |
| MW_RDI                       | NE_COMMU_BREAK                | R_LOF                         | S1_SYN_CHANGE                 |
| MW_RDI                       | R_LOF                         | R_LOF                         | LAG_MEMBER_DOWN               |
| TU_AIS                       | LTI                           | R_LOF                         | PLA_MEMBER_DOWN               |
| TU_AIS                       | CLK_NO_TRACE_MODE             | R_LOF                         | ETH_LINK_DOWN                 |
| TU_AIS                       | S1_SYN_CHANGE                 | R_LOF                         | NE_COMMU_BREAK                |
| RADIO_RSL_LOW                | LTI                           | R_LOF                         | CPU_STATUS                    |
| RADIO_RSL_LOW                | S1_SYN_CHANGE                 | R_LOF                         | DECADE                         |
| RADIO_RSL_LOW                | LAG_MEMBER_DOWN               | R_LOF                         | PERIOD                         |
| RADIO_RSL_LOW                | PLA_MEMBER_DOWN               | R_LOF                         | TIME                            |
| RADIO_RSL_LOW                | MW_RDI                        | R_LOF                         | LTI                            |
| RADIO_RSL_LOW                | MW_LOF                        | HARD_BAD                      | S1_SYN_CHANGE                 |
| RADIO_RSL_LOW                | MW_BER_SD                     | HARD_BAD                      | S1_SYN_CHANGE                 |
| RADIO_RSL_LOW                | ETH_LINK_DOWN                 | HARD_BAD                      | LAG_MEMBER_DOWN               |
| RADIO_RSL_LOW                | NE_COMMU_BREAK                | HARD_BAD                      | PLA_MEMBER_DOWN               |
| RADIO_RSL_LOW                | R_LOF                         | HARD_BAD                      | ETH_LOS                       |
| BD_STATUS                    | S1_SYN_CHANGE                 | HARD_BAD                      | MW_RDI                        |
| BD_STATUS                    | LAG_MEMBER_DOWN               | HARD_BAD                      | MW_LOF                        |
| BD_STATUS                    | PLA_MEMBER_DOWN               | HARD_BAD                      | NE_COMMU_BREAK                |
| BD_STATUS                    | ETH_LOS                       | HARD_BAD                      | R_LOF                          |
| BD_STATUS                    | MW_RDI                        | HARD_BAD                      | CPU_STATUS                    |
| BD_STATUS                    | MW_LOF                        | HARD_BAD                      | DECADE                         |
| BD_STATUS                    | ETH_LINK_DOWN                 | HARD_BAD                      | PERIOD                         |
| BD_STATUS                    | RADIO_RSL_LOW                 | HARD_BAD                      | TIME                            |
| BD_STATUS                    | TU_AIS                        | HARD_BAD                      | S1_SYN_CHANGE                 |
| NE_COMMU_BREAK               | LTI                           | ETH_LOS                       | LTI                            |
| NE_COMMU_BREAK               | CLK_NO_TRACE_MODE             | ETH_LOS                       | S1_SYN_CHANGE                 |
| NE_COMMU_BREAK               | S1_SYN_CHANGE                 | ETH_LOS                       | PLA_MEMBER_DOWN               |
| NE_COMMU_BREAK               | LAG_MEMBER_DOWN               | ETH_LOS                       | MW_RDI                        |
| NE_COMMU_BREAK               | PLA_MEMBER_DOWN               | ETH_LOS                       | MW_LOF                        |
| NE_COMMU_BREAK               | ETH_LINK_DOWN                 | ETH_LOS                       | NE_COMMU_BREAK                |
| NE_COMMU_BREAK               | NE_NOT_LOGIN                  | ETH_LOS                       | MW_RDI                        |
| ETH_LINK_DOWN                | LTI                           | ETH_LOS                       | MW_LOF                        |
| ETH_LINK_DOWN                | CLK_NO_TRACE_MODE             | ETH_LOS                       | MW_RDI                        |
| ETH_LINK_DOWN                | S1_SYN_CHANGE                 | ETH_LOS                       | MW_LOF                        |
| S1_SYNCHANGE                 | LTI                           | ETH_LOS                       | MW_LOF                        |
| POWER_ALM                    | BD_STATUS                     | ETH_LOS                       | MW_RDI                        |
| POWER_ALM                    | ETH_LOS                       | ETH_LOS                       | MW_RDI                        |
| POWER_ALM                    | MW_RDI                        | MW_RDI                        | MW_RDI                        |
| POWER_ALM                    | MW_LOF                        | MW_RDI                        | MW_RDI                        |

Fig. 3. Recall in the Sensitivity Experiments with Variance