# Causal Structure Learning in Hawkes Processes with Complex Latent Confounder Networks

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Multivariate Hawkes process provides a powerful framework for modeling temporal dependencies and event-driven interactions in complex systems. While existing methods primarily focus on uncovering causal structures among observed subprocesses, real-world systems are often only partially observed, with latent subprocesses posing significant challenges. In this paper, we show that continuous-time event sequences can be represented by a discrete-time causal model as the time interval shrinks, and we leverage this insight to establish necessary and sufficient conditions for identifying latent subprocesses and the causal influences. Accordingly, we propose a two-phase iterative algorithm that alternates between inferring causal relationships among discovered subprocesses and uncovering new latent subprocesses, guided by path-based conditions that guarantee identifiability. Experiments on both synthetic and real-world datasets show that our method effectively recovers causal structures despite the presence of latent subprocesses.

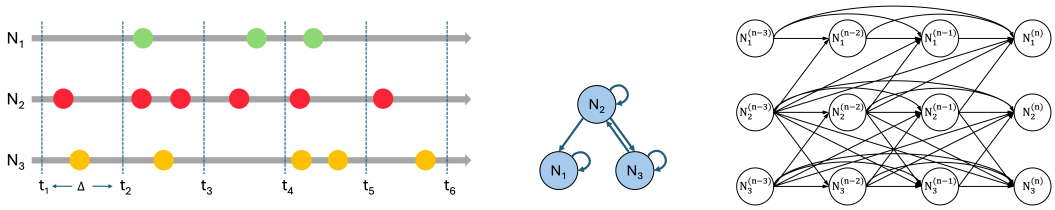
## 1 Introduction

Causal discovery in complex systems is crucial in domains such as social networks [57], neuroscience [4], and finance [20]. Multivariate Hawkes processes [19, 31] have become a powerful tool for modeling temporal dependencies and event-driven interactions. Most existing methods [52, 14, 39, 24] rely on Granger causality [26] and maximum likelihood estimation [48], or on pre-binned likelihood approaches [42, 6, 37]. However, these methods operate under the *sufficiency assumption* that all task-relevant subprocesses are observed. In practice, many components remain unmeasured (e.g., unrecorded neurons in spike train data [22]), creating latent confounders that hinder reliable causal discovery. Existing strategies for missing data [40] do not identify *entirely unobserved* subprocesses, making this an important open challenge. A detailed review is deferred to Appendix A.

In this work, we address the largely unexplored problem of learning causal structures in Hawkes processes with latent subprocesses. Our framework leverages a discrete-time representation and rank constraints on cross-covariance matrices to enable both causal discovery and latent subprocess identification. Specifically, we contribute:

- A principled framework for identifying latent subprocesses without prior knowledge of their existence or number.
- Necessary and sufficient conditions linking discretized Hawkes representations to causal influence, enabling discovery of both observed and latent subprocesses.
- A two-phase iterative algorithm that alternates between structure recovery and latent subprocess discovery, with practical identifiability guarantees.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.




(a) (b) (c)

Figure 1: Illustration of a multivariate Hawkes process with three subprocesses $N_1, N_2, N_3$. (a) A point process representation, where the continuous timeline is partitioned into intervals of length $\Delta$. (b) The corresponding summary causal graph, which is the central object of this paper. Each node represents a subprocess, with causal relations $N_1 \leftarrow N_2 \rightleftharpoons N_3$, and self-loops on all nodes. (c) The window causal graph, depicting the underlying time-lagged causal mechanism. Each node denotes the count in a time interval $\Delta$, modeled as a weighted sum of lagged parent nodes and an uncorrelated noise, as in Eq. 2. **Note**: This paper focuses on cases where some subprocesses are latent.

# 2 Partially Observed Multivariate Hawkes Process-based Causal Model

## 2.1 Multivariate Hawkes Process

A multivariate Hawkes process is a self-exciting point process modeling temporal dependencies among events via a set of counting processes $\mathcal{N}_\mathcal{G} = \{N_i\}_{i=1}^{l}$, where $N_i(t)$ records the number of type-$i$ events up to time $t$.

**Definition 2.1** (Multivariate Hawkes Process [19, 31])**.** For each $i \in \{1, \ldots, l\}$, the intensity of $N_i$ is

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{l} \int_0^t \phi_{ij}(t-s)\, dN_j(s), \tag{1}$$

where $\mu_i$ is the background rate and $\phi_{ij}(s) \geq 0$ the excitation kernel measuring the decaying influence of past type-$j$ events on $N_i$. Stationarity requires the spectral radius of $\Phi_{ij} = \int_0^\infty \phi_{ij}(s)ds$ to be less than one.

For each subprocess $N_i$, we define its *parent cause set* $\mathcal{P}_\mathcal{G} \subseteq \mathcal{N}_\mathcal{G}$ as the minimal set such that $\lambda_i(t)$ depends only on histories of $\mathcal{P}_\mathcal{G}$ and not on others. Equivalently, $N_i$ is *locally independent* [12] of $\mathcal{N}_\mathcal{G} \setminus \mathcal{P}_\mathcal{G}$ given $\mathcal{P}_\mathcal{G}$. Further background and derivations appear in Appendix B.

## 2.2 Model Definition

We formalize our framework as a graphical causal model for multivariate Hawkes processes, where nodes represent subprocesses and directed edges correspond to nonzero excitation functions. The goal is to recover both observed and latent subprocesses and their causal relations.

**Definition 2.2** (Partially Observed Multivariate Hawkes Process-based Causal Model (PO-MHP))**.** Let $\mathcal{G} = (\mathcal{N}_\mathcal{G}, \mathcal{E}_\mathcal{G})$ be a directed graph, where each node $N_i \in \mathcal{N}_\mathcal{G}$ represents a subprocess. A directed edge $E_{ij}$ exists iff $\int_0^t \phi_{ij}(t-s)\, dN_j(s) > 0$. The node set consists of $p$ observed nodes $\mathcal{O}_\mathcal{G} = \{O_i\}_{i=1}^{p}$ and $q$ latent nodes $\mathcal{L}_\mathcal{G} = \{L_i\}_{i=1}^{q}$.

The PO-MHP model naturally allows cycles and self-loops, as well as edges between observed and latent subprocesses.

**Definition 2.3** (Causal Effect)**.** For any $N_i, N_j \in \mathcal{N}_\mathcal{G}$, if a directed path exists from $N_i$ to $N_j$, then $N_i$ is a *cause* of $N_j$ and $N_j$ is an *effect* of $N_i$.

**Definition 2.4** (Parent Cause Set)**.** For $N_i \in \mathcal{N}_\mathcal{G}$, the minimal set $\mathcal{P}_\mathcal{G} \subseteq \mathcal{N}_\mathcal{G} \setminus \{N_i\}$ is called its *parent cause set* if every directed path to $N_i$ passes through some node in $\mathcal{P}_\mathcal{G}$. If $N_i$ has a self-loop, it is also included in $\mathcal{P}_\mathcal{G}$.

**Remark.** $N_i$ is locally independent of $\mathcal{N}_\mathcal{G} \setminus \mathcal{P}_\mathcal{G}$ given $\mathcal{P}_\mathcal{G}$ if and only if $\mathcal{P}_\mathcal{G}$ is its parent cause set.

# 3 Structure Identification in Partially Observed Hawkes Processes

## 3.1 From Continuous-Time to Discrete-Time Representation

Directly inferring causal structure from the continuous-time formulation in Eq. 1 is difficult, especially with latent subprocesses. Instead of MLE-based approaches [52, 14, 39, 24, 42, 6, 37], we establish an explicit reduction from Hawkes dynamics to a discrete-time linear autoregressive model, which enables time-aware rank tests for causal discovery.

**Theorem 3.1** (Hawkes Process as a Linear Autoregressive Model)**.** *Let $\mathcal{N}_\mathcal{G} = \{N_i\}_{i=1}^l$ be a stationary multivariate Hawkes process with intensities $\{\mu_i\}$ and excitation functions $\{\phi_{ij}(s)\}$. Define discretized event counts*

$$N_i^{(n)} \coloneqq N_i(n\Delta) - N_i((n-1)\Delta), \quad N_i^{(0)} = 0,$$

*for bin width $\Delta \in (0, \delta)$. As $\Delta \to 0$, the process admits the linear autoregressive representation*

$$N_i^{(n)} = \sum_{j=1}^{l} \sum_{k=1}^{n} \theta_{ij}^{(k)} N_j^{(n-k)} + \varepsilon_i^{(n)} + \theta_i^{(0)}, \tag{2}$$

*where $\theta_i^{(0)} = \Delta\mu_i$, $\theta_{ij}^{(k)} = \int_{(k-1)\Delta}^{k\Delta} \phi_{ij}(s)ds$, and $\varepsilon_i^{(n)}$ is white noise.*

This discrete-time view shows that each current variable $N_i^{(n)}$ is a weighted sum of lagged variables plus noise, enabling causal inference via cross-covariance rank conditions. Proofs are deferred to Appendix G. In practice, only a finite number of lags need be considered, since excitation functions decay and $\theta_{ij}^{(k)}$ vanish for large $k$; we choose $m$ lags exceeding this effective support [27, 36].

## 3.2 Structure Discovery Through Rank Constraints

We link statistical properties of Hawkes data to the discretized variables of window causal graph, which in turn identifies the summary graph—even with latent subprocesses. Under the linear representation in Eq. 2 with white noise, the causal structure induces characteristic *low-rank* patterns in cross-covariance matrices of observed variables.

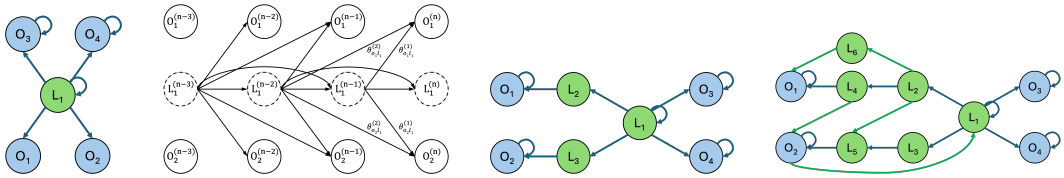
**Lemma 3.2** (D-separation $\Leftrightarrow$ Rank in the Window Graph)**.** *Consider the window causal graph of a PO-MHP. For any disjoint variable sets $\mathbf{A}_v$, $\mathbf{B}_v$ and $\mathbf{C}_v$, $\mathbf{C}_v$ d-separates $\mathbf{A}_v$ and $\mathbf{B}_v$, iff $\operatorname{rank}(\Sigma_{\mathbf{A}_v \cup \mathbf{C}_v, \mathbf{B}_v \cup \mathbf{C}_v}) = |\mathbf{C}_v|$, where $\Sigma_{\mathbf{A}_v \cup \mathbf{C}_v, \mathbf{B}_v \cup \mathbf{C}_v}$ denotes the **cross-covariance matrix** between $\mathbf{A}_v \cup \mathbf{C}_v$ and $\mathbf{B}_v \cup \mathbf{C}_v$, and $|\mathbf{C}_v|$ is the **cardinality** of $\mathbf{C}_v$.*

**Proposition 3.3** (Identifying Observed Parent Cause Set)**.** *Let $\mathcal{O}_\mathcal{G} = \{O_i\}_{i=1}^p$ be observed subprocesses (latent subprocesses may exist). For target $O_1$, the following are equivalent: (i) In the summary graph, the set $\mathcal{P}_\mathcal{G} \subseteq \mathcal{O}_\mathcal{G}$ is the parent cause set of the subprocess $O_1$; (ii) In the window graph, with the observed variable set $\mathbf{O}_v \coloneqq \{O_i^{(j)}\}_{i\in\{1,2,\ldots,p\}}^{j\in\{n-m,\ldots,n\}}$, $\mathcal{P}_\mathcal{G}$ is the minimal set such that lagged variable set $\mathbf{P}_v \coloneqq \{O_i^{(j)}\}_{O_i\in\mathcal{P}_\mathcal{G}}^{j\in\{n-m,\ldots,n-1\}}$ contains all parent variables of the current variable $O_1^{(n)}$; (iii) $\mathcal{P}_\mathcal{G}$ is the minimal set such that variable set $\mathbf{P}_v$ d-separates $O_1^{(n)}$ from the rest $\mathbf{O}_v \backslash \{\mathbf{P}_v \cup O_1^{(n)}\}$. (iv) $\mathcal{P}_\mathcal{G}$ is the minimal set such that $\operatorname{rank}(\Sigma_{O_1^{(n)} \cup \mathbf{P}_v,\ \mathbf{O}_v \backslash O_1^{(n)}}) = |\mathbf{P}_v|$.*

Proposition 3.3 depends only on observed variables and thus identifies the observed parent cause set $\mathcal{P}_\mathcal{G}$ of any target $O_1$, regardless of latent subprocesses, implying local independence of $O_1$ given $\mathcal{P}_\mathcal{G}$.

**Intermediate Latent Subprocesses.** Latent nodes lying on directed paths between $O_1$ and its identified observed parents are in general unidentifiable (their effects can be absorbed by observed parents). For Hawkes, however, once $\mathcal{P}_\mathcal{G}$ is identified, the discrete-time structure allows counting the number of such intermediate latent nodes under mild conditions; details are in Appendix C.

**Latent Confounders.** A latent confounder is a latent node that must be included in the parent cause set to render an observed effect locally independent of others (e.g., $O_1 \leftarrow L_1 \rightarrow O_2$ in Figure 2a). Rank conditions in Proposition 3.3 alone cannot reveal such $L_1$ because it is unobserved.



(a) (b) (c) (d)

Figure 2: Latent-confounder examples. (a) Summary graph with a latent confounder $L_1$ for $O_1, O_2$; some nodes have self-loops. (b) Corresponding window graph with two effective lags. (c)–(d) More complex latent paths from $L_1$ to $\{O_1, O_2\}$ via intermediate latent subprocesses.

**Assumption 1** (Excitation function)**.** Assume excitation function $\phi_{ij}(s) = a_{ij}w(s)$ with a time lag $s$ related decay function $w(s)$ and coefficients $a_{ij}$ representing the peer influence between event types.

This covers exponential decay $\alpha_{ij}e^{-\beta s}$ and other normalized decays [5]. We also assume rank-faithfulness to exclude measure-zero degeneracies (Appendix D).

Under Assumption 1, excitation coefficients in (2) decompose as $\theta_{ij}^{(k)} = a_{ij}\int_{(k-1)\Delta}^{k\Delta} w(s)ds$, so the decay part of $k$-dependence is shared across edges. In the setting of Figure 2a with two lags ($m = 2$), the current variables $(O_1^{(n)}, O_2^{(n)})$ depend linearly on $(L_1^{(n-1)}, L_1^{(n-2)})$ with a rank-1 coefficient matrix; hence $\operatorname{rank}\left(\Sigma_{\{O_1^{(n)},O_2^{(n)}\},\,\{O_i^{(j)}\}_{i\in\{3,4\}}^{j\in\{n-m,\ldots,n\}}}\right) = 1$, indicating one latent confounder affecting $O_1$ and $O_2$ (formalized in Proposition 3.5; proof in Appendix K). If $O_1, O_2$ have self-loops, indirect paths via their lagged variables increase rank; including their observed lags restores identifiability, yielding $\operatorname{rank}\left(\Sigma_{\{O_i^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\ldots,n\}},\,\{O_i^{(j)}\}_{i\in\{3,4\}}^{j\in\{n-m,\ldots,n\}}\cup\{O_i^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\ldots,n-1\}}}\right) = 2m+1$, where $2m$ accounts for observed lags of $O_1, O_2$ and the $+1$ corresponds to a single latent confounder (see Appendix E). We introduce a path situation to capture all graphical configurations that could induce rank deficiency.

**Definition 3.4** (Symmetric Acyclic Path Situation)**.** Let $L_1$ be a latent confounder for observed set $\mathcal{O}_{\mathcal{G}1}$. The following hold: (i) there exist directed paths from $L_1$ to each node in $\mathcal{O}_{\mathcal{G}1}$ containing only intermediate latent nodes (no observed nodes on the paths, and endpoints are not reused as intermediates); (ii) all such paths have equal length; (iii) the paths are acyclic and intermediate latent nodes have no self-loops.

The structures in Figures 2c and 2d satisfy Definition 3.4; adding or removing intermediate latent nodes asymmetrically or forming cycles breaks it. The next result leverages Definition 3.4 to detect a latent confounder from observed effects.

**Proposition 3.5** (Identifying a Latent Confounder from Observed Effects)**.** *Consider a PO-MHP with excitation function $\phi_{ij}(s) = a_{ij}w(s)$ and rank-faithfulness. Let $\mathbf{O}_v = \{O_i^{(j)}\}_{i\in\{1,\ldots,p\}}^{j\in\{n-m,\ldots,n\}}$. For two observed subprocesses $O_1, O_2$, $\operatorname{rank}\left(\Sigma_{\{O_i^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\ldots,n\}},\,\mathbf{O}_v\setminus\{O_1^{(n)},O_2^{(n)}\}}\right) = 2m+1$ iff there exists a latent confounder $L_1$ in the parent cause sets of $\{O_1, O_2\}$ such that conditioning on $\mathcal{P}'_{\mathcal{G}} = L_1 \cup \{O_1, O_2\}$ renders $\{O_1, O_2\}$ locally independent of $\mathcal{O}_{\mathcal{G}} \setminus \mathcal{P}'_{\mathcal{G}}$, and $L_1$ with $\{O_1, O_2\}$ satisfy the Definition 3.4.*

**Surrogate-Based Recovery of Remaining Relations.** Once a latent confounder $L_1$ is detected from its observed effects (Proposition 3.5), we recover the remaining relations by treating one observed effect as an *observed surrogate* of $L_1$ and grouping its *observed siblings*. Intuitively, under the excitation function assumption and rank-faithfulness, conditioning on the surrogate (and its siblings) isolates the local influence of the underlying latent node, so that the parent-cause tests reduce to rank conditions on blocks of cross-covariances, analogous to the observed-only case. This surrogate view also composes: it enables testing relations between an inferred latent and an observed node, as well as between two inferred latents via their surrogates. The formal statements–including the surrogate definition and two theorems covering (i) parent set identification when latent causes are involved and (ii) discovery of new latent subprocesses causally related to inferred latent subprocesses–together

with Fig. 7, are deferred to Appendix F. In practice, we therefore: (1) detect latent confounders via the $2m+1$ rank signature; (2) select surrogates and siblings; (3) apply the surrogate-based rank tests to complete the graph among observed and inferred latent subprocesses.

# 4 Rank-based Discovery Algorithm

In this section, we present a two-phase iterative algorithm that leverages the identification theorems to iteratively identify causal relationships among discovered subprocesses and discover new latent subprocesses. Let $\mathcal{A}_\mathcal{G}$ denote the *active process set*, consisting of subprocesses whose parent causes are yet to be identified. Initially, $\mathcal{A}_\mathcal{G}$ is set to the observed subprocess set $\mathcal{O}_\mathcal{G}$ and is progressively updated throughout the procedure. Additionally, due to the existence of cycles in the summary causal graph, observed subprocesses previously identified as effects may still serve as causes for other subprocesses in $\mathcal{A}_\mathcal{G}$, and thus remain under investigation. The overall procedure is in Algorithm 1.

**Algorithm 1** Two-Phase Iterative Discovery Algorithm

**Input:** Observed subprocess set $\mathcal{O}_\mathcal{G}$
**Output:** Causal graph $\mathcal{G}$

Initialize partial causal graph $\mathcal{G} \coloneqq \emptyset$, active process set $\mathcal{A}_\mathcal{G} \coloneqq \mathcal{O}_\mathcal{G}$.
**repeat**
  $(\mathcal{G}, \mathcal{A}_\mathcal{G}) \leftarrow$ Identifying Causal Relations $(\mathcal{G}, \mathcal{A}_\mathcal{G}, \mathcal{O}_\mathcal{G})$. //phase I
  $(\mathcal{G}, \mathcal{A}_\mathcal{G}) \leftarrow$ Discovering New Latent Subprocesses $(\mathcal{G}, \mathcal{A}_\mathcal{G}, \mathcal{O}_\mathcal{G})$. //phase II
**until** $\mathcal{A}_\mathcal{G}$ is empty or no updates occur.
**return:** $\mathcal{G}$

**Phase I: Identifying Causal Relations** Each iteration begins with Phase I, which aims to identify the causal structure for under-investigated subprocesses (both latent and observed) in $\mathcal{A}_\mathcal{G}$. In this phase, we systematically iterate over each subprocess in $\mathcal{A}_\mathcal{G}$ and attempt to identify its parent causes using the current $\mathcal{A}_\mathcal{G} \cup \mathcal{O}_\mathcal{G}$. If a subprocess's parent cause set is fully contained within this set, it can be identified using Proposition 3.3 and Theorem F.2. Once its parent cause set is identified, the subprocess is removed from $\mathcal{A}_\mathcal{G}$. This phase continues until no further updates occur. Details of this phase are provided in Algorithm 2 in Appendix O.1.

**Phase II: Discovering New Latent Subprocesses** When no more subprocesses in $\mathcal{A}_\mathcal{G}$ can be resolved using Phase I, we enter Phase II. This phase seeks to discover new latent confounder subprocesses by exhaustively checking all pairs in $\mathcal{A}_\mathcal{G}$ using Proposition 3.5 and Theorem F.3. Identified latent confounders are merged if they overlap in subprocesses, implying they share the same latent parent cause. $\mathcal{A}_\mathcal{G}$ is then updated to add new latent subprocesses and remove their effects, and the algorithm returns to Phase I in the next iteration. The procedure continues until $\mathcal{A}_\mathcal{G}$ is empty or remains unchanged. Detailed steps are provided in Algorithm 3 in Appendix O.2.

**Theorem 4.1** (Identifiability of the Causal Graph)**.** *Consider a PO-MHP with excitation function $\phi_{ij}(s) = a_{ij}w(s)$ and rank faithfulness. If each latent confounder subprocess, along with all its observed surrogates, satisfies Definition 3.4, then the causal graph over the observed subprocesses and latent confounder subprocesses can be identified. In particular, when no latent subprocesses exist, the causal graph is fully identifiable through only Phase I of the algorithm.*

Moreover, the computational complexity depends on the number of subprocesses (including latent confounders) and the density of the underlying causal graph, which together determine the number of iterations required for complete graph discovery. A detailed complexity analysis is in Appendix P.

# 5 Experiments

**Synthetic Data** We compare our method against six strong baselines. SHP [37] and THP [6] are discrete-time (binned) Hawkes methods, while NPHC [1] is a cumulant-based approach. Because existing Hawkes-based methods do not identify latent subprocesses without prior knowledge, we also include two rank-based methods designed for i.i.d. linear models—Hier. Rank [21] and RLCD [13]—and we further add LPCMCI [16] as a time-series baseline that handles exogenous

latent confounders. For these three methods (Hier. Rank, RLCD, LPCMCI), we apply the discretized (binned) Hawkes data on them. For our method, we evaluate both on event sequences generated by the Hawkes process in Eq. (1) and on data generated directly from the discrete-time model in Eq. (2). We test across six synthetic graph families: the fully observed graph in Fig. 1b and five structures with latent subprocesses in Figs. 2a and 7a–7d. We report average F1-score over ten runs on a personal PC (CPU). Additional details and further results (larger graphs, sensitivity to $\Delta$, and robustness to rank-faithfulness violations) appear in Appendix Q. As shown in Fig. 3, our method consistently outperforms the baselines on both fully and partially observed graphs. Notably, latent cases typically require larger sample sizes: because the spectral radius of a stationary Hawkes process is $< 1$, causal influences attenuate along latent paths, which in turn demands more data for reliable detection.

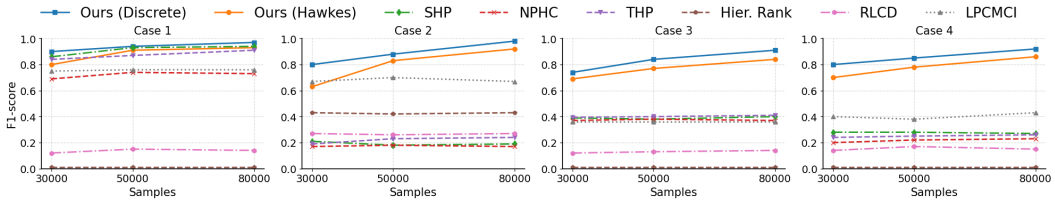



Figure 3: F1-score comparisons for first four synthetic causal graphs (Cases 1–4), corresponding to the structures in Figs. 1b, 2a, 7a and 7b. See Appendix Q.3 for additional cases.

**Real-world Data** We evaluate on a public cellular network dataset [37] with expert-validated ground truth. The corpus contains 18 alarm types collected from 55 devices ($\approx$ 35k events over eight months); not every device exhibits all alarms. We focus on `device_id=8`, which contains the alarms relevant to the subgraph under study. For evaluation, we consider a five-alarm subgraph (`Alarm_ids=0-3` and `7`) and treat `Alarm_id=7` as latent via manual exclusion. Notably, `Alarm_id=1` and `Alarm_id=3` are both observed effects of the latent subprocess (`Alarm_id=7`), which enables its recovery from observed data. Our inferred graph (Fig. 4) correctly recovers the latent subprocess and its major influences; the only discrepancy from the ground truth is a single missing edge, `Alarm_id=1` $\rightarrow$ `Alarm_id=3`. Moreover, on this sub-dataset our method quantitatively outperforms representative baselines; see Appendix Q.4 for details.

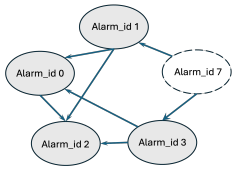



Figure 4: Inferred causal subgraph from the cellular network dataset, where `Alarm_id=7` is successfully identified as a latent subprocess.

## 6 Conclusion and Future Work

In this paper, we proposed a principled framework for structure learning in partially observed multivariate Hawkes processes (PO-MHP). By leveraging sub-covariance rank constraints and a carefully designed path constraint, our method effectively identifies both causal relationships among observed subprocesses and latent confounders influencing them. Specifically, we established necessary and sufficient conditions for inferring latent subprocesses and identifying causal relations, and developed a two-phase iterative algorithm with identifiability guarantees to recover the full causal graph. Notably, our approach naturally extends to discrete time series data, given its foundation in the discretized representation of Hawkes processes. Future work includes relaxing the identification conditions to broaden applicability, and applying our method to diverse real-world datasets for deeper domain insights.

## References


[1] Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate hawkes integrated cumulants. *Journal of Machine Learning Research*, 18(192):1–28, 2018.

[2] E. Bacry, M. Bompaire, S. Gaïffas, and S. Poulsen. tick: a Python library for statistical learning, with a particular emphasis on time-dependent modeling. *ArXiv e-prints*, July 2017.

[3] Sumanta Basu, Ali Shojaie, and George Michailidis. Network granger causality with inherent grouping structure. *The Journal of Machine Learning Research*, 16(1):417–453, 2015.

[4] Anna Bonnet, Charlotte Dion-Blanc, François Gindraud, and Sarah Lemler. Neuronal network inference and membrane potential model using multivariate hawkes processes. *Journal of Neuroscience Methods*, 372:109550, 2022.

[5] Ronald S Burt. Decay functions. *Social networks*, 22(1):1–28, 2000.

[6] Ruichu Cai, Siyu Wu, Jie Qiao, Zhifeng Hao, Keli Zhang, and Xi Zhang. Thps: Topological hawkes processes for learning causal structure on event sequences. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):479–493, 2022.

[7] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[8] Kacper Chwialkowski and Arthur Gretton. A kernel independence test for random processes. In *International Conference on Machine Learning*, pages 1422–1430. PMLR, 2014.

[9] Tom Claassen, Joris Mooij, and Tom Heskes. Learning sparse causal models is not np-hard. *arXiv preprint arXiv:1309.6824*, 2013.

[10] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[11] Hadi Daneshmand, Manuel Gomez-Rodriguez, Le Song, and Bernhard Schoelkopf. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In *International conference on machine learning*, pages 793–801. PMLR, 2014.

[12] Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):245–264, 2008.

[13] Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. *arXiv preprint arXiv:2312.11001*, 2023.

[14] Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017.

[15] Mehrdad Farajtabar, Nan Du, Manuel Gomez Rodriguez, Isabel Valera, Hongyuan Zha, and Le Song. Shaping social activity by incentivizing users. *Advances in neural information processing systems*, 27, 2014.

[16] Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in Neural Information Processing Systems*, 33:12615–12625, 2020.

[17] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[18] Asela Gunawardana, Christopher Meek, and Puyang Xu. A model for temporal dependencies in event streams. *Advances in neural information processing systems*, 24, 2011.

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

[20] Alan G Hawkes. Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198, 2018.

[21] Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. *Advances in neural information processing systems*, 35:5549–5561, 2022.

[22] Haiping Huang. Effects of hidden nodes on network structure inference. *Journal of Physics A: Mathematical and Theoretical*, 48(35):355002, 2015.

[23] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

[24] Tsuyoshi Idé, Georgios Kollias, Dzung Phan, and Naoki Abe. Cardinality-regularized hawkes-granger model. *Advances in Neural Information Processing Systems*, 34:2682–2694, 2021.

[25] Songyao Jin, Feng Xie, Guangyi Chen, Biwei Huang, Zhengming Chen, Xinshuai Dong, and Kun Zhang. Structural estimation of partially observed linear non-gaussian acyclic model: A practical approach with identifiability. In *The Twelfth International Conference on Learning Representations*, 2023.

[26] Sanggyun Kim, David Putrino, Soumya Ghosh, and Emery N Brown. A granger causality measure for point process models of ensemble neural spiking activity. *PLoS computational biology*, 7(3):e1001110, 2011.

[27] Matthias Kirchner. An estimation procedure for the hawkes process. *Quantitative Finance*, 17(4):571–595, September 2016.

[28] Matthias Kirchner. Hawkes and INAR($\infty$) processes. *Stochastic Processes and their Applications*, 126(8):2494–2525, 2016.

[29] Matthias Kirchner. *Perspectives on Hawkes processes*. ETH Zurich, 2017.

[30] Shinsuke Koyama. Coarse-grained hawkes processes. *Preprints*, 2025.

[31] Patrick J Laub, Thomas Taimre, and Philip K Pollett. Hawkes processes. *arXiv preprint arXiv:1507.02822*, 2015.

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

[33] Dixin Luo, Hongteng Xu, Yi Zhen, Xia Ning, Hongyuan Zha, Xiaokang Yang, and Wenjun Zhang. Multi-task multi-dimensional hawkes processes for modeling event sequences. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3685–3691, 2015.

[34] Christopher Meek. Toward learning graphical and causal process models. In *CI@ UAI*, pages 43–48, 2014.

[35] Judea Pearl. *Causality*. Cambridge university press, 2009.

[36] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 26, 2013.

[37] Jie Qiao, Ruichu Cai, Siyu Wu, Yu Xiang, Keli Zhang, and Zhifeng Hao. Structural hawkes processes for learning causal structure from discrete-time event sequences. *arXiv preprint arXiv:2305.05986*, 2023.

[38] Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1388–1397. Pmlr, 2020.

[39] Farnood Salehi, William Trouleau, Matthias Grossglauser, and Patrick Thiran. Learning hawkes processes from a handful of events. *Advances in neural information processing systems*, 32, 2019.

[40] Christian Shelton, Zhen Qin, and Chandini Shetty. Hawkes process inference with missing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[41] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[42] Leigh Shlomovich, Edward AK Cohen, Niall Adams, and Lekha Patel. Parameter estimation of binned hawkes processes. *Journal of Computational and Graphical Statistics*, 31(4):990–1000, 2022.

[43] Peter Spirtes. An anytime algorithm for causal inference. In *International Workshop on Artificial Intelligence and Statistics*, pages 278–285. PMLR, 2001.

[44] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

[45] Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 499–506, 1995.

[46] Peter L Spirtes. Calculation of entailed rank constraints in partially non-linear and cyclic models. *arXiv preprint arXiv:1309.7004*, 2013.

[47] Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for gaussian graphical models. *The Annals of Statistics*, 38(3), June 2010.

[48] Alejandro Veen and Frederic P Schoenberg. Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624, 2008.

[49] T w Anderson. *An introduction to multivariate statistical analysis*. Wiley & Sons, 1974.

[50] Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in neural information processing systems*, 33:14891–14902, 2020.

[51] Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Identification of linear non-gaussian latent hierarchical structure. In *International Conference on Machine Learning*, pages 24370–24387. PMLR, 2022.

[52] Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. Learning granger causality for hawkes processes. In *International conference on machine learning*, pages 1717–1726. PMLR, 2016.

[53] Hongteng Xu, Yi Zhen, and Hongyuan Zha. Trailer generation via a point process-based visual attractiveness model. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[54] Junchi Yan, Chao Zhang, Hongyuan Zha, Min Gong, Changhua Sun, Jin Huang, Stephen Chu, and Xiaokang Yang. On machine learning towards predictive sales pipeline analytics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[55] Wei Zhang, Thomas Panum, Somesh Jha, Prasad Chalasani, and David Page. Cause: Learning granger causality from event sequences using attribution methods. In *International Conference on Machine Learning*, pages 11235–11245. PMLR, 2020.

[56] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1513–1522, 2015.

[57] Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial intelligence and statistics*, pages 641–649. PMLR, 2013.

# Appendix

# A Related Work

This work is closely related to three areas: point processes, Hawkes processes, and causal discovery methods.

**Point Processes.** Extensive efforts have been devoted to understanding temporal dependencies in point processes. Meek [34] introduced a graphical framework for general point processes, leveraging $\delta^*$-separation and process independence to connect graphical representations with statistical properties. Gunawardana and Meek [18] proposed a one-dimensional point process model with piecewise-constant conditional intensity, utilizing a closed-form Bayesian approach to infer temporal dependencies between event types. Chwialkowski and Gretton [8] developed a kernel-based independence test applicable to general random processes, providing a nonparametric perspective on dependency learning.

Several studies have focused on specific structures within point processes. Basu et al. [3] investigated Granger causality for discrete transition processes while incorporating inherent grouping structures. Daneshmand et al. [11] proposed a continuous-time diffusion network inference method based on a parametric cascade generative process, advancing the modeling of temporal influence. In the context of marked point processes, Didelez [12] introduced a class of graphical models capable of capturing local independence over different marks, offering a more generalized approach to analyzing dependencies in complex systems.

**Hawkes Processes.** Hawkes processes [19, 31] constitute a class of point processes with self-exciting intensities that capture how past events modulate the likelihood of future events. Much work on learning temporal dependencies in Hawkes processes builds on Granger causality [17]. Representative extensions adopt predefined excitation kernels, including exponential [15, 57, 54, 6], power-law [56], and nonparametric forms [32, 33].

Regularization plays a central role in structure learning for Hawkes processes. Xu et al. [52] expand kernels on basis functions and use sparse-group lasso for estimation. Zhou et al. [57] propose a convex program with nuclear- and $\ell_1$-norms to promote low-rank and sparsity. Ide et al. [24] introduce cardinality-regularized Hawkes with an $\ell_0$ penalty. Nonparametric approaches include estimating integrated kernels [1] and deep models with attribution for Granger inference [55]. However, most of these methods target relations among *observed* subprocesses and do not address *truly latent* components.

When only binned counts are available, a line of work fits Hawkes from discretized event sequences. Shlomovich et al. [42] develop an EM procedure with importance sampling to estimate parameters from binned data when exact timestamps are unavailable. Qiao et al. (SHP) [37] learn causal structure from discrete-time event sequences via sparsity-regularized likelihood over bin counts. Cai et al. (THPs) [6] incorporate topological constraints to recover causal influences on discretized sequences. These discrete/binned approaches generally assume full observability and do not identify the existence or number of *latent* subprocesses.

**Causal Discovery Methods.** Causal discovery [35] aims to uncover causal relations from data and has been studied extensively under i.i.d. assumptions with DAG structures. Classical families include constraint-based methods (e.g., PC [44]), score-based methods (e.g., GES [7]), and functional approaches (e.g., LiNGAM [41]).

Latent variables present significant challenges to these methods. To address this, extensions such as the FCI algorithm [45, 43] and its variants [10, 9] leverage conditional independence constraints to infer partial causal structures in the presence of independent (i.e., exogenous) latent confounders.

Recent advances have extended these methods to handle causally related latent confounders. Representative examples include Huang et al. [21] and Dong et al. [13], which identify equivalence classes in linear models by leveraging second-order (rank) statistics. However, the result graphs of their approaches are usually equivalent classes of the ground truth graph, and these approaches typically rely on structural conditions that are *not* natural in discretized Hawkes settings: (i) hierarchical latent structures [21] (e.g., no observed-to-observed edges and no observed-to-latent edges), and (ii) cardinality constraints [21, 13] (e.g., $|\text{children}| > |\text{parents}|$ for latent groups). In time-series obtained from Hawkes processes, the induced autoregressive representation is *dense* across many lags, so observed surrogates are often fewer than the effective latent "parents," violating such cardinality requirements; moreover, endogenous latent confounders (latent variables influenced by observed

processes) naturally happen in our setting. Furthermore, Xie et al. [50, 51] and Jin et al. [25] utilized higher-order statistics to accurately identify causal graphs even in the presence of latent confounders. But they still have unfeasible cardinality constraints, and they still assume i.i.d. samples, which can introduce spurious dependencies and invalidate guarantees when temporal constraints are ignored.

There are also extensions of constraint-based discovery to time series (e.g., SVAR-based LiNGAM [23]) and PC-style temporal methods (e.g., PCMCI [38], LPCMCI [16]). These rely on conditional independence tests over lagged variables and again presuppose assumptions (weak autocorrelation, exogenous latent variables) that are misaligned with Hawkes dynamics, where dense cross-lag effects and endogenous latent variables are common.

#### A.0.1 Detailed Relation to a Binned Hawkes process Estimation Method

Shlomovich et al. [42] address parameter estimation for binned Hawkes processes via a modified EM algorithm when only bin counts $N_t = N((t+1)\Delta) - N(t\Delta)$ are observed and exact event times are unavailable. The bin counts are treated as observed data and the unobserved event times $\mathcal{T}$ as latent variables (their Eq. 6). Because direct Monte Carlo sampling of $\mathcal{T}$ is intractable in Hawkes models, they employ importance sampling to simulate within-bin timestamps that match the observed counts, thereby maximizing the (binned) likelihood (see their Sec. 2).

Our goal and methodology differ. Leveraging the link between INAR and linear autoregressive models, Theorem 3.1 establishes an explicit linear structural representation for discretized multivariate Hawkes processes. This connection enables causal discovery directly over binned variables—including the identification of latent confounder subprocesses—with identifiability guarantees (Propositions 3.3 and 3.5; Theorems F.2 and F.3). In contrast to likelihood maximization based on simulated event times, our framework uses time-aware rank constraints on cross-covariances to recover causal structure. To the best of our knowledge, prior work has not provided a direct, theoretically grounded reduction from Hawkes processes to linear structural models for the purpose of causal discovery.

#### A.0.2 Detailed Relation to Rank-Based Latent Discovery in i.i.d. Models

Huang et al. [21] (and related works by Xie et al. [51] and Dong et al. [13]) study latent structure discovery under i.i.d. assumptions and continuous variables. Our problem differs substantively: we aim to recover causal structure among *observed and latent subprocesses* in multivariate Hawkes processes, where each subprocess is a point process and inference is performed on discretized representations.

**Different Data Domain and Causal Assumptions.** Huang et al. [21] (and Xie et al. [51]) assume a latent hierarchical structure, specifically: (i) there are no direct causal links among observed variables, and all dependencies among observed variables arise exclusively from their latent confounder variables; and (ii) observed variables cannot cause latent variables, i.e., endogenous latent confounders are ruled out (see Eq. 1 and Definition 1 in [21], and Eq. 1, 2 and Definition 1 in [51]). Neither assumption is needed in our framework. We allow both direct observed-to-observed edges (see Proposition Proposition 3.3 in our paper) and the existence of endogenous latent confounder subprocesses that can be caused by observed subprocesses (see Theorem F.2 in our paper).

**Cardinality Requirements vs. Hawkes Density.** Huang et al. [21], Xie et al. [51], and Dong et al. [13] rely on a cardinality condition of the form $|\text{children}| > |\text{parents}|$ for certain latent sets (cf. Definition 4 in [21], Condition 1 in [51], Definition 5 in [13]). This is generally incompatible with discretized Hawkes processes, whose autoregressive representation is inherently dense (Eq. 2 in our paper): if a latent $L_1$ causes $O_2$, then each discretized variable $O_2^{(n)}$ is influenced by *many* lags of $L_1$ (potentially hundreds or thousands in practice), making the required $|\text{children}| > |\text{parents}|$ condition fail systematically. Our method avoids such cardinality assumptions: leveraging the separable excitation (Assumption 1), we place lagged observed variables on both sides of carefully chosen cross-covariance blocks so that rank deficiency reliably signals latent confounders (lines 199–216; Proposition 3.5; Theorem F.3).

**Time-Aware vs. i.i.d. Causal Discovery.** The above i.i.d. methods do not exploit temporal order and, in principle, can test variables at time $n$ as putative parents of variables at time $n-1$. Our procedure is explicitly time-aware: candidate parents for $t = n$ are restricted to appropriate lags

(Propositions 3.3 and 3.5; Theorems F.2 and F.3), aligning identification with Hawkes dynamics. This distinction mirrors PC [44] (i.i.d.) vs. PCMCI [38] (time series).

# B Multivariate Hawkes Process Details

Before introducing multivariate Hawkes process, we first describe the temporal point process and counting process briefly. A temporal point process is a random process whose realization consists of a list of discrete events in time $\{T_1, T_2, \dots\}$ taking values in $[0, \infty)$. Another equivalent representation is the counting process, $N_1 = \{N_1(t) | t \in [0, \infty)\}$, where $N_1(t)$ records the number of events *before* time $t$ and $N_1(0) = 0$. A multivariate point process with $l$ types of events is represented by $l$ counting processes $\{N_i\}_{i=1}^{l}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. $N_i = \{N_i(t) | t \in [0, \infty)\}$, where $N_i(t)$ is the number of type-$i$ events occurring *before* time $t$ and $N_i(0) = 0$. $\mathbf{U} = \{1, \dots, l\}$ (sometimes abbreviated as $[l]$) represents the set of event types. $\Omega = [0, \infty) \times \mathbf{U}$ is the sample space. $\mathcal{F} = \mathcal{F}(t)$ is a filtration, that is, a non-descreasing family of $\sigma$-algebras which for each time point $t \in \mathbb{R}$, represent the set of event sequences the processes can realize *before* time $t$. $\mathbb{P}$ is the probability measure. Point processes can be characterized by the conditional intensity function, which models patterns of interest, such as self-triggering or self-correcting behaviors [53]. The conditional intensity function is defined as the expected instantaneous rate of type-$i$ events occurring at time $t$, given the event history:

$$\lambda_i(t) = \lim_{h \to 0} \frac{\mathbb{E}[N_i(t+h) - N_i(t) | \mathcal{H}(t)]}{h}, \tag{3}$$

where $\mathcal{H}(t) = \{(t_k, i) | t_k < t, i \in \mathbf{U}\}$ collects historical events of all types *before* time t. The multivariate Hawkes process is a class of multivariate point processes characterized by a self-triggering pattern as defined in Definition 2.1.

# C Identifying Intermediate Latent Subprocesses

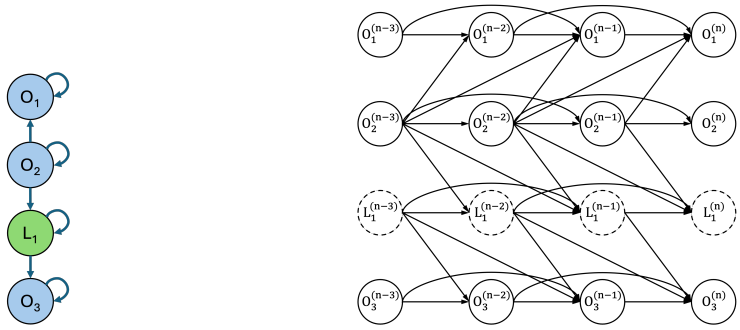



Figure 5: Example of an intermediate latent subprocess on the directed path from $O_2$ to $O_1$. (a) The summary causal graph, where $L_1$ is the intermediate latent subprocess. (b) The corresponding window causal graph with two effective lag variables.

As shown in the summary causal graph in Fig. 5a, $L_1$ is an intermediate latent subprocess on the directed path fromthe observed subprocess $O_2$ to $O_3$. According to Proposition 3.3, $L_1$ is not identifiable and its effect is attributed to $O_2$, leading to the inference that $O_2$ is the parent cause of $O_3$. This is because the influence of $L_1$ is indistinguishable from that of $O_2$ and can be effectively merged into $O_2$.

Consider now the corresponding window causal graph shown in Fig. 5b. The observed variable set is given by $\mathbf{O}_v := \{O_i^{(j)}\}_{i \in \{1,2,3\}}^{j \in \{n-m, \dots, n\}}$, where $m = 3$ exceeds the number of effective lag variables (which is 2 in this example). Instead of conditioning on all three lagged variables $\{O_2^{(n-1)}, O_2^{(n-2)}, O_2^{(n-3)}\}$ of $O_2$, we exclude $O_2^{(n-1)}$ and condition only on $\{O_2^{(n-2)}, O_2^{(n-3)}\}$. In this case, $O_3^{(n)}$ becomes d-separated from the remaining variables in $\mathbf{O}_v$. This property arises because, due to the presence of the intermediate latent subprocess $L_1$, $O_2^{(n-1)}$ no longer has a direct

influence on $O_3^{(n)}$. The following corollary formalizes a general method for identifying the number of intermediate latent subprocesses that may exist between an observed subprocess and each of its inferred observed parent causes.

**Corollary C.1** (Identifying Intermediate Latent Subprocesses)**.** *Let $\mathcal{O}_\mathcal{G} \coloneqq \{O_i\}_{i=1}^p$ denote the observed subprocesses, with the corresponding observed variable set $\mathbf{O}_v \coloneqq \{O_i^{(j)}\}_{i\in\{1,2,\dots,p\}}^{j\in\{n-m,\dots,n\}}$. Consider an observed subprocess $O_1$ and its inferred observed parent cause set $\mathcal{P}_\mathcal{G} \subseteq \mathcal{O}_\mathcal{G}$. For any $O_2 \in \mathcal{P}_\mathcal{G}$, let $h$ be the largest value such that the lagged variable set $\mathbf{P}_v \coloneqq \{O_i^{(j)}\}_{O_i\in\mathcal{P}_\mathcal{G}}^{j\in\{n-m,\dots,n-1\}} \backslash \{O_2^{(j)}\}_{j\in\{n-h,\dots,n-1\}}$ d-separates $O_1^{(n)}$ from the remaining variables $\mathbf{O}_v \backslash \{\mathbf{P}_v \cup O_1^{(n)}\}$. Equivalently, $h$ is the largest value such that:*

$$\operatorname{rank}\left(\Sigma_{\{O_1^{(n)}\}\cup\mathbf{P}_v,\, \mathbf{O}_v\backslash\{O_1^{(n)}\}}\right) = |\mathbf{P}_v|.$$

*This is equivalent to stating that the shortest directed path from $O_2$ to $O_1$ that does not pass through any other observed subprocess consists of $h$ latent subprocesses.*

*Remark* C.2. In Corollary C.1, $O_1$ and $O_2$ may refer to the same subprocess in cases where Proposition 3.3 infers that $O_1$ has a self-loop. In such cases, Corollary C.1 can be used to determine whether this self-loop represents a direct self-excitation or is mediated through intermediate latent subprocesses.

*Proof.* Let $\mathcal{O}_\mathcal{G} \coloneqq \{O_i\}_{i=1}^p$ and $\mathbf{O}_v \coloneqq \{O_i^{(j)}\}_{i\in\{1,2,\dots,p\}}^{j\in\{n-m,\dots,n\}}$. Consider an observed subprocess $O_1$ and its inferred parent cause set $\mathcal{P}_\mathcal{G}$. For any $O_2 \in \mathcal{P}_\mathcal{G}$, assume the shortest directed path from $O_2$ to $O_1$ consists of $h$ latent subprocesses. This implies that the lagged variables $\{O_2^{(j)}\}_{j\in\{n-h,\dots,n-1\}}$ do not influence $O_1^{(n)}$, while the variables $\{O_2^{(j)}\}_{j\in\{n-m,\dots,n-h\}}$ do.

Thus, the variable set $\mathbf{P}_v = \{O_i^{(j)}\}_{O_i\in\mathcal{P}_\mathcal{G}}^{j\in\{n-m,\dots,n-1\}} \setminus \{O_2^{(j)}\}_{j\in\{n-h,\dots,n-1\}}$ is the minimal set that d-separates $O_1^{(n)}$ from the remaining variables. By Lemma 3.2, this implies:

$$\operatorname{rank}\left(\Sigma_{\{O_1^{(n)}\}\cup\mathbf{P}_v,\, \mathbf{O}_v\backslash\{O_1^{(n)}\}}\right) = |\mathbf{P}_v|.$$

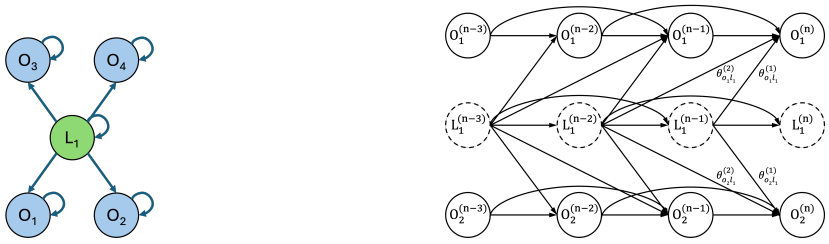
This completes the proof. □

# D Rank Faithfulness for the Hawkes Process

**Assumption 2** (Rank Faithfulness for the Hawkes Process)**.** A probability distribution $p$ is rank faithful to the graph $\mathcal{G}$ if every rank constraint on any sub-covariance matrix that holds in $p$ is entailed by every linear structural model (as defined in Eq. 1) with respect to $\mathcal{G}$ and the excitation function $\phi_{ij}(s) = a_{ij}w(t)$, $\forall i, j \in \{1, \dots, l\}$.

The rank faithfulness assumption is widely adopted in the causal discovery literature for i.i.d. data [46, 21]. In our setting, it concerns only the excitation function coefficients $a_{ij}$, and prior studies have shown that violations of this assumption occur only in degenerate cases of Lebesgue measure zero. Specifically, it fails only in rare pathological scenarios, such as when multiple $a_{ij}$ coefficients involving those of latent subprocesses are exactly equal across different subprocesses in a manner that induces rank deficiency—situations that are highly unlikely to arise in practical applications.

To empirically assess the robustness of our method to potential violations of rank faithfulness, we conduct a sensitivity analysis where, for each synthetic graph, we choose the exponential excitation function $\phi_{ij}(s) = \alpha_{ij}e^{-\beta s}$ and deliberately assign identical $a_{ij}$ values to two randomly selected edges, thereby artificially increasing the risk of the violation of rank faithfulness. The results, reported in Table 3 in Appendix Q.3, demonstrate that our method remains robust even under such perturbations.




(a) (b)

Figure 6: Illustration of self-Looped observed subprocesses under latent confounder influence. (a) Summary causal graph where $O_1, O_2, O_3$, and $O_4$ are observed subprocesses, and $L_1$ is a latent confounder subprocess. All subprocesses have self-loops. (b) Corresponding window causal graph for (a), illustrating the discretized causal mechanisms among $O_1, O_2$, and $L_1$, with two effective lag variables.

# E Accounting for Self-Looped Observed Subprocesses under Latent Confounder Influence

Consider Fig. 6, where $O_1$ and $O_2$ also have self-loops. As shown in Fig. 6b, these self-loops introduce additional indirect effects, where the lagged latent variables $\{L_1^{(j)}\}_{j\in\{n-m,\dots,n-1\}}$ propagate their influence to the current variables $O_1^{(n)}$ and $O_2^{(n)}$ through the observed lagged variables $\{O_i^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\dots,n-1\}}$.

Fortunately, since these lagged variables are observed, they can be explicitly incorporated into the structural equations and, correspondingly, into the covariance matrix. Considering the window graph in Fig. 6b with $m$ effective lag variables, the structural equations for the observed variables $\{O_i^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\dots,n\}}$ can be written as:

$$\begin{bmatrix} O_1^{(n)} \\ O_1^{(n-1)} \\ \cdots \\ O_1^{(n-m)} \\ O_2^{(n)} \\ O_2^{(n-1)} \\ \cdots \\ O_2^{(n-m)} \end{bmatrix} = \mathbf{E} \begin{bmatrix} L_1^{(n-1)} \\ \cdots \\ L_1^{(n-m)} \\ O_1^{(n-1)} \\ \cdots \\ O_1^{(n-m)} \\ O_2^{(n-1)} \\ \cdots \\ O_2^{(n-m)} \end{bmatrix} + \begin{bmatrix} \epsilon_{o_1}^{(n)} + \theta_{o_1}^{(0)} \\ \epsilon_{o_1}^{(n-1)} + \theta_{o_1}^{(0)} \\ \cdots \\ \epsilon_{o_1}^{(n-m)} + \theta_{o_1}^{(0)} \\ \epsilon_{o_2}^{(n)} + \theta_{o_1}^{(0)} \\ \epsilon_{o_2}^{(n-1)} + \theta_{o_1}^{(0)} \\ \cdots \\ \epsilon_{o_2}^{(n-m)} + \theta_{o_1}^{(0)} \end{bmatrix}, \tag{4}$$

$$\mathbf{E} = \begin{bmatrix} a_{o_1 l_1} \int_0^{\Delta} w(s)ds & \cdots & a_{o_1 l_1} \int_{(m-1)\Delta}^{m\Delta} w(s)ds & 1 & \overset{1}{\cdots} & 1 & 0 & \overset{0}{\cdots} & 0 \\ 0 & \overset{0}{\cdots} & 0 & 1 & \overset{0}{\cdots} & 0 & 0 & \overset{0}{\cdots} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \overset{0}{\cdots} & 0 & 0 & \overset{0}{\cdots} & 1 & 0 & \overset{0}{\cdots} & 0 \\ a_{o_2 l_1} \int_0^{\Delta} w(s)ds & \cdots & a_{o_2 l_1} \int_{(m-1)\Delta}^{m\Delta} w(s)ds & 0 & \overset{0}{\cdots} & 0 & 1 & \overset{1}{\cdots} & 1 \\ 0 & \overset{0}{\cdots} & 0 & 0 & \overset{0}{\cdots} & 0 & 1 & \overset{0}{\cdots} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \overset{0}{\cdots} & 0 & 0 & \overset{0}{\cdots} & 0 & 0 & \overset{0}{\cdots} & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \}\text{m} \\ \\ \\ \\ \}\text{m} \end{matrix}. \tag{5}$$

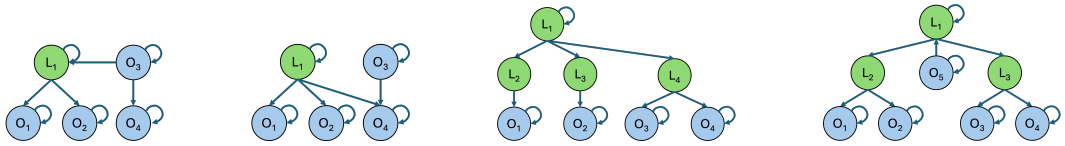
It is straightforward to see that the rank of the coefficient matrix $\mathbf{E}$ is $2m+1$. Accordingly, by including these observed lagged variables in the cross-covariance matrix, we obtain:

$$\operatorname{rank}\left(\Sigma_{\{O_i^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\dots,n\}},\ \{O_i^{(j)}\}_{i\in\{3,4\}}^{j\in\{n-m,\dots,n\}} \cup \{O_i^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\dots,n-1\}}}\right) = 2m+1,$$

where $2m$ corresponds to the observed lagged variables of $O_1$ and $O_2$, and 1 corresponds to the latent confounder subprocess $L_1$. For a formal proof, see Proposition 3.5 and Appendix K. This result implies the presence of a latent confounder subprocess $L_1$, such that the set $\{L_1, O_1, O_2\}$ forms the parent cause set of $\{O_1, O_2\}$. Conditioning on this set renders $\{O_1, O_2\}$ locally independent of $O_3$ and $O_4$.

# F Surrogate-Based Recovery of Latent Structure

Proposition 3.5 allows us to infer the existence of a latent confounder from its observed effects. This raises an important question: *How can we systematically infer the remaining causal relations involving the inferred latent subprocesses?* This challenge is illustrated by the four summary graphs in Fig. 7. In the following, we show how the observed effects can serve as surrogates for their associated latent confounders, enabling the recovery of the remaining causal structure.




(a) (b) (c) (d)

Figure 7: Illustrative examples of interactions among inferred latent confounder and the remaining observed subprocesses. In (a)–(c), assume $L_1$ has been inferred via its observed effects $\{O_1, O_2\}$. (a) $O_3$ causes $L_1$. (b) Both $L_1$ and $O_3$ cause $O_4$. (c) $L_1$ causes $L_4$, where $L_4$ can be inferred from $\{O_3, O_4\}$. (d) $L_1$ serves as the latent confounder of both latent confounder $L_2$ and $L_3$.

**Definition F.1** (Observed Effects as Surrogates)**.** For each latent subprocess $L_1$ inferred from its observed effects $\{O_1, O_2\}$, we define one of its observed effects, denoted as $\mathcal{D}e(L_1) \coloneqq O_1$, to serve as an *observed surrogate* of $L_1$. This surrogate is chosen such that there exists a directed path from $L_1$ to $\mathcal{D}e(L_1)$ that does not pass through any other observed subprocesses. We further define $\mathcal{S}ib(\mathcal{D}e(L_1))$ as the set of *observed siblings* of $\mathcal{D}e(L_1)$, containing all known other observed subprocesses affected by $L_1$ through paths that also do not pass through other observed subprocesses.

For any observed subprocess $O_1$, we adopt the unified notation $\mathcal{D}e(O_1) = O_1$, and correspondingly, $\mathcal{S}ib(\mathcal{D}e(O_1)) = \emptyset$. Moreover, $\mathcal{S}ib(\mathcal{D}e(L_1))$ represents the minimal set of observed subprocesses required to isolate the local influence of $L_1$ on the rest of the system, except through $\mathcal{D}e(L_1)$.

**Theorem F.2** (Identifying Parent Cause Set with Latent Confounder Involved)**.** *Consider a PO-MHP with excitation function $\phi_{ij}(s) = a_{ij}w(s)$ and rank faithfulness. The system $\mathcal{N}_\mathcal{G} \coloneqq \mathcal{O}_\mathcal{G} \cup \mathcal{L}_\mathcal{G}$ consists of observed subprocesses $\mathcal{O}_\mathcal{G} \coloneqq \{O_i\}_{i=1}^p$, and inferred latent confounder processes $\mathcal{L}_\mathcal{G}$ whose parent cause sets are yet to be identified. Let $\mathbf{O}_v \coloneqq \{O_i^{(j)}\}_{i\in\{1,\dots,p\}}^{j\in\{n-m,\dots,n\}}$ denote the corresponding observed variable set. For a subprocess $N_1 \in \mathcal{N}_\mathcal{G}$ and a candidate parent cause set $\mathcal{P}'_\mathcal{G} \subseteq \mathcal{N}_\mathcal{G}$, when either $N_1$ is latent, or $\mathcal{P}'_\mathcal{G}$ contains latent subprocesses, or both, the following condition holds: $\mathcal{P}'_\mathcal{G}$ is the minimal set such that* $\operatorname{rank}(\Sigma_{\mathbf{A}_v,\mathbf{B}_v}) = |\mathbf{A}_v| - 1$, *where* $\mathbf{A}_v \coloneqq \{\mathcal{D}e(N_1)^{(j)}, \mathcal{D}e(L_i)^{(j)}\}_{L_i\in\mathcal{P}'_\mathcal{G}}^{j\in\{n-m,\dots,n\}} \cup \{O_i^{(j)}\}_{O_i\in\mathcal{P}'_\mathcal{G}}^{j\in\{n-m,\dots,n-1\}} \cup \{O_i^{(j)}\}_{O_i\in\mathcal{S}ib(\mathcal{D}e(N_1))\cup\{\mathcal{S}ib(\mathcal{D}e(L_i))\}_{L_i\in\mathcal{P}'_\mathcal{G}}}^{j\in\{n-m,\dots,n\}}$ *and* $\mathbf{B}_v = \mathbf{O}_v \setminus \left(\mathcal{D}e(N_1)^{(n)} \cup \{\mathcal{D}e(L_i)^{(n)}\}_{L_i\in\mathcal{P}'_\mathcal{G}}\right)$, *if and only if $\mathcal{P}'_\mathcal{G}$ is a subset of the parent cause set of $N_1$ such that: conditioning on $\mathcal{S}_\mathcal{G} \coloneqq \mathcal{P}'_\mathcal{G} \cup \mathcal{D}e(N_1) \cup \{\mathcal{D}e(L_i)\}_{L_i\in\mathcal{P}'_\mathcal{G}} \cup \mathcal{S}ib(\mathcal{D}e(N_1)) \cup \{\mathcal{S}ib(\mathcal{D}e(L_i))\}_{L_i\in\mathcal{P}'_\mathcal{G}}$ renders $N_i$ locally independent of $\mathcal{N}_\mathcal{G}\setminus\mathcal{S}_\mathcal{G}$; for each $L_i \in \mathcal{P}'_\mathcal{G}$, the latent confounder $L_i$ with observed effects $\{\mathcal{D}e(N_1), \mathcal{D}e(L_i)\}$ satisfies Definition 3.4; and, all possible observed surrogates of $N_i$ in $\mathcal{O}_\mathcal{G}$ have been identified so as to be added into the observed sibling set.*

With Theorem F.2 (and Proposition 3.3), we can identify arbitrary causal relations among both observed and inferred latent subprocesses. This naturally raises a final question: *How can we further infer new latent subprocesses that are causally related to inferred latent subprocesses, as in Fig. 7d?* As shown in the following theorem, the observed surrogate of a latent subprocess can still be leveraged for such inference.

**Theorem F.3** (Identifying Latent Confounder from Latent Confounder )**.** *Consider a PO-MHP with excitation function $\phi_{ij}(s) = a_{ij}w(s)$ and rank faithfulness. The system $\mathcal{N}_\mathcal{G} \coloneqq \mathcal{O}_\mathcal{G} \cup \mathcal{L}_\mathcal{G}$ consists of observed subprocesses $\mathcal{O}_\mathcal{G} \coloneqq \{O_i\}_{i=1}^p$, and inferred latent confounder processes $\mathcal{L}_\mathcal{G}$ whose parent cause sets remain unidentified by Theorem F.2. Let $\mathbf{O}_v \coloneqq \{O_i^{(j)}\}_{i\in\{1,\dots,p\}}^{j\in\{n-m,\dots,n\}}$ denote the corresponding observed variable set. For any two subprocesses $N_1, N_2 \subseteq \mathcal{N}_\mathcal{G}$ (either observed or latent),* $\operatorname{rank}(\Sigma_{\mathbf{A}_v,\mathbf{B}_v}) = |\mathbf{A}_v| - 1$*, where* $\mathbf{A}_v \coloneqq \{\mathcal{D}e(N_i)^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\dots,n\}} \cup \{O_i^{(j)}\}_{O_i\in\mathcal{S}ib(\mathcal{D}e(N_1))\cup\mathcal{S}ib(\mathcal{D}e(N_2))}^{j\in\{n-m,\dots,n\}}$ *, and* $\mathbf{B}_v \coloneqq \mathbf{O}_v\backslash\{\mathcal{D}e(N_1)^{(n)}, \mathcal{D}e(N_2)^{(n)}\}$ *, if and only if there exits a latent confounder subprocess $L_1$ in the parent cause set of $\{N_1, N_2\}$ such that: conditioning on $\mathcal{P}'_\mathcal{G} \coloneqq L_1 \cup \{N_i\}_{i\in\{1,2\}} \cup \{\mathcal{S}ib(\mathcal{D}e(N_i))\}_{i\in\{1,2\}}$ renders $\{N_1, N_2\}$ locally independent of $\mathcal{N}_\mathcal{G}\backslash\mathcal{P}'_\mathcal{G}$; $L_1$ with $\{\mathcal{D}e(N_1), \mathcal{D}e(N_2)\}$ satisfies Definition 3.4; and all possible observed surrogates of $\{N_1, N_2\}$ in $\mathcal{O}_\mathcal{G}$ have been identified so as to be added into the observed sibling set.*

Theorem F.2 and Theorem F.3 are extensions of Proposition 3.3 and Proposition 3.5, respectively. These extend the framework by replacing latent subprocesses with their observed surrogates when evaluating the rank of the relevant sub-covariance matrices.

# G Proof of Theorem 3.1

*Proof.* To prove Theorem 3.1, we proceed in three steps. First, we define the multivariate INAR sequence (Definition G.1) and show that it admits a linear autoregressive model representation (Proposition G.3). Then, in Theorem G.5, we establish that this multivariate INAR counting process converges weakly to a multivariate Hawkes process as the bin size $\Delta \to 0$, with the correspondence between the parameters of both models made explicit. The details are as follows:

**Step 1: Definition of the Multivariate INAR model.** We begin by introducing the multivariate INAR model, adapted from Definition 20 in the paper B. Hawkes forests in [29].

**Definition G.1** (Multivariate integer-valued autoregressive model [29])**.** An multivariate integer-valued autoregressive time series(multivariate INAR) is a sequence of $\mathbb{N}_0$-valued random variables $\mathbf{X}_v = \{X_1^{(n)}, X_2^{(n)}, \dots, X_l^{(n)}\}_{n\in\mathbb{N}_0}$ with $X_i^{(0)} = 0$, defined as:

$$X_i^{(n)} = \sum_{j=1}^{l}\sum_{k=1}^{n}\sum_{h=1}^{X_j^{(n-k)}} \xi_h^{(\theta_{ij}^{(k)})} + \epsilon_i^{(n)}, \quad i \in \{1,\dots,l\}, n \in \mathbb{N}_0, \tag{6}$$

where the reproduction coefficients $\theta_{ij}^{(k)} \geq 0$ with the subcritical matrix $[\sum_{k=1}^n \theta_{ij}^{(k)}]_{(i,j)\in\{1,\dots,l\}}$, and the immigration coefficients $\theta_i^{(0)} \geq 0$. $\epsilon_i^{(n)} \overset{iid}{\sim} \text{Pois}(\theta_i^{(0)})$ and $\xi_h^{(\theta_{ij}^{(k)})} \overset{iid}{\sim} \text{Pois}(\theta_{ij}^{(k)})$ are mutually independent and also independent of $\epsilon_i^{(n)}$.

*Remark* G.2. Definition G.1 follows Definition 20 in paper B. Hawkes Forests in [29], but with adapted notation to match Theorem 3.1. Key correspondences include: $d \to l$, $i \to j$, $j \to i$, $l \to h$, $(\mathbf{X}_n)_{n\in\mathbb{Z}} \to \mathbf{X}_v$, $X_n^{(j)} \to X_i^{(n)}$, $\xi_{n,l}^{(i,j,k)} \to \xi_h^{(\theta_{ij}^{(k)})}$, $\epsilon_n^{(j)} \to \epsilon_i^{(n)}$, $\alpha_{i,j,k} \to \theta_{ij}^{(k)}$, $\alpha_{0,j} \to \theta_i^{(0)}$. We also restrict indices to $n \in \mathbb{N}_0$ to match our Hawkes process formulation (Definition 2.1); this is purely notational and does not affect the model semantics, as the indices are used to describe relative positions within the time series.

**Step 2: Linear autoregressive representation of the INAR model.** The multivariate INAR sequence admits an equivalent linear autoregressive representation, as shown in Proposition G.3, corresponding to Proposition 3.1 in [27]. The current variable $X_i^{(n)}$ is expressed as a weighted sum of all lag variables $X_j^{n-k}$, plus a constant term $\theta_i^{(0)}$ and a stationary white-noise term $\varepsilon_i^{(n)}$.

**Proposition G.3.** *Let $\mathbf{X}_v$ be a $l$-dimensional INAR sequence as in Definition G.1 with immigration coefficients $\theta_i^{(0)} \geq 0$, reproduction coefficients $\theta_{ij}^{(k)} \geq 0$, and $X_i^{(0)} = 0$. Then*

$$\varepsilon_i^{(n)} \coloneqq X_i^{(n)} - \theta_i^{(0)} - \sum_{j=1}^{l}\sum_{k=1}^{n} \theta_{i,j}^k X_j^{(n-k)}, \quad n \in \mathbb{N}_0, \tag{7}$$

*defines a white-noise sequence, i.e.,* $(\varepsilon_i^{(n)})$ *is stationary,* $\mathbb{E}[\varepsilon_i^{(n)}] = 0$*,* $i \in \{1, \dots, l\}$*,* $n \in \mathbb{N}_0$*. Moreover, let the* $l \times l$ *noise matrices* $\mathbf{u}_n \mathbf{u}_{n'}^\top := [\varepsilon_i^{(n)} \varepsilon_j^{(n)}]_{(i,j)\in\{1,\dots,l\}}$ *and reproduction-coefficient matrices* $A_k := [\theta_{ij}^{(k)}]_{(i,j)\in\{1,\dots,l\}}$*, we have:*

$$\mathbb{E}[\mathbf{u}_n \mathbf{u}_{n'}^\top] = \begin{cases} \operatorname{diag}\left( \left( I_{l\times l} - \sum_{k=1}^n A_k \right)^{-1} \right), & n = n', \\ 0_{l\times l}, & n \neq n'. \end{cases} \tag{8}$$

*Remark* G.4. Proposition G.3 is adapted from Proposition 3.1 of [27], which also appears as Proposition 6 of the same paper in the author's doctoral thesis [29]. The original formulation uses full vector and matrix notation; here, we present each dimension separately for consistency with our notation. Moreover, we adapted notations as in Remark G.2.

**Step 3: Convergence of the INAR to a Hawkes process.** Finally, we show that the multivariate INAR process converges to a multivariate Hawkes process as $\Delta \to 0$. The corresponding parameters of the INAR and the Hawkes process are also stated in the below theorem.

**Theorem G.5** (Multivariate INAR converging to multivariate Hawkes process [29])**.** *Let* $\mathcal{N}_{\mathcal{G}1} = \{N_i\}_{i=1}^l$ *be a stationary multivariate Hawkes process with background intensities* $\{\mu_i\}_{i=1}^l$*, and piecewise-continuous excitation functions* $\{\phi_{ij}(s) \geq 0, \forall s \in (0,\infty)\}_{i=1}^l$*. For bin width* $\Delta \in (0, \delta)$*, let* $\mathbf{X}_v = \{X_1^{(n)}, X_2^{(n)}, \dots, X_l^{(n)})_{n\in\mathbb{N}_0}$ *be an multivariate* INAR *sequence with:*

$$\theta_i^{(0)} = \Delta\mu_i, \quad \theta_{ij}^{(k)} = \int_{(k-1)\Delta}^{k\Delta} \phi_{ij}(s)ds,$$

*and* $X_i^{(0)} = 0$*. From the sequences* $\mathbf{X}_v$*, we define a family of point processes* $\mathcal{N}_{\mathcal{G}2} = \{N_i^\Delta\}_{i=1}^l$*, where for each* $N_i^\Delta$*,*

$$N_i^\Delta(t) := \sum_{n:n\Delta\leq t} X_i^{(n)}, \quad t \in [0,\infty). \tag{9}$$

*Then,* $\mathcal{N}_{\mathcal{G}2}$ *converges weakly to* $\mathcal{N}_{\mathcal{G}1}$ *in distribution, as* $\Delta \to 0$*.*

*Remark* G.6. Theorem G.5 is a simplified version of Theorem 25 in [29]. The original proof proceeds via convergence of Hawkes forests (constructed via branching random walks), showing that the Hawkes process is a limit of INAR-based approximating forests. The convergence of Hawkes process and INAR comes from the convergence of Hawkes forest and the approximating forest with corresponding parameters. We adapt it here with a direct correspondence between Hawkes and INAR parameters, and restrict domains to $t \in [0,\infty)$ and $n \in \mathbb{N}_0$ for consistency and clarification. Typically, Hawkes process results hold for both domains [31, Remark 2], since variable $t$ and $n$ is used only to calibrate relative positions. Moreover, besides the notation changes in Remark G.2, we adopt: $\mathbf{N}_\mathbf{F} \to \mathcal{N}_{\mathcal{G}1}$, $\mathbf{N}_{\mathbf{F}^{(\Delta)}} \to \mathcal{N}_{\mathcal{G}2}$, the reproduction intensities $h_{i,j} = w_{i,j}m_{i,j} \to$ excitation function $\phi_{ij}$.

*Remark* G.7. The constant $\delta$ in the Theorem G.5 comes from the moment structure of the INAR sequence. For details, see Theorem 2 in [28] and Corollary 24 in paper B. Hawkes forests in [29].

**In summary:** The linear autoregressive representation of the multivariate INAR model is established in Proposition G.3, based on the model definition provided in Definition G.1. The convergence of the multivariate INAR process to the multivariate Hawkes process, along with the correspondence of their parameters, is presented in Theorem G.5. Together, these results validate the discrete-time linear formulation stated in Theorem 3.1. This completes the proof.

□

# H Proof of Lemma 3.2

*Proof.* The proof of Lemma 3.2 is based on Proposition 2.2 and Theorem 2.4 from [47], which we restate here for completeness.

**Proposition H.1** (Rank Characterization of Conditional Independence [47]). *Let $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ be a multivariate normal random vector, and let $A$, $B$, and $C$ be disjoint subsets of indices. Then the conditional independence statement $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_C$ holds if and only if the cross-covariance matrix $\Sigma_{A\cup C, B\cup C}$ has rank $|C|$.*

Although this result was originally established for linear acyclic models with independent Gaussian noise, it relies solely on second-order properties (variance and covariance) of the data and leverages path analysis rooted in the independence of noise terms. Consequently, this result remains valid for linear models with arbitrary noise distributions, since the argument applies to any distribution with finite second moments.

**Theorem H.2** (Conditional Independence in Directed Graphical Models [47]). *In a directed graph $\mathcal{G}$, a set $C$ d-separates $A$ and $B$ if and only if the conditional independence statement $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_C$ holds for every distribution that is Markov with respect to $\mathcal{G}$.*

Combining the two results, we obtain the following: For any linear acyclic causal model with disjoint variable sets $\mathbf{A}_v$, $\mathbf{B}_v$, and $\mathbf{C}_v$, the set $\mathbf{C}_v$ d-separates $\mathbf{A}_v$ and $\mathbf{B}_v$ in the associated causal graph if and only if:

$$\operatorname{rank}(\Sigma_{\mathbf{A}_v\cup\mathbf{C}_v, \mathbf{B}_v\cup\mathbf{C}_v}) = |\mathbf{C}_v|.$$

This equivalence confirms that the d-separation criterion in the causal graph corresponds to a rank condition on the cross-covariance matrix $\Sigma_{\mathbf{A}_v\cup\mathbf{C}_v, \mathbf{B}_v\cup\mathbf{C}_v}$.

Since the window causal graph in PO-MHP is a DAG with linear causal relations and serially uncorrelated white noise, the above rank condition applies directly to the window causal graph in the PO-MHP framework. This completes the proof. □

# I Proof of Proposition 3.3

*Proof.* For any subprocess $O_1$, we prove the equivalence of the four statements step by step.

**(1) ⇔ (2):** If $\mathcal{P}_\mathcal{G}$ is the parent cause set of $O_1$ in the summary graph, by construction of the window causal graph, it equivalent to that the corresponding lagged variable set $\mathbf{P}_v$ contains all direct parent variables of $O_1^{(n)}$. This follows from the fact that, in the window graph, directed edges exist from the effective lag variables of each parent subprocess to $O_1^{(n)}$. Moreover, by definition of the parent cause set, $\mathcal{P}_\mathcal{G}$ is minimal with this property.

**(2) ⇔ (3):** If $\mathbf{P}_v$ contains all direct parents of $O_1^{(n)}$ in the window graph, by the Markov property of DAGs, $\mathbf{P}_v$ d-separates $O_1^{(n)}$ from all other observed variables in $\mathbf{O}_v \setminus \left(\mathbf{P}_v \cup \{O_1^{(n)}\}\right)$. Reversly, if $\mathbf{P}_v$ d-separates $O_1^{(n)}$ from all other observed variables in $\mathbf{O}_v \setminus \left(\mathbf{P}_v \cup \{O_1^{(n)}\}\right)$, by the Granger causality-events in the future cannot causally influence events in the past, $\mathbf{P}_v$ should contain all direct parents of $O_1^{(n)}$ in the window graph. Moreover, by definition of the parent cause set, $\mathcal{P}_\mathcal{G}$ is minimal with this property.

**(3) ⇔ (4):** By applying Lemma 3.2, the d-separation between $O_1^{(n)}$ and the rest of the variables, conditioned on $\mathbf{P}_v$, is equivalent to the rank constraint:

$$\operatorname{rank}\left(\Sigma_{\{O_1^{(n)}\}\cup\mathbf{P}_v,\ \mathbf{O}_v\setminus\{O_1^{(n)}\}}\right) = |\mathbf{P}_v|.$$

**(4) ⇔ (1):** Assume the rank condition holds for $\mathbf{P}_v$. By Lemma 3.2, this implies that $\mathbf{P}_v$ d-separates $O_1^{(n)}$ from all other variables in the window graph. Translating back to the summary graph, this implies that $\mathcal{P}_\mathcal{G}$ is the minimal parent cause set of $O_1$, as no smaller set can block all paths to $O_1$.

Thus, all statements are equivalent. This completes the proof. □

# J Preliminaries for Proofs of Proposition 3.5 and Theorems F.2 and F.3

To establish this result, we rely on the concepts of trek separation (t-separation) and d-separation introduced by [47], which provide powerful tools for analyzing latent structures in linear causal models.

**Definition J.1** (Trek [47])**.** A trek in the DAG $\mathcal{G}$ from variable $V_i$ to variable $V_j$ is an ordered pair of directed paths $(\mathbf{P_1}, \mathbf{P_2})$ where $\mathbf{P_1}$ has sink $V_i$, $\mathbf{P_2}$ has sink $V_j$, and both $\mathbf{P_1}$ and $\mathbf{P_2}$ have the same source $V_k$. The common source $V_k$ is called the top of the trek, denoted top$(\mathbf{P_1}, \mathbf{P_2})$. Note that one or both of $\mathbf{P_1}$ and $\mathbf{P_2}$ may consist of a single variable, that is, a path with no edges. A trek $(\mathbf{P_1}, \mathbf{P_2})$ is *simple* if the only common variable among $\mathbf{P_1}$ and $\mathbf{P_2}$ is the common source top$(\mathbf{P_1}, \mathbf{P_2})$. We let $\mathcal{T}(V_i, V_j)$ and $\mathcal{S}(V_i, V_j)$ denote the sets of all treks and all simple treks from $V_i$ to $V_j$, respectively.

**Definition J.2** (T-separation [47])**.** Let $\mathbf{A}_v$, $\mathbf{B}_v$, $\mathbf{C_A}$, and $\mathbf{C_B}$ be four subsets of total variable set $\mathbf{V}_v$. We say the ordered pair $(\mathbf{C_A}, \mathbf{C_B})$ t-separates $\mathbf{A}_v$ from $\mathbf{B}_v$ if, for every trek $(\tau_1; \tau_2)$ from a variable in $\mathbf{A}_v$ to a variable in $\mathbf{B}_v$, either $\tau_1$ contains a variable in $\mathbf{C_A}$ or $\tau_2$ contains a variable in $\mathbf{C_B}$.

**Theorem J.3** (Trek separation for directed graphical models [47])**.** *The sub-matrix $\sum_{\mathbf{A},\mathbf{B}}$ has rank less than equal to $r$ for all covariance matrices consistent with the graph $\mathcal{G}$ if and only if there exist subsets $\mathbf{C}_A, \mathbf{C}_B \subset \mathbf{V}_G$ with $|\mathbf{C}_A| + |\mathbf{C}_B| \leq r$ such that $(\mathbf{C}_A, \mathbf{C}_B)$ t-separates $\mathbf{A}$ from $\mathbf{B}$. Consequently,*

$$\operatorname{rank}(\Sigma_{\mathbf{A},\mathbf{B}}) \leq \min\{|\mathbf{C}_A| + |\mathbf{C}_B| : (\mathbf{C}_A, \mathbf{C}_B) \text{ t-separates } \mathbf{A} \text{ from } \mathbf{B}\}$$

*and equality holds for generic covariance matrices consistent with $\mathcal{G}$.*

**Corollary J.4** (T-separation and D-separation [47])**.** *A set $\mathbf{C}$ d-separates $\mathbf{A}$ and $\mathbf{B}$ in $\mathcal{G}$ if and only if there is a partition $\mathbf{C} = \mathbf{C}_A \cup \mathbf{C}_B$ such that $(\mathbf{C}_A, \mathbf{C}_B)$ t-separates $\mathbf{A} \cup \mathbf{C}$ from $\mathbf{B} \cup \mathbf{C}$.*

Therefore, when $\mathbf{C_A}$ and $\mathbf{C_B}$ are disjoint, the combined set $\mathbf{C_A} \cup \mathbf{C_B}$ also serves as a d-separator between $\mathbf{A}$ and $\mathbf{B}$. Moreover, since the window graph in the Hawkes process is a DAG with linear relations, the above results can be directly applied after suitable adaptation to the Hawkes process setting.

# K Proof of Proposition 3.5

*Proof.* We prove both directions of the equivalence.

**($\Leftarrow$) If such a latent confounder $L_1$ exists, the rank condition holds.** Suppose there exists a latent confounder $L_1$ that is one common parent cause in the parent cause set of $\{O_1, O_2\}$, and that $L_1$ together with $\{O_1, O_2\}$ makes them locally independent of other subprocesses.

Given that $L_1$ and its paths to $O_1$ and $O_2$ satisfy Definition 3.4, the contribution of $L_1$ to both $O_1$ and $O_2$ in the window graph occurs through the same number of latent intermediates, resulting in an aligned contribution across time lags. In this setup, the influence of $L_1$ will appear as a shared component across the observed variables $\{O_i^{(j)}\}_{j\in\{n-m,\ldots,n\}}^{i\in\{1,2\}}$.

Consider the window graph with $m$ considered effective lag variables. Following the logic of trek separation, in the window graph with $m$ effective lag variables, the minimal choke set $\mathbf{C_A}$ that t-separates $O_1^{(n)}, O_2^{(n)}$ from the rest is given by:

$$\mathbf{C}_A \coloneqq \{L_1^{(j)}\}_{j\in\{n-m,\ldots,n-1\}} \cup \{O_i^{(n)}\}_{i\in\{1,2\}}^{j\in\{n-m,\ldots,n-1\}}.$$

It is equivalent to that $\mathbf{C}_A$ is the minimal set that d-separates $\{O_1^{(n)}, O_2^{(n)}\}$ from the $\mathbf{O}_v \backslash \{O_1^{(n)}, O_2^{(n)}\}$.

Thus, by Theorem J.3, the generic rank of the cross-covariance matrix is bounded above by $|\mathbf{C_A}| = 2m + m = 3m$, where $2m$ comes from observed lag variables of $\{O_1, O_2\}$ and $m$ comes from latent lag variables of $L_1$. However, due to the structure of the excitation function $\phi_{ij}(s) = a_{ij}w(s)$, the latent subprocess $L_1$ contributes effectively as a *single* shared component across all its lag variables, reducing the effective rank from $m$ to 1.

To explain this, we first write the structural equations for the observed variables $\{O_i^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\dots,n\}}$ as the linear regression on those check points as:

$$\begin{bmatrix} O_1^{(n)} \\ O_1^{(n-1)} \\ \cdots \\ O_1^{(n-m)} \\ O_2^{(n)} \\ O_2^{(n-1)} \\ \cdots \\ O_2^{(n-m)} \end{bmatrix} = \mathbf{E} \begin{bmatrix} L_1^{(n-1)} \\ \cdots \\ L_1^{(n-m)} \\ O_1^{(n-1)} \\ \cdots \\ O_1^{(n-m)} \\ O_2^{(n-1)} \\ \cdots \\ O_2^{(n-m)} \end{bmatrix} + \begin{bmatrix} \epsilon_{o_1}^{(n)} + \theta_{o_1}^{(0)} \\ \epsilon_{o_1}^{(n-1)} + \theta_{o_1}^{(0)} \\ \cdots \\ \epsilon_{o_1}^{(n-m)} + \theta_{o_1}^{(0)} \\ \epsilon_{o_2}^{(n)} + \theta_{o_1}^{(0)} \\ \epsilon_{o_2}^{(n-1)} + \theta_{o_1}^{(0)} \\ \cdots \\ \epsilon_{o_2}^{(n-m)} + \theta_{o_1}^{(0)} \end{bmatrix}, \tag{10}$$

$$\mathbf{E} = \begin{bmatrix} a_{o_1 l_1} \int_0^{\Delta} w(s)ds & \cdots & a_{o_1 l_1} \int_{(m-1)\Delta}^{m\Delta} w(s)ds & 1 & \overset{1}{\cdots} & 1 & 0 & \overset{0}{\cdots} & 0 \\ 0 & \overset{0}{\cdots} & 0 & 1 & \overset{0}{\cdots} & 0 & 0 & \overset{0}{\cdots} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \overset{0}{\cdots} & 0 & 0 & \overset{0}{\cdots} & 1 & 0 & \overset{0}{\cdots} & 0 \\ a_{o_2 l_1} \int_0^{\Delta} w(s)ds & \cdots & a_{o_2 l_1} \int_{(m-1)\Delta}^{m\Delta} w(s)ds & 0 & \overset{0}{\cdots} & 0 & 1 & \overset{1}{\cdots} & 1 \\ 0 & \overset{0}{\cdots} & 0 & 0 & \overset{0}{\cdots} & 0 & 1 & \overset{0}{\cdots} & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \overset{0}{\cdots} & 0 & 0 & \overset{0}{\cdots} & 0 & 0 & \overset{0}{\cdots} & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \}\text{m} \\ \\ \\ \\ \}\text{m} \end{matrix}. \tag{11}$$

It is straightforward to see that the rank of the coefficient matrix $\mathbf{E}$ is $2m+1$, because the two row corresponding to $O_1^{(n)}$ and $O_2^{(n)}$ in $\mathbf{E}$ are linearly dependent (proportional to each other).

Furthermore, the cross-covariance matrix of $\{O_i^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\dots,n\}}$ and $\mathbf{O}_v\backslash\{O_1^{(n)}, O_2^{(n)}\}$, i.e., $\Sigma_{\{O_i^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\dots,n\}},\, \mathbf{O}_v\backslash\{O_1^{(n)},O_2^{(n)}\}}$ can be written as $\mathbf{E}\mathbf{C}_A\mathbf{C}_A^{\top}\mathbf{F}^{\top}$ where $\mathbf{E}$ and $\mathbf{F}$ are coefficient matrix by regressing variables on those choke points. The $\mathrm{rank}(\mathbf{C}_A\mathbf{C}_A^{\top}\mathbf{F}^{\top})$ has full column rank, because $\mathbf{F}$ calculated from regressing all the rest variables $\mathbf{O}_v\backslash\{O_1^{(n)}, O_2^{(n)}\}$ on $\mathbf{C}_A$ and without blocking lagged variables, no shrinkage of rank occurs. Consequently, the rank of the cross-covariance matrix $\mathrm{rank}\left(\Sigma_{\{O_i^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\dots,n\}},\, \mathbf{O}_v\backslash\{O_1^{(n)},O_2^{(n)}\}}\right) = rank\left(\mathbf{E}\mathbf{C}_A\mathbf{C}_A^{\top}\mathbf{F}^{\top}\right) = \mathrm{rank}(\mathbf{E}) = 2m+1$ (The following theorem proofs also adopt a similar way).

Thus, the total rank becomes:

$$\mathrm{rank} = 2m \text{ (from observed lags of } O_1 \text{ and } O_2) + 1 \text{ (from } L_1) = 2m+1.$$

**($\Rightarrow$) If the rank condition holds, there exists a latent confounder $L_1$ satisfying the claimed properties.** Conversely, assume the observed rank condition:

$$\mathrm{rank}\left(\Sigma_{\{O_i^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\dots,n\}},\, \mathbf{O}_v\backslash\{O_1^{(n)},O_2^{(n)}\}}\right) = 2m+1.$$

By construction of the window graph (Eq. 2), if there were no latent confounder between $O_1$ and $O_2$, the rank would be at most $2m$, corresponding to the observed lag variables of $O_1$ and $O_2$. The observed rank being strictly $2m+1$ thus implies the presence of an additional latent variable influencing both $O_1$ and $O_2$.

Due to the rank faithfulness assumption (Assumption 2), such a rank elevation uniquely corresponds to a latent subprocess $L_1$ acting as a parent cause of both $O_1$ and $O_2$. Furthermore, for the rank increment to be exactly one, the causal paths from $L_1$ to $O_1$ and $O_2$ must satisfy the *symmetric path situation* (Definition 3.4): i.e., the paths only involve intermediate latent subprocesses of the same depth without self-loops, ensuring that the contribution of $L_1$ introduces a single additional rank component shared by both $O_1$ and $O_2$ at the same temporal lag level.

Finally, by construction, conditioning on $\mathcal{P}'_{\mathcal{G}} \coloneqq L_1 \cup \{O_1, O_2\}$ removes all causal influence from $L_1$, rendering $\{O_1, O_2\}$ locally independent of the remaining observed subprocesses.

This completes the proof. □

# L Proof of Theorem F.2

*Proof.* We prove both directions of the equivalence.

**(⇐) If such a parent cause set $\mathcal{P}'_{\mathcal{G}}$ exists, the rank condition holds.** Assume that $\mathcal{P}'_{\mathcal{G}}$ is the minimal set of subprocesses such that:

- $\mathcal{P}'_{\mathcal{G}}$ is a subset of the parent cause set of $N_1$.
- Conditioning on $\mathcal{S}_{\mathcal{G}} \coloneqq \mathcal{P}'_{\mathcal{G}} \cup \mathcal{D}e(N_1) \cup \{\mathcal{D}e(L_i)\}_{L_i \in \mathcal{P}'_{\mathcal{G}}} \cup \mathcal{S}ib(\mathcal{D}e(N_1)) \cup \{\mathcal{S}ib(\mathcal{D}e(L_i))\}_{L_i \in \mathcal{P}'_{\mathcal{G}}}$ renders $N_1$ locally independent of all other subprocesses in the system.
- All possible observed surrogates of $N_i$ in $\mathcal{O}_{\mathcal{G}}$ have been identified.
- For each $L_i \in \mathcal{P}'_{\mathcal{G}}$, the relationship between $L_i$ and its observed effects $\{\mathcal{D}e(N_1), \mathcal{D}e(L_i)\}$ satisfies Definition 3.4.

In this setup, the lagged variables of $\mathcal{D}e(N_1)$ and $\mathcal{D}e(L_i)$, as well as the lagged and current variables of their observed siblings $\mathcal{S}ib(\mathcal{D}e(N_1))$ and $\mathcal{S}ib(\mathcal{D}e(L_i))_{L_i \in \mathcal{P}'_{\mathcal{G}}}$, appear in both $\mathbf{A}_v$ and $\mathbf{B}_v$. The rank contribution from these *observed variables* is deterministically: $|\mathbf{O}_{v1}| \coloneqq \left| \{\mathcal{D}e(N_1)^{(j)}, \mathcal{D}e(L_i)^{(j)}\}_{L_i \in \mathcal{P}'_{\mathcal{G}}}^{j \in \{n-m, \dots, n-1\}} \cup \{O_i^{(j)}\}_{O_i \in \mathcal{P}'_{\mathcal{G}}}^{j \in \{n-m, \dots, n-1\}} \cup \{O_i^{(j)}\}_{O_i \in \mathcal{S}ib(\mathcal{D}e(N_1)) \cup \{\mathcal{S}ib(\mathcal{D}e(L_i))\}_{L_i \in \mathcal{P}'_{\mathcal{G}}}}^{j \in \{n-m, \dots, n\}} \right|$. The remaining part of $\mathbf{A}_v$, i.e., $\mathbf{A}_v \backslash \mathbf{O}_{v1}$, consists of the current variables $\{\mathcal{D}e(N_1)^{(n)}, \mathcal{D}e(L_i)^{(n)}\}_{L_i \in \mathcal{P}'_{\mathcal{G}}}$.

Given the symmetric path structure (Definition 3.4), each latent confounder $L_i \in \mathcal{P}'_{\mathcal{G}}$ contributes exactly one shared latent component, as the influence propagates through symmetric, acyclic paths. Due to the specific excitation function $\phi_{ij}(s) = a_{ij} w(s)$, this results in precisely one rank contribution per latent subprocess, regardless of the number of lagged variables.

Thus, the latent contribution adds exactly:

$$|\mathbf{O}_{v2}| \coloneqq \left| \{L_i\}_{L_i \in \mathcal{P}'_{\mathcal{G}}} \right| = \left| \mathcal{D}e(L_i)^{(n)}\}_{L_i \in \mathcal{P}'_{\mathcal{G}}} \right|$$

rank-one components.

Combining both observed and latent contributions, the total rank becomes:

$$|\mathbf{O}_{v1}| + |\mathbf{O}_{v2}| = |\mathbf{O}_{v1}| + |\mathbf{O}_{v2}| + 1 \,(\text{from } \mathcal{D}e(N_1)^{(n)}) - 1 = |\mathbf{A}_v| - 1.$$

**(⇒) The rank condition implies the claimed causal structure and local independence.** Assume that $\mathcal{P}'_{\mathcal{G}}$ is the minimal set such that:

$$\operatorname{rank}\left(\Sigma_{\mathbf{A}_v, \mathbf{B}_v}\right) = |\mathbf{A}_v| - 1$$

By the theory of trek separation (Theorem J.3), such a rank deficiency implies that the information flow between $\mathbf{A}_v$ and $\mathbf{B}_v$ must pass through a set of choke points, corresponding to the candidate parent causes in $\mathcal{P}'_{\mathcal{G}}$.

If no latent confounders existed, or if $\mathcal{P}'_{\mathcal{G}}$ were not part of the parent cause set of $N_1$, the rank would be exactly $|\mathbf{O}_{v1}|$, solely contributed by the lagged variables of observed surrogates and both the current and lagged variables of their siblings.

Since all possible observed surrogates of $N_i$ in $\mathcal{O}_{\mathcal{G}}$ have been identified, the extra deficiency of rank (i.e., $|\mathbf{O}_{v2}|$) thus directly implies the existence of latent subprocesses contributing shared rank-one components. By the rank faithfulness, this observed rank pattern is only consistent with the existence of latent subprocesses $\{L_i\}_{L_i \in \mathcal{P}'_{\mathcal{G}}}$ that act as confounders between $\mathcal{D}e(N_1)$ and their respective observed effects, and these latent subprocesses are members of the parent cause set of $N_1$.

For the rank deficit per latent subprocess to be exactly one, the contribution from each latent subprocess must propagate through symmetric acyclic paths, consistent with Definition 3.4, ensuring a single rank-one component contribution per latent subprocess. Moreover, the inclusion of the observed surrogates and their siblings ensures that no alternative paths can explain the dependency patterns. Thus, $\mathcal{P}'_{\mathcal{G}}$ must be the subset of parent causes, satisfying the conditional local independence of $N_1$ given $\mathcal{S}_{\mathcal{G}}$.

Therefore, the rank condition is both necessary and sufficient to identify $\mathcal{P}'_{\mathcal{G}}$ as the subset of parent causes of $N_1$, considering both observed and latent subprocesses. This completes the proof. □

# M Proof of Theorem F.3

*Proof.* We prove both directions of the equivalence.

**(⇐) If such a latent confounder $L_1$ exists, the rank condition holds.** Assume there exists a latent confounder subprocess $L_1$ such that:

- $L_1$ is a common parent cause of $\{N_1, N_2\}$.
- Conditioning on $\mathcal{P}'_{\mathcal{G}} \coloneqq L_1 \cup N_1, N_2 \cup \mathcal{S}ib(\mathcal{D}e(N_i))_{i\in\{1,2\}}$ renders $\{N_1, N_2\}$ locally independent of the rest of the system $\mathcal{N}_{\mathcal{G}} \backslash \mathcal{P}'_{\mathcal{G}}$.
- All possible observed surrogates of $\{N_1, N_2\}$ in $\mathcal{O}_{\mathcal{G}}$ have been identified.
- $L_1$ and its observed effects $\{\mathcal{D}e(N_1), \mathcal{D}e(N_2)\}$ satisfy Definition 3.4.

By the Definition 3.4, the causal influence from $L_1$ to $\{\mathcal{D}e(N_1), \mathcal{D}e(N_2)\}$ is symmetric and only propagates through the same number of intermediate latent subprocesses without self-loops. Under this condition, the contributions of $L_1$ to the observed surrogates $\{\mathcal{D}e(N_1), \mathcal{D}e(N_2)\}$ appear as a rank-one component across the lagged variables of these subprocesses, aligned in time.

Thus, in the window graph, the latent influence from $L_1$ will introduce exactly one additional rank component across the observed variable set $\mathbf{A}_v$ beyond the rank contribution from the observed lagged variables themselves.

Formally, following the arguments for Proposition 3.5, the rank of $\Sigma_{\mathbf{A}_v,\mathbf{B}_v}$ is determined by the minimal set of choke points that t-separate $\mathbf{A}_v$ from $\mathbf{B}_v$ in the window graph. Given the assumed structure:

- The lagged variables of $\{\mathcal{D}e(N_1), \mathcal{D}e(N_2)\}$ and both the current and lagged variables of their observed siblings, denoted as $\mathbf{O}_{v1} \coloneqq \{\mathcal{D}e(N_i)^{(j)}\}_{i\in\{1,2\}}^{j\in\{n-m,\ldots,n-1\}} \cup \{O_i^{(j)}\}_{O_i\in\mathbf{Sib}(\mathcal{D}e(N_1))\cup\mathbf{Sib}(\mathcal{D}e(N_2))}^{j\in\{n-m,\ldots,n\}}$, appear in both $\mathbf{A}_v$ and $\mathbf{B}_v$, contributing deterministically $|\mathbf{O}_{v1}|$ to the rank.
- The influence from $L_1$ propagates symmetrically to both $\mathcal{D}e(N_1)$ and $\mathcal{D}e(N_2)$ through acyclic paths composed exclusively of latent subprocesses, per Definition 3.4. As a result, due to the excitation function $\phi_{ij}(s) = a_{ij}w(s)$, the total rank contribution from $L_1$ is exactly one.

Therefore, the total rank becomes:

$$\operatorname{rank}\left(\Sigma_{\mathbf{A}_v,\mathbf{B}_v}\right) = |\mathbf{O}_{v1}| + 1 = |\mathbf{A}_v| - 1$$

**(⇒) If the rank condition holds, such a latent confounder $L_1$ must exist.** Now assume the observed rank condition:

$$\operatorname{rank}\left(\Sigma_{\mathbf{A}_v,\mathbf{B}_v}\right) = |\mathbf{A}_v| - 1$$

We know that parent cause sets of all inferred latent confounder processes in $\mathcal{N}_{\mathcal{G}}$ remain unidentified even after applying Theorem F.2. In the absence of any new latent confounder, the maximum possible rank would be $|\mathbf{O}_{v1}|$, corresponding solely to the contributions of the lagged variables of $\{\mathcal{D}e(N_1), \mathcal{D}e(N_2)\}$ and both the current and lagged variables of their observed siblings. The observed rank being exactly $|\mathbf{O}_{v1}| + 1 = |\mathbf{A}_v| - 1$ implies the existence of an additional latent source influencing both $N_1, N_2$ and their observed surrogates.

Due to the rank faithfulness, this increment must be attributed to a unique latent subprocess $L_1$ that acts as a confounder for $N_1$ and $N_2$. Moreover, the fact that the rank increment is only one implies that the paths from $L_1$ to $N_1, N_2$ must satisfy the symmetric and acyclic conditions in Definition 3.4, ensuring that the influence of $L_1$ is captured as a rank-one shared component at the observed surrogates level.

Moreover, the inclusion of the observed surrogates and their siblings ensures that all other possible paths and confounding structures are blocked, enforcing $\mathcal{P}'_{\mathcal{G}} := L_1 \cup \{N_1, N_2\} \cup \{\mathcal{S}ib(\mathcal{D}e(N_i))\}_{i\in\{1,2\}}$ in ensuring local independence and all possible observed surrogates of $\{N_1, N_2\}$ in $\mathcal{O}_{\mathcal{G}}$ have been identified.

Thus, the rank pattern is both necessary and sufficient to imply the existence of $L_1$ and the claimed causal and conditional independence structure. This completes the proof. □

# N Proof of Theorem 4.1

*Proof.* We prove the theorem by considering the two cases separately: (i) the system contains no latent subprocesses, and (ii) the system contains latent subprocesses that satisfy Definition 3.4.

**Case (i): No latent subprocesses.** In this case, the system consists solely of observed subprocesses $\mathcal{O}_{\mathcal{G}}$. Since there are no latent confounders, Phase I alone is sufficient for identifiability. This follows directly from Proposition 3.3, which ensures that for each observed subprocess, its parent cause set can be uniquely identified by checking the rank condition of the relevant cross-covariance matrices. Specifically, since all subprocesses are observed and no latent subprocesses confound their relationships, the rank condition provides a unique solution. Thus, the entire causal graph can be identified solely through Phase I.

**Case (ii): Presence of latent subprocesses satisfying Definition 3.4.** In the general case where latent subprocesses exist, the algorithm relies on the synergy between Phase I and Phase II.

- **Phase I** iteratively identifies the parent cause set for each subprocess (including both observed and previously discovered latent subprocesses) whose parent cause set is fully contained in the current set of known subprocesses. By Proposition 3.3 and Theorem F.2, this identification is guaranteed when no latent confounders intervene or when latent confounders are already represented by their observed surrogates.
- **Phase II** handles the discovery of new latent confounder subprocesses by systematically applying Proposition 3.5 and Theorem F.3. The identifiability is guaranteed under the condition that all latent confounders and their associated observed effects satisfy Definition 3.4. This condition ensures that each latent confounder contributes a unique, identifiable rank-1 pattern in the cross-covariance matrix of its observed surrogates and their siblings, enabling its detection through the rank conditions established in the theorems.

**Termination and completeness.** The algorithm alternates between Phase I and Phase II. Since each iteration either identifies a new parent cause set or discovers a new latent subprocess, and given the finite number of subprocesses (including latent ones), the algorithm must eventually terminate.

By construction:

- All observed subprocesses will eventually have their parent cause sets identified through Phase I.
- All latent subprocesses satisfying Definition 3.4 will be identified through Phase II and incorporated into the active set for further investigation.
- The recursive application of the identification theorems ensures that no causal relationships (either between observed, latent, or between observed and latent) will remain unidentified under the conditions.
- If Definition 3.4 fails for any latent, the algorithm terminates without fabricating that latent or any edges it would entail, thereby returning only the identifiable portion of the causal graph (sound abstention).

Thus, under excitation function $\phi_{ij}(s) = a_{ij}w(s)$ and rank faithfulness, the entire causal graph consisting of both observed subprocesses and latent confounders can be identified. This completes the proof. □

# O Details of identification algorithm

## O.1 Phase I

The detailed algorithm for Phase I is in Algorithm 2.

**Algorithm 2** Identifying Causal Relations

**Input:** Partial causal graph $\mathcal{G}$, Active subprocess set $\mathcal{A}_\mathcal{G}$, Observed subprocess set $\mathcal{O}_\mathcal{G}$
**Output:** Partial causal graph $\mathcal{G}$, Active subprocess set $\mathcal{A}_\mathcal{G}$

1: **repeat**
2: Select a subprocess $N_1$ from $\mathcal{A}_\mathcal{G}$.
3: **for** $Len = 1$ **to** $|\mathcal{A}_\mathcal{G} \cup \mathcal{O}_\mathcal{G}|$ **do**
4: **repeat**
5: Select subset $\mathcal{P}'_\mathcal{G} \subseteq \mathcal{A}_\mathcal{G} \cup \mathcal{O}_\mathcal{G}$ such that $|\mathcal{P}'_\mathcal{G}| = Len$.
6: **if** $(\mathcal{A}_\mathcal{G} \cup \mathcal{O}_\mathcal{G}, \mathcal{P}'_\mathcal{G}, N_1)$ satisfies Proposition 3.3 and Theorem F.2 **then**
7: $\mathcal{A}_\mathcal{G} = \mathcal{A}_\mathcal{G} \backslash N_1$, and update $\mathcal{G}$.
8: Return to line 2.
9: **end if**
10: **until** All subsets of $\mathcal{A}_\mathcal{G} \cup \mathcal{O}_\mathcal{G}$ with size $Len$ selected.
11: **end for**
12: **until** $\mathcal{A}_\mathcal{G}$ is not updated or $|\mathcal{A}_\mathcal{G}| \leq 1$.
13: **return:** $\mathcal{G}$, $\mathcal{A}_\mathcal{G}$

## O.2 Phase II

The detailed algorithm for Phase II is in Algorithm 3.

**Algorithm 3** DiscoveringNewLatentComponentProcesses

**Input:** Partial causal graph $\mathcal{G}$, Active subprocess set $\mathcal{A}_\mathcal{G}$, Observed subprocess set $\mathcal{O}_\mathcal{G}$
**Output:** Partial causal graph $\mathcal{G}$, Active subprocess set $\mathcal{A}_\mathcal{G}$

1: Initialize cluster set $\mathbb{C} := \emptyset$ and the group size $Len = 2$.
2: **repeat**
3: Select a subset $\mathcal{Y}_\mathcal{G}$ from $\mathcal{A}_\mathcal{G}$ such that $|\mathcal{Y}_\mathcal{G}| = Len$.
4: **if** $(\mathcal{A}_\mathcal{G} \cup \mathcal{O}_\mathcal{G}, \mathcal{Y}_\mathcal{G})$ satisfies Proposition 3.5 and Theorem F.3 **then**
5: Add $\mathcal{Y}_\mathcal{G}$ into $\mathbb{C}$.
6: **end if**
7: **until** All subset of $\mathcal{A}_\mathcal{G}$ with size $Len$ selected.
8: Merge all the overlapping sets in $\mathbb{C}$.
9: **for** each merged set $\mathcal{C}_i \in \mathbb{C}$ **do**
10: Introduce a new latent subprocess $L_j$.
11: $\mathcal{A}_\mathcal{G} = \mathcal{A}_\mathcal{G} \cup L_j \backslash \mathcal{C}_i$, and update $\mathcal{G}$.
12: **end for**
13: **return:** $\mathcal{G}$, $\mathcal{A}_\mathcal{G}$

# P Computational Complexity of the Algorithm

In this section, we analyze the computational complexity of our two-phase iterative algorithm, which alternates between: (1) inferring causal relationships among discovered subprocesses and (2) identifying new latent subprocesses. Let $n$ denote the number of processes in the *active process set* $\mathcal{A}_\mathcal{G}$ and $m$ denote the number of subprocesses in the augmented process set $\mathcal{T}_\mathcal{G} := \mathcal{A}_\mathcal{G} \cup \mathcal{O}_\mathcal{G}$ at the start of each phase. Assume each test is an oracle test.

**Phase I: Inferring Causal Relationships**

For each component process $N_1 \in \mathcal{A}_\mathcal{G}$, we evaluate subsets of $\mathcal{T}_\mathcal{G}$ starting from subsets of size 1 up to the size of $\mathcal{T}_\mathcal{G}$, stopping when the test result is positive. In the worst case, for each $N_1$, we need to

evaluate all subsets of $\mathcal{T}_{\mathcal{G}}$, which requires $\sum_{k=1}^{m} \binom{m}{k}$ tests. For one subprocess $N_1 \in \mathcal{A}_{\mathcal{G}}$, if its parent cause set is found, $\mathcal{A}_{\mathcal{G}}$ is updated. After that, the algorithm will restart to go over all the subprocesses in $\mathcal{A}_{\mathcal{G}}$ to make sure no parent cause set of subprocesses in $\mathcal{A}_{\mathcal{G}}$ can be found. In the worst case, the algorithm find parent cause set for the last component process in $\mathcal{A}_{\mathcal{G}}$ each time. The complexity of Phase I is upper bounded by: $\mathcal{O}\left(n! \sum_{k=1}^{m} \binom{m}{k}\right)$.

**Phase II: Identifying New Latent Subprocesses**

In this phase, we test all subsets of $\mathcal{A}_{\mathcal{G}}$ of size 2. Since there are $\binom{n}{2}$ such subsets, the complexity of Phase II is upper bounded by: $\mathcal{O}\left(\binom{n}{2}\right)$.

**Overall Complexity**

The total complexity of the algorithm depends on the number of (both observed and latent) subprocesses and the structural density of the causal graph, as these factors determine the number of iterations required for the algorithm to run. Combining the two phases, for each iteration, the overall complexity is approximately upper bounded by: $\mathcal{O}\left(n! \sum_{k=1}^{m} \binom{m}{k} + \binom{n}{2}\right)$.

In practical scenarios, the structural density of the causal graph and sparsity of dependencies may reduce the number of required iterations and tests, leading to improved efficiency compared to this worst-case analysis.

# Q More Details of Experiments

## Q.1 Synthetic Data Generation and Implementation

We evaluate our method on two types of synthetic data: event sequences generated by the Hawkes process in Eq. (1), and discrete-time data generated directly from the discrete-time model in Eq. (2)

**Hawkes Process Data**: We generate event sequences using the `tick` library [2], an efficient framework for simulating multivariate Hawkes processes. The excitation function is set as exponential kernel $\phi_{ij}(s) = \alpha_{ij} e^{-\beta s}$, where $\beta$ is fixed at 1. $\alpha_{ij}$ is sampled uniformly from $[0.8, 0.99]$ except for Case 1. Because of the cycles between $N_2$ and $N_3$ of Case 1, large $\alpha_{ij}$ may lead to nonstationarity. Thus, we sample $\alpha_{ij}$ uniformly from $[0.40, 0.80]$ specifically for Case 1. To ensure stationarity and avoid explosive behavior, we verify the spectral radius of the integrated excitation matrix after generating $\alpha_{ij}$. To discretize the sequences for our method, we select the time bins of length 0.1 and consider 600 effective lag time bins as discretized lag variables for the calculation sub-covariance matrix. The sample size corresponds to the number of discrete data points.

**Discrete-Time Series Data**: To assess our method under ideal discrete-time conditions (i.e., exactly satisfying Theorem 3.1), we generate data directly from Eq. (2). The excitation function is set as exponential kernel $\phi_{ij}(s) = \alpha_{ij} e^{-\beta s}$. The coefficients $\alpha_{ij}$ and decay parameter $\beta$ are set as above. Similar to the Hawkes data, we verify the spectral radius to ensure stationarity. The noise terms are drawn from independent Gaussian distributions. We set the number of effective lag variables to 200. The sample size corresponds to the number of discrete data points.

**Preprocessing and Rank Deficiency Testing**: For each trial, we standardize the discretized data to ensure fair comparison. To test for rank deficiency, we use canonical correlation analysis (CCA) [49], following the procedure in [21]. We use the grid search to find the best rank test threshold. We also conduct a empirical sensitivity analysis for test threshold. The result is in Appendix Q.3. A threshold of 0.10 provides a good balance across multiple scenarios.

**Data Usage for Baselines**: For Hawkes process-based methods (SHP [37], THP [6], and NPHC [1]), we use the raw Hawkes process data produced by the `tick` library. For rank-based methods designed for i.i.d. data with linear relations (Hier. Rank [21] and RLCD [13]), we use the discretized Hawkes process data.

We run all the experiments on a personal PC (CPU).

## Q.2 Evaluation Metrics

We evaluate the accuracy of causal structure recovery using the standard F1-score, which combines precision and recall.

Causal relationships among both latent and observed subprocesses are represented by an adjacency matrix, where each entry is either 1 or 0, indicating the presence or absence of a directed edge, respectively. Specifically, $\text{Adj}\mathcal{G}(i, j) = 1$ denotes a directed edge from the $j$-th subprocess to the $i$-th subprocess, while $\text{Adj}\mathcal{G}(i, j) = 0$ indicates no such edge.

We measure the similarity between the estimated and ground-truth adjacency matrices using the F1-score. First, we compute precision, defined as

$$\text{precision} = \frac{\text{true positives}}{\text{total inferred positives}},$$

which represents the proportion of correctly inferred edges among all predicted edges. Next, we calculate recall, defined as

$$\text{recall} = \frac{\text{true positives}}{\text{total ground-truth positives}},$$

which captures the proportion of correctly inferred edges relative to the true causal edges. The F1-score, given by

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}},$$

harmonizes precision and recall to provide a balanced measure of structural recovery.

**Practical Considerations**

In practice, the indices of latent subprocesses in the estimated (summary) graph may not correspond to those in the ground truth. To address this, following Huang et al. [21], we permute the latent subprocess indices in the estimated graph and select the permutation that minimizes the difference from the true graph. When the number of estimated latent subprocesses is smaller than the true number, we add isolated latent nodes to balance the comparison. Conversely, if the estimate exceeds the true number, we select the subset that best matches the true latent subprocesses.

Additionally, since our inferred summary graph simplifies the underlying causal structure, by omitting intermediate latent subprocesses and redundant edges as formalized in our theorems and Definition 3.4, we adjust the ground-truth adjacency matrix to this idealized representation before comparison. This ensures a fair evaluation of causal discovery.

For baselines designed for i.i.d. data with linear relations (i.e., Hier. Rank [21] and RLCD [13]), their output graphs capture relationships among discretized variables, rather than subprocesses. To enable fair comparison, we regard an edge $N_1 \to N_2$ as correctly identified if more than half of the considered variables associated with $N_1$ have inferred edges to those of $N_2$.

### Q.3 Additional Experimental Results

**Comparisons on Cases 5 and 6** Fig. 8 shows the F1-score comparisons for Cases 5 and 6, which correspond to intricate latent confounder structures illustrated in Fig.7c and Fig.7d. These cases involve interactions between latent confounders. The results indicate that our method maintains strong performance even under these challenging causal configurations.

**Sensitivity to Time Discretization Interval** We evaluate the sensitivity of our method to the choice of the discretization interval $\Delta$ with decay parameter $\beta = 1$ in the exponential excitation function $\phi_{ij}(s) = \alpha_{ij} e^{-\beta s}$. As shown in Table 1, when $\Delta$ is set to 0.01 or 0.05, our method achieves consistently high F1-scores across all cases, confirming that the discretized representation sufficiently preserves the temporal dynamics of the underlying Hawkes process. Even at $\Delta = 0.1$, the performance remains stable. However, when $\Delta$ increases to 0.3, we observe a sharp drop in performance, highlighting that overly coarse discretization leads to significant loss of temporal resolution, impairing the estimation of causal structures. The result shows the need to choose a small bin width $\Delta$ relative to the typical support of the excitation function [27, 30].

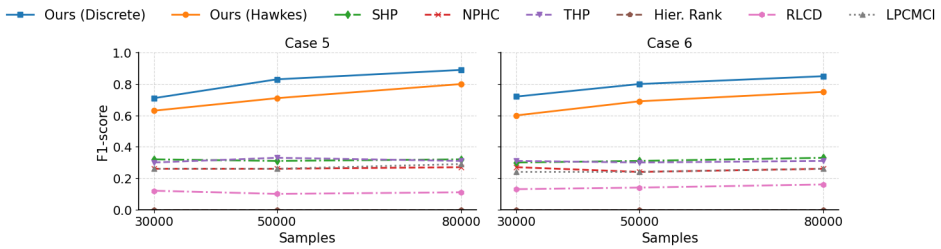



Figure 8: F1-score comparisons on the remaining two causal graphs (Cases 5 and 6), involving latent confounder interactions. Case 5 and Case 6 correspond to the causal structures in Figs. 7c and 7d, respectively.

**Sensitivity to Rank-Test Threshold** We evaluate the sensitivity of our method to the threshold $\tau$ used in the rank test (i.e., the cutoff deciding rank deficiency). We vary $\tau \in \{0.01, 0.05, 0.10, 0.20\}$ and assess three representative cases. Each experiment uses 30k Hawkes samples generated by the `tick` library under an exponential excitation function $\phi_{ij}(s) = \alpha_{ij} e^{-\beta s}$ with $\beta = 1$ and time interval $\Delta = 0.1$; results are averaged over ten runs. As shown in Table 2, in the fully observed setting (Case 1) precision remains 1.00 while recall decreases as $\tau$ increases, whereas in latent settings (Cases 2–3) a moderately larger threshold improves precision because of the attenuation of causal influences through the latent subprocesses. Overall, a threshold of 0.10 provides a good balance across different scenarios.

**Robustness to Violations of Rank Faithfulness** To test robustness under violations of rank faithfulness, we randomly select two edges in each synthetic graph and assign them identical coefficients $\alpha_{ij}$ for the exponential excitation function $\phi_{ij}(s) = \alpha_{ij} e^{-\beta s}$ in every run. This manipulation introduces potential linear dependencies in the cross-covariance matrix, which could challenge rank-based methods. As presented in Table 3, despite the induced degeneracy, our method maintains strong performance, especially as the sample size increases. These results suggest that our approach is robust to moderate violations of rank faithfulness in practical scenarios.

Table 1: Performance of our method under varying $\Delta$ values using 80k Hawkes process samples generated by the `tick` library with decay parameter $\beta = 1$ in the exponential excitation function. Case 1–3 correspond to Figs. 1b, 2a, and 7a, respectively. Results are averaged over ten runs. Performance remains stable and high when $\Delta \leq 0.1$, but degrades significantly at $\Delta = 0.3$ due to the loss of fine-grained temporal information.

| | **Precision** | | | **Recall** | | | **F1-Score** | | |
|---|---|---|---|---|---|---|---|---|---|
| $\Delta$ | Case 1 | Case 2 | Case 3 | Case 1 | Case 2 | Case 3 | Case 1 | Case 2 | Case 3 |
| 0.01 | 0.98 | 0.91 | 0.84 | 0.92 | 0.93 | 0.83 | 0.93 | 0.92 | 0.84 |
| 0.05 | 1.00 | 0.96 | 0.83 | 0.84 | 0.98 | 0.82 | 0.90 | 0.97 | 0.82 |
| 0.10 | 1.00 | 0.91 | 0.86 | 0.87 | 0.93 | 0.83 | 0.93 | 0.92 | 0.84 |
| 0.30 | 0.50 | 0.55 | 0.50 | 0.17 | 0.63 | 0.33 | 0.25 | 0.59 | 0.40 |

**Evaluation on a larger and more complex causal graph** We further evaluate our method on a larger causal graph with 14 subprocesses, as shown in Fig. 9. Table 4 reports the F1-scores averaged over ten runs. Despite the increased complexity, our method successfully recovers the underlying causal structure with high accuracy.

**Scalability and Runtime Profiling** We profile runtime on three representative synthetic graphs (Cases 1–3) and two real-world settings. All runs were executed on an AMD EPYC 9454 CPU. The first real-world setting follows our main paper: a five-alarm subgraph (`Alarm_ids=0-3` with one latent `Alarm_id=7`) from *device_id = 8*. The second merges all devices into a single multivariate event sequence with all 18 alarms to gauge scaling with graph size. We observe that Case 1 is fastest as no latent confounders are present and Phase I suffices. Case 2 introduces latent confounders, requiring

Table 2: Sensitivity to the rank-test threshold $\tau$. Each entry is averaged over ten runs on 30k samples generated with an exponential kernel ($\beta = 1$). Case 1–3 correspond to Figs. 1b, 2a, and 7a, respectively. Overall, a threshold of 0.10 provides a good balance across different scenarios.

| | **Precision** | | | **Recall** | | | $F_1$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **Threshold** $\tau$ | Case 1 | Case 2 | Case 3 | Case 1 | Case 2 | Case 3 | Case 1 | Case 2 | Case 3 |
| 0.01 | 1.00 | 0.42 | 0.57 | 0.80 | 0.53 | 0.50 | 0.88 | 0.47 | 0.53 |
| 0.05 | 1.00 | 0.62 | 0.62 | 0.64 | 0.73 | 0.54 | 0.77 | 0.67 | 0.57 |
| 0.10 | 1.00 | 0.66 | 0.72 | 0.60 | 0.75 | 0.65 | 0.74 | 0.71 | 0.68 |
| 0.20 | 1.00 | 0.76 | 0.68 | 0.47 | 0.85 | 0.63 | 0.62 | 0.80 | 0.65 |

Table 3: Performance of our method when, in each run, two edges in each graph are randomly assigned identical coefficients $\alpha_{ij}$ for the exponential excitation function, increasing the risk of rank deficiency. Hawkes process samples are generated by the `tick` library. Case 1–3 correspond to Figs. 1b, 2a, and 7a, respectively. Results are averaged over ten runs. Despite these perturbations, our method maintains strong performance, demonstrating robustness to such violations.

| | **Precision** | | | **Recall** | | | **F1-Score** | | |
|---|---|---|---|---|---|---|---|---|---|
| #Samples | Case 1 | Case 2 | Case 3 | Case 1 | Case 2 | Case 3 | Case 1 | Case 2 | Case 3 |
| 30k | 0.87 | 0.60 | 0.72 | 0.87 | 0.75 | 0.71 | 0.87 | 0.67 | 0.71 |
| 50k | 0.92 | 0.83 | 0.76 | 0.84 | 0.82 | 0.73 | 0.87 | 0.82 | 0.74 |
| 80k | 0.95 | 0.84 | 0.83 | 0.90 | 0.83 | 0.80 | 0.92 | 0.83 | 0.81 |

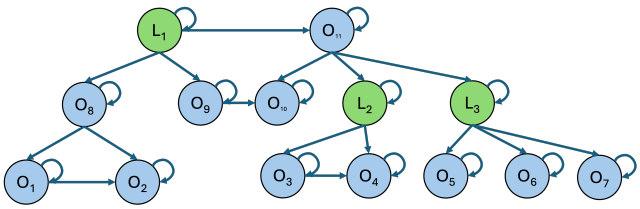



Figure 9: Illustration of a larger causal graph consisting of 14 subprocesses, used to evaluate scalability and robustness.

Table 4: Performance of our method on the larger causal graph in Fig. 9, using Hawkes process data generated by the `tick` library. Results are averaged over ten runs. The method consistently recovers the causal structure with improving accuracy as sample size increases.

| **Sample Size** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|
| 30k | 0.65 | 0.52 | 0.58 |
| 50k | 0.71 | 0.58 | 0.64 |
| 80k | 0.80 | 0.71 | 0.75 |

both phases in the first iteration and increasing runtime. Case 3 is slowest among synthetic cases because the latent confounder is itself caused by an observed subprocess, triggering an additional iteration to identify its observed parent. For real data, merging all devices markedly increases runtime as the sequence spans all 18 alarms and may deviate from a homogeneous Hawkes mechanism. A phase-wise complexity breakdown is provided in Appendix P, which offers further insight into the scalability of the algorithm.

Table 5: Runtime across synthetic and real-world settings.

| **Graph Type** | **Runtime (s)** |
|---|---|
| Case 1 | 227.80 |
| Case 2 | 1036.01 |
| Case 3 | 2603.95 |
| Real Dataset (`Alarm_ids=0-3`, *device_id* = *8*) | 1364.71 |
| Real Dataset (all devices merged; 18 alarms) | 20914.29 |

### Q.4 Analysis of Real-world Dataset Results

We evaluate our method on a real-world cellular network dataset [37], which includes expert-validated ground-truth causal relationships. The dataset comprises 18 distinct alarm types and $\sim$35,000 recorded alarm events collected over eight months from an operational telecommunication network. This benchmark has been widely used in prior work (e.g., the PCIC 2021 causal discovery track and [37]), where performance for many methods is available and top F1-scores are reported up to $\approx 0.6$.

For our evaluation, we focus on a subgraph involving five alarm types (`Alarm_ids=0-3` and 7), where `Alarm_id=7` is manually excluded and treated as a latent subprocess. Both `Alarm_id=1` and `Alarm_id=3` are observed effects of this latent subprocess, providing an opportunity to assess our method's ability to infer latent confounders. The ground-truth causal subgraph is shown in Figure 10. Compared with our inferred causal graph, the ground truth contains an additional edge from `Alarm_id=1` to `Alarm_id=3`. However, as noted in Definition 3.4, causal edges between observed effects of a latent confounder are permissible in our framework.

During inference, using Proposition 3.3 and Theorem F.2, we correctly identify `Alarm_ids=0,1,3` as the parent causes of `Alarm_id=2`, and `Alarm_ids=1,3` as the parent causes of `Alarm_id=0`. The parent cause sets of `Alarm_id=1` and `Alarm_id=3` cannot be fully explained by the observed subprocesses alone. This necessitates the existence of a latent confounder influencing both, leading to the successful identification of `Alarm_id=7` as a latent subprocess.

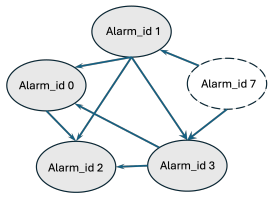



Figure 10: Ground-truth causal subgraph from the metropolitan cellular network dataset. `Alarm_id=7` is treated as a latent subprocess.

**Baselines and protocol.** We compare against representative Hawkes-based methods (SHP [37], THP [6], NPHC [1]), two rank-based latent-variable methods originally for i.i.d. data (Hier. Rank [21], RLCD [13]), and a time-series method for exogenous latents (LPCMCI [16]). Following our rebuttal, LPCMCI is newly included. For fairness, all baselines are run on the same sub-dataset (`Alarm_ids=0-3` and 7 from *device_id* = *8*) used by our method, with each method evaluated over ten runs and averaged.

**Results on the sub-dataset.** Our method achieves the best F1-score when the data conforms to a single multivariate Hawkes process (per-device setting). Table 6 reports the average F1-scores.

Table 6: F1-scores on the cellular network sub-dataset (`Alarm_ids=0-3` and `7`, *device_id = 8*) where `Alarm_id=7` is manually excluded and treated as a latent subprocess; averages over 10 runs.

| **Algorithm** | **F1-score** |
|---|---|
| SHP | 0.49 |
| THP | 0.48 |
| NPHC | 0.42 |
| Hier. Rank | 0.00 |
| RLCD | 0.39 |
| LPCMCI | 0.43 |
| **Ours** | **0.76** |

**Merged-devices analysis.** For completeness, we also merge events from all 55 devices into a single multivariate sequence with all 18 alarm types and analyze it with our method. This setting violates the assumption that samples share the same generative mechanism (devices can be heterogeneous), and it yields a much lower F1-score (0.17). This illustrates why per-device analysis is more compatible with our assumptions, whereas merged-device data can confound structure learning.

**Dataset description.** The dataset records 34,838 alarm events from a metropolitan cellular network [37], covering 18 alarm types and 55 devices. Each record contains:

- **Alarm ID**: one of 18 alarm types,
- **Device ID**: one of 55 devices,
- **Start Timestamp**: time when the alarm was triggered,
- **End Timestamp**: time when the alarm was resolved.

For causal analysis, we sort events by alarm type and use the start timestamp as the event time, yielding a temporally ordered sequence suitable for inference.

# R Limitations

Our method recovers the causal structure of the *discretized* time-series representation of a multivariate Hawkes process; the correspondence to the underlying continuous-time (PO-)MHP holds in the limit as $\Delta \to 0$. When the observational resolution is coarse (large finite $\Delta$), this approximation may not fully capture the continuous-time dynamics. We therefore recommend choosing $\Delta$ small relative to the typical support of the excitation kernel and provide a sensitivity analysis in Table 1.

# S NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions listed in the abstract and introduction correspond to the content of the sections that follow.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explain each assumption in detail and perform sensitivity experiments to violations of the assumptions in Appendix Q.3.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For each theoretical result, we clearly state the corresponding assumptions and attach the corresponding proofs in the appendix. All theorems, formulas, and proofs in the paper are numbered and cross-referenced.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Details of the algorithm are in Appendix O and we explain the experimental setup in detail in Appendix Q.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The dataset is publicly available and we attach a demonstration code in the supplementary material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details of the algorithm are in Appendix O and we explain the experimental setup in detail in Appendix Q.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: In our paper, the source of volatility is mainly synthetic data generated, not the model itself. Once data is fixed, our method become deterministic. We report averaged results over multiple independent trials aiming to reduce the randomness in data generation. Moreover, as our focus is on structural identifiability rather than predictive performance variance, we prioritized reporting the main trends across multiple scenarios.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our experiment can be run on a personal PC (CPU).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: They are discussed in the introduction and conclusion part.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The real-world dataset we used is publicly accessible, and we cite the source regarding the dataset.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.