IEEE Transactions on Information Theory

Learning network-structured dependence from non-stationary multivariate point process data

Journal:	*IEEE Transactions on Information Theory*
Manuscript ID	IT-23-0396.R1
Manuscript Type:	Regular Manuscript
Date Submitted by the Author:	02-Dec-2023
Complete List of Authors:	Gao, Muhong ; Chinese Academy of Sciences, Academy of Mathematics and Systems Science Zhang, Chunming; University of Wisconsin-Madison, Zhou, Jie; Capital Normal University School of Mathematical Sciences
Keywords:	generalized linear model, intensity function, M-estimation, multivariate counting process, network structure
Subject Category:	Machine Learning and Statistics

Learning network-structured dependence from non-stationary multivariate point process data

Muhong Gao *gaomh@amss.ac.cn*
Chunming Zhang *cmzhang@stat.wisc.edu*
Jie Zhou *zhoujie@amss.ac.cn*

December 2, 2023

Abstract

Learning the sparse network-structured dependence among nodes from *multivariate point process* data $\{\boldsymbol{T}_i\}_{i\in\mathcal{V}}$ has wide applications in information transmission, social science, and computational neuroscience. This paper develops new continuous-time stochastic models of the *conditional intensity processes* $\{\lambda_i(t \mid \mathscr{F}_t) : t \geq 0\}_{i\in\mathcal{V}}$ that depend on the past event counts of parent nodes to learn the network structure underlying an array of non-stationary *multivariate counting processes* $\{\boldsymbol{N}(t) : t \geq 0\}$ for $\{\boldsymbol{T}_i\}_{i\in\mathcal{V}}$. The stochastic mechanism of the model is crucial for statistical inference of graph parameters relevant to structure recovery but does not satisfy the key assumptions underlying commonly used processes such as the Poisson process, Cox process, Hawkes process, queuing model, and the piecewise deterministic Markov process. This inspires us to introduce a new *marked point process for intensity discontinuities*, derive the compact representations of their conditional distributions, and demonstrate the cyclicity property of $\boldsymbol{N}(t)$ driven by recurrence time points. These new theoretical properties further enable us to establish statistical consistency and convergence properties of the proposed penalized M-estimators for graph parameters under mild regularity conditions. Simulation evaluations demonstrate the computational simplicity of the proposed method and its increased estimation accuracy compared to existing methods. Real multiple neuron spike train recordings are analyzed to infer connectivity in neuronal networks.

Key words and phrases: consistency; generalized linear model; intensity function; M-estimation; multivariate counting process; network structure;

Short title: Learning network-structured dependence from point process data

1 Introduction

Structured multivariate point process data, ranging from neuron multiple spike trains, file access patterns and failure events in server farms, queuing networks to social networks, has wide applications. Inference of the network structure underlying such multivariate point processes and addressing queries based on the learned structure are important issues. For example, learning the structure of cooperative activity between multiple neurons is an important task in understanding neural spike activity and identifying patterns of information transmission and storage in cortical circuits [33, 19, 6, 21, 7]. Analogously, learning the access patterns of files can be exploited for developing faster file access systems.

Typically, multivariate *point processes* refer to random processes of occurrences of a particular event (such as neuron spike firing) in time and geographical spaces, recorded at V nodes as $\{\boldsymbol{T}_1, \ldots, \boldsymbol{T}_V\}$, where

$$\boldsymbol{T}_i = (T_{i,1},\ \ldots,\ T_{i,\mathrm{N}_i})^\top \quad \text{with } 0 < T_{i,1} < \cdots < T_{i,\mathrm{N}_i} \le \mathrm{T}, \quad \text{for } i \in \mathcal{V}, \tag{1.1}$$

correspond to series of time points $T_{i,\ell}$ of the ℓth event, $\ell = 1, \ldots, \mathrm{N}_i$, arriving at the ith node in an experiment with time length T, where the superscript $\top$ denotes transpose, and $\mathcal{V} = \{1, \ldots, V\}$ is the node set. The corresponding *counting process* $N_i(t) = \sum_{\ell \ge 1} \mathrm{I}(0 \le T_{i,\ell} \le t)$ counts the number of events that occur up to time t at node $i \in \mathcal{V}$, where $\mathrm{I}(\cdot)$ denotes the indicator operator. An important objective is to extract the dependency structure among nodes within the network from V sequences of time series. This dependence network, also recognized as the "local independence graph" [1, 14], visually represents the dependence relationship of historical events from parent nodes on the current events of child nodes. Figure 1 showcases the network-structured dependence (in the left panel) of multivariate point process data at 5 nodes (in the right panel).

Due to the stochastic nature of the point process data $\{T_{i,\ell}\}$ in (1.1) for event occurrences, two types of methods are relevant for modeling multivariate point process data. (a) The discrete-time modeling approach includes the dynamic Bayesian network [12, 28] and variants of generalized linear models (GLM) [6, 40, 44, 46]. This approach partitions

the time axis into equally spaced time bins and transforms the series of event times into a sequence of event bin counts, empirically modelled by Poisson distributions. However, a major drawback is the tradeoff between discrete approximation error and the loss of information. (b) In contrast, the continuous-time approach aims to depict physical processes more accurately but faces substantial challenges in modeling both the time-varying part of *intensity processes* and the sparsity feature underlying the network structure. Several specific continuous-time point process models have been developed, such as the Cox process [26, 30], inhomogeneous Poisson process [35], the linear Hawkes process [9, 16, 42], and the non-linear Hawkes process [20, 39]. Other recent works analyzing point process data include [13, 8, 34, 43], with [43] focusing on spatiotemporal data (e.g., crime data) and [13, 8, 34] focusing on interaction data (e.g., E-mail/text messages). In particular, [8, 34] focus on identifying uniform effects (e.g., homophily, dyadic, and triadic effects) in a predetermined network.

To capture the unknown dependency structure between point process data represented in (1.1) both qualitatively and quantitatively, we aim to develop new network structure learning methods that integrate the utility of continuous-time and discrete-time modeling. Specifically, we build new continuous-time GLM-type stochastic models (3.1) for the *conditional intensity processes* $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$, where each $\lambda_i(t \mid \mathscr{F}_t)$ depends on short-term past events of all other nodes up to time t (in contrast to [26, 30], where the conditional intensity is a separate stochastic process independent of the past events) and incorporates the magnitude and direction of interaction effects in graph parameters. By employing penalized M-estimation of parameters in the graph structure (as in (5.8)), we obtain a sparse network. In contrast, the consideration of sparsity was not present in [13, 8, 34, 43]. Our method captures both excitatory and inhibitory effects between nodes, distinguishing itself from the linear Hawkes process [9, 16], specifically tailored for modeling excitatory effects to ensure a non-negative intensity function. Furthermore, our method does not require partitioning the data into bins, making it partition-free and avoiding the subjective choice of bin width associated with the discrete-time approach.

Addressing the theoretical challenges arising from statistical learning procedures in continuous-time stochastic modeling remains a central issue. To the best of our knowledge, there are limited theoretical studies at the intersection of continuous-time point processes and network-structured learning methods. Traditional tools for establishing stochastic convergence and statistical consistency are not directly applicable in the context of statistical

estimation from point process data. This is because the loss function (e.g., in (5.4)) for parameter estimation primarily relies on the non-standard dependence structure of counting processes $\{N_i(t)\}_{i\in\mathcal{V}}$ associated with the point process data $\{T_{i,\ell}\}$. While works for the non-linear Hawkes process [20, 10], queuing models [23, 27], and the piecewise deterministic Markov process [2] provide insights, they rely on specific assumptions and properties that do not hold for our model (3.1). Refer to Sections 4.1.3 and 4.2 for more detailed discussions.

This paper aims to contribute to several aspects that are central to statistical inference for a wide array of non-stationary multivariate point process data encountered in various applications.

(i) We introduce a new tool called the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I}) = (\{\breve{T}_\ell\}_{\ell\geq 1}, \{I_\ell\}_{\ell\geq 1})$ for capturing intensity discontinuities (see Section 4.1). This tool involves compiling all the discontinuity points of the intensity processes $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ into a single sequence $\{\breve{T}_\ell\}_{\ell\geq 1}$, where each $\breve{T}_\ell$ is accompanied by a unique categorical mark $I_\ell \in \mathcal{V}\cup\{0\}$. We derive the probability distributions of $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ and establish a series of probabilistic properties. These results for $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ also provide valuable insights into the probabilistic properties of the original counting processes $\{N_i(t)\}_{i\in\mathcal{V}}$ and enable the development of a new simulation algorithm for generating synthetic data.

(ii) We establish the cyclicity property of $\{N_i(t)\}_{i\in\mathcal{V}}$ driven by a sequence of recurrence time points $R_1 < R_2 < \cdots$ (see Section 4.2). This property demonstrates that our counting processes $\{N_i(t)\}_{i\in\mathcal{V}}$, upon reaching each recurrence time point $t = R_\ell$, initiate a renewed cyclic procedure independent of the event history, as illustrated in Figure 3 of Appendix A. Building on this property, we further derive the asymptotic mean stationarity of $\{N_i(t)\}_{i\in\mathcal{V}}$.

(iii) All these probabilistic results are essential for deriving the statistical properties, such as the consistency of the proposed penalized M-estimation in structure learning, in Section 5.

The validity of our proposed penalization method for inferring network-structured dependencies is supported by extensive simulation studies, and its practical utility in the analysis of real-world multivariate point process data is illustrated with a prefrontal cortex spike train dataset.

The rest of the paper is arranged as follows. Section 2 reviews the multivariate point process and outlines the proposed continuous-time modeling framework. Section 3 presents our proposed model for $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$, and Section 4 investigates related probabilistic properties of $\{N_i(t)\}_{i\in\mathcal{V}}$. Section 5 addresses statistical properties related to the proposed

network recovery procedure. Section 6 illustrates simulation evaluations of the proposed method, and Section 7 analyzes real spike train data. Section 8 briefly discusses and concludes the paper. Appendices A and B collect all figures and tables, technical details, and derivations.

2 Multivariate point process in our setup

We start with a brief review of the point process. For a more comprehensive discussion, refer to [11]. Denote by $\mathcal{V} = \{1, \ldots, V\}$ the set of nodes. Throughout the paper, we focus on the setting where the number V of nodes is a fixed constant. For each node $i \in \mathcal{V}$, we define the univariate *point process* as $\{T_{i,\ell}\}_{\ell \geq 1}$ on the probability space $(\Omega, \mathscr{F}, \mathrm{P})$, where the time-ordered sequence of event time points at node i is denoted as

$$0 < T_{i,1} < T_{i,2} < \cdots . \tag{2.1}$$

For $t \geq 0$, we use $N_i(t)$ to represent the event counts in the time interval $[0, t]$:

$$N_i(t) = \sum_{\ell \geq 1} \mathrm{I}(0 \leq T_{i,\ell} \leq t). \tag{2.2}$$

The term $\{N_i(t)\}_{t \geq 0}$ refers to the *counting process* of $\{T_{i,\ell}\}_{\ell \geq 1}$. More generally, we denote the event counts in any Borel set $\mathcal{T} \in \mathcal{B}(\mathbb{R})$ as:

$$N_i(\mathcal{T}) = \sum_{\ell \geq 1} \mathrm{I}(T_{i,\ell} \in \mathcal{T}), \tag{2.3}$$

which, for $\mathcal{T} = [0, t]$, reduces to $N_i(t)$ as defined in (2.2).

According to (2.2), a point process $\{T_{i,\ell}\}_{\ell \geq 1}$ uniquely defines a counting process $\{N_i(t)\}_{t \geq 0}$. Conversely, $\{N_i(t)\}_{t \geq 0}$ uniquely yields a point process, due to the identity $T_{i,\ell} = \inf\{t > T_{i,\ell-1} : N_i(t) > N_i(T_{i,\ell-1})\}$. Thus, the counting process $\{N_i(t)\}_{t \geq 0}$ and the point process $\{T_{i,\ell}\}_{\ell \geq 1}$ are equivalent to each other.

For the multivariate setting with V nodes, we define the vector $\boldsymbol{N}(t) = (N_1(t), \ldots, N_V(t))^\top$, and call $\{\boldsymbol{N}(t)\}_{t \geq 0}$ the *multivariate counting process*, corresponding to the *multivariate point process* $\{T_{i,\ell} : \ell \geq 1\}_{i \in \mathcal{V}}$. For each $t \geq 0$, let $\mathscr{F}_t \subseteq \mathscr{F}$ be the smallest sub σ-algebra that contains all the information of the multivariate counting process in the history up to time t, formally defined as

$$\mathscr{F}_t = \sigma\big(\{N_i(s) : s \in [0, t],\ i \in \mathcal{V}\}\big). \tag{2.4}$$

From (2.4), it is seen that

$$\mathscr{F}_{t_1} \subseteq \mathscr{F}_{t_2} \subseteq \cdots, \quad \text{for any } 0 \leq t_1 \leq t_2 \leq \cdots. \tag{2.5}$$

We refer to the sequence of σ-algebras $\{\mathscr{F}_t\}_{t\geq 0}$ in (2.4), satisfying the property (2.5), as the *filtration generated by* $\{\boldsymbol{N}(t)\}_{t\geq 0}$, and call $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\geq 0}, \mathrm{P})$ the corresponding *filtered probability space*.

2.1 Total intensity process of $\boldsymbol{N}(t)$

For a single node i, the stochastic character of a counting process $N_i(t)$ is captured by the corresponding *intensity process* (also called *conditional intensity function*) $\lambda_i(t \mid \mathscr{F}_t)$, which measures the instantaneous rate of event occurrence at node i. In this paper, we adopt the definition of the intensity process from [38]:

$$\begin{aligned} \lambda_i(t \mid \mathscr{F}_t) &= \lim_{\Delta \downarrow 0} \Delta^{-1} \mathrm{P}\big(N_i(t+\Delta) = N_i(t) + 1 \mid \mathscr{F}_t\big) && (2.6) \\ &= \lim_{\Delta \downarrow 0} \Delta^{-1} \mathrm{P}\big(N_i(t+\Delta) \neq N_i(t) \mid \mathscr{F}_t\big), \quad \text{almost surely (a.s.)} && (2.7) \end{aligned}$$

for $i \in \mathcal{V}$ and $t \geq 0$.

For the multivariate case with V nodes, we similarly define the *total intensity process* of $\boldsymbol{N}(t)$ as:

$$\begin{aligned} \lambda^{\mathrm{sum}}(t \mid \mathscr{F}_t) &= \lim_{\Delta \downarrow 0} \Delta^{-1} \mathrm{P}\big(\cup_{i\in\mathcal{V}} \{N_i(t+\Delta) = N_i(t) + 1\} \mid \mathscr{F}_t\big) && (2.8) \\ &= \lim_{\Delta \downarrow 0} \Delta^{-1} \mathrm{P}\big(\boldsymbol{N}(t+\Delta) \neq \boldsymbol{N}(t) \mid \mathscr{F}_t\big), \quad \text{a.s.}, && (2.9) \end{aligned}$$

where, for $s \neq t$, the event $\{\boldsymbol{N}(s) \neq \boldsymbol{N}(t)\}$ denotes $\cup_{i\in\mathcal{V}}\{N_i(s) \neq N_i(t)\}$.

Remark 1 *Our definition of the intensity process in* (2.6) *and* (2.7)*, following* [38]*, assumes that* $\lim_{\Delta\downarrow 0} \Delta^{-1}\mathrm{P}(N_i(t+\Delta) = N_i(t)+1 \mid \mathscr{F}_t) = \lim_{\Delta\downarrow 0} \Delta^{-1}\mathrm{P}(N_i(t+\Delta) \neq N_i(t) \mid \mathscr{F}_t)$ *holds* a.s. *for every* $i \in \mathcal{V}$ *and* $t \geq 0$*. This assumption essentially means that simultaneous events from a single node are not allowed in our point process. As shown in Appendix* B*, any multivariate counting process with identical limits* (2.6) *and* (2.7) *also has identical limits* (2.8) *and* (2.9) a.s. *for every* $t \geq 0$*.*

2.2 Orthogonality of martingales of $\boldsymbol{N}(t)$

For the multivariate setting in our study, the structure of the counting process $\boldsymbol{N}(t)$ cannot be fully described by solely presenting the intensity processes $\lambda_i(t \mid \mathscr{F}_t)$ at individual nodes

i. Additionally, it is necessary to clarify how the increments of event counts, $N_i(t+\Delta)-N_i(t)$ and $N_j(t+\Delta)-N_j(t)$, are correlated between any pair of distinct nodes i and j. For $\boldsymbol{N}(t)$ in Definition 1 below, we introduce the notion of orthogonality of martingales (OM) which refers to the case where $N_i(t+\Delta)-N_i(t)$ and $N_j(t+\Delta)-N_j(t)$, conditional on $\mathscr{F}_t$, are asymptotically independent for all $i \neq j$.

Definition 1 (Orthogonality of martingales (OM)) *A multivariate counting process $\boldsymbol{N}(t)$ satisfies the* OM *condition if, for any two distinct nodes $i, j \in \mathcal{V}$ and any time $t \geq 0$:*

$$\begin{aligned} & \lim_{\Delta \downarrow 0} \Delta^{-2} \mathrm{P}\big(N_i(t+\Delta) = N_i(t)+1, N_j(t+\Delta) = N_j(t)+1 \mid \mathscr{F}_t\big) \\ = \; & \lim_{\Delta \downarrow 0} \Delta^{-2} \mathrm{P}\big(N_i(t+\Delta) = N_i(t)+1 \mid \mathscr{F}_t\big) \cdot \mathrm{P}\big(N_j(t+\Delta) = N_j(t)+1 \mid \mathscr{F}_t\big) \\ = \; & \lambda_i(t \mid \mathscr{F}_t)\lambda_j(t \mid \mathscr{F}_t), \quad \text{a.s.}. \end{aligned} \tag{2.10}$$

Lemma 1 shows that for a multivariate counting process $\boldsymbol{N}(t)$ that satisfies the OM condition, the total intensity process $\lambda^{\mathrm{sum}}(t \mid \mathscr{F}_t)$ in (2.8) and (2.9) equals the sum of all intensity processes $\lambda_i(t \mid \mathscr{F}_t)$ over individual nodes $i \in \mathcal{V}$. In the remainder of the paper, we consistently assume the OM condition for the multivariate counting process $\boldsymbol{N}(t)$.

Lemma 1 (Total intensity of the multivariate counting process $\boldsymbol{N}(t)$) *Assume conditions* A1 *and* A2 *in Appendix* B. *Assume that* $\mathrm{P}(\lambda_i(t \mid \mathscr{F}_t) < \infty) = 1$ *for all $i \in \mathcal{V}$ and $t \geq 0$. If $\boldsymbol{N}(t)$ satisfies the* OM *condition, then for any $t \geq 0$, the total intensity process $\lambda^{\mathrm{sum}}(t \mid \mathscr{F}_t)$ defined in* (2.8) *and* (2.9) *satisfies*

$$\lambda^{\mathrm{sum}}(t \mid \mathscr{F}_t) = \sum_{i=1}^{V} \lambda_i(t \mid \mathscr{F}_t), \quad \text{a.s.}. \tag{2.11}$$

3 Statistical model for $\lambda_i(t \mid \mathscr{F}_t)$ with network structure

We propose a continuous-time GLM-type modeling for $\lambda_i(t \mid \mathscr{F}_t)$:

$$\lambda_i(t \mid \mathscr{F}_t) \;=\; \exp\Big\{\beta_{0;i} + \sum_{j \in \mathcal{V}} \beta_{j,i}\, x_j(t)\Big\}, \quad i \in \mathcal{V}, \quad t \geq 0. \tag{3.1}$$

The parameters $\beta_{0;i}$ and $\beta_{j,i}$, along with the covariates $x_j(t)$, have the following interpretations:

Baseline intensity parameter $\beta_{0;i}$. Since the background intensity may vary over nodes, we include a bias term $\beta_{0;i}$ in (3.1) to associate the baseline intensity parameter with each node i.

Connection strength parameter $\beta_{j,i}$. The connection strength parameter $\beta_{j,i}$ in (3.1) quantifies the magnitude and direction of a parent node j's influence on the child node i, denoted as $(\boldsymbol{j}) \xrightarrow{\beta_{j,i}} (\boldsymbol{i})$. Specifically:

$$\begin{array}{lll} \beta_{j,i} > 0: & \text{Excitatory} & \text{effect from node } j \text{ to node } i; \\ \beta_{j,i} = 0: & \text{No} & \text{effect from node } j \text{ to node } i; \\ \beta_{j,i} < 0: & \text{Inhibitory} & \text{effect from node } j \text{ to node } i. \end{array}$$

For interpretability, we assume $\beta_{i,i} = 0$ for all $i \in \mathcal{V}$, meaning there is no self-effect. The network graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ can be obtained from all pairs of nodes (j, i) with non-zero connection parameters $\beta_{j,i}$ in the edge set:

$$\mathcal{E} \;=\; \{(j,i) \in \mathcal{V} \times \mathcal{V} : \beta_{j,i} \neq 0,\, j \neq i\} = \mathcal{E}_+ \cup \mathcal{E}_-. \tag{3.2}$$

This distinguishes the edge set for excitatory effects:

$$\mathcal{E}_+ \;=\; \{(j,i) \in \mathcal{V} \times \mathcal{V} : \beta_{j,i} > 0,\, j \neq i\}, \tag{3.3}$$

from the edge set for inhibitory effects:

$$\mathcal{E}_- \;=\; \{(j,i) \in \mathcal{V} \times \mathcal{V} : \beta_{j,i} < 0,\, j \neq i\}. \tag{3.4}$$

The configuration of this graph $\mathcal{G}$ reveals the interaction effects between nodes, and learning such a graph structure through statistical estimation methods is the main goal of this paper.

Regression covariates $x_j(t)$. Regression covariates $x_j(t)$ aim to represent the effect from other nodes $j \in \mathcal{V}$ on node i within a short period of time until t. We formulate $x_j(t)$ as follows:

$$x_j(t) \;=\; g\big(r_{j,\phi}(t)\big), \tag{3.5}$$

where $r_{j,\phi}(t)$ is the empirical rate during a short time interval of width $\phi \in (0, \infty)$:

$$r_{j,\phi}(t) = N_j\big(\big(t-\phi, t\big]\big)/\phi. \tag{3.6}$$

Here, $g(\cdot) : [0,\infty) \to [0,\infty)$ is a non-linear shape function that is continuous, non-negative, and monotonically increasing, with $g(0) = 0$. It is worth noting that within the modelling framework (3.1) for $\lambda_i(t \mid \mathscr{F}_t)$, the function g is not restricted to be bounded. Condition A5 assumes a bounded $g(\cdot)$ to facilitate the analytical derivation of theoretical results, such as probabilistic properties of $\boldsymbol{N}(t)$ and asymptotic properties of parameter estimators. However, this assumption may be relaxed in certain cases. For practical choices of the shape function g and the time-lag constant ϕ, refer to Appendix A.1. Additionally, for empirical performances in data analysis, parameter estimation, and structure learning, see Sections 6–7.

3.1 Connection of model (3.1) with other models

The proposed model (3.1) employs the GLM-type framework to link the intensity process $\lambda_i(t \mid \mathscr{F}_t)$ with both historical data and the network structure. This provides a novel continuous-time approach for modeling multivariate point process data. Note that the exponential link function in our model (3.1) is convex and twice-differentiable, aiding theoretical analysis and computational efficiency, distinguishing it from other non-linear link functions such as ReLU or sigmoid functions. As shown below, by selecting two specific choices of the shape function g in (3.5) (combined with (3.6)), model (3.1) establishes connections with two existing models.

Example 1: $g(x) = x$. Then model (3.1) becomes:

$$\begin{aligned} \lambda_i(t \mid \mathscr{F}_t) &= \exp\Big\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \beta_{j,i}\, r_{j,\phi}(t)\Big\} \\ &= \exp\Big\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \int_{-\infty}^{t} \frac{1}{\phi}\, \beta_{j,i}\, \mathrm{I}(0 \le t-u < \phi)\, \mathrm{d}N_j(u)\Big\}, \end{aligned} \tag{3.7}$$

which is a special case of the general multivariate non-linear Hawkes process [5]:

$$\lambda_i(t \mid \mathscr{F}_t) = \varphi\Big(\beta_{0;i} + \sum_{j\in\mathcal{V}} \int_{-\infty}^{t} \omega_{j,i}(t-u)\, \mathrm{d}N_j(u)\Big), \tag{3.8}$$

when we set the non-linear link function $\varphi(\cdot) = \exp(\cdot)$, the interaction function $\omega_{j,i}(u) = \beta_{j,i}\, \mathrm{I}(0 \le u < \phi)/\phi$, and assume $\beta_{i,i} = 0$.

Example 2: $g(x) = \log(1+x)$. Then model (3.1) becomes:

$$\begin{aligned} \lambda_i(t \mid \mathscr{F}_t) &= \exp\Big(\beta_{0;i} + \sum_{j\in\mathcal{V}} \beta_{j,i}\, \log\{1 + r_{j,\phi}(t)\}\Big) \\ &= \exp(\beta_{0;i}) \prod_{j\in\mathcal{V}} \{1 + r_{j,\phi}(t)\}^{\beta_{j,i}}, \end{aligned} \tag{3.9}$$

which agrees with [35]. In comparison to Example 1, the shape-function $g(x) = \log(1+x)$ in Example 2 is relatively flat. This moderates the steepness of the exponential link function and down-weights the influence of excessively large intensities. Therefore, (3.9) is expected to better represent the dynamics of multivariate point process data in real applications.

Distinction from Markov processes: A general stochastic process is Markovian if, conditional on the past and present states, the probability of transitioning to a future state depends solely on the present state, but not on the past history ([37], p. 132). In our case,

the counting process $\{\boldsymbol{N}(t)\}_{t\geq 0}$ associated with the intensity processes $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ in model (3.1) (together with (3.5) and (3.6)), is *not* Markovian. This is because the intensities $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ depend not only on the current state of $\boldsymbol{N}(t)$ but also on the past states of $\boldsymbol{N}((t-\phi, t))$. This distinction highlights the clear difference between our model and other Markovian models of stochastic processes, such as the Markov multi-state model [3] commonly used in survival analysis, the versatile Markovian point process [29] used for modeling queuing systems, or the piecewise deterministic Markov process [2] used for modeling physical processes of particle motions.

4 Properties of the proposed intensity model

In this section, we investigate the probabilistic properties of the counting process $\boldsymbol{N}(t)$ associated with the intensity processes $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ in model (3.1). These results are essentially required for deriving our statistical properties (Theorems 5–7 and Corollary 1 in Section 5).

A distinctive feature of our intensity processes $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ in (3.1) is that they are piecewise-constant functions of time t (as to be shown in Section 4.1.1). In other words, unlike many other models, our $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ do not change continuously over time, yielding a countable number of discontinuity points in $(0, \infty)$ from all nodes. The discontinuity points of $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ play an important role in characterizing the stochastic features of our intensity processes. We begin by investigating the set of discontinuity points of $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ in Section 4.1.

4.1 Marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for intensity discontinuities

In this section, we conduct a step-by-step analysis based on the discontinuity points of $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ in (3.1). Section 4.1.1 demonstrates the piecewise-constant nature of $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$. Section 4.1.2 defines the "marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for intensity discontinuities," which proves to be equivalent (as shown in (4.7) and (4.8)) for studying the point process $\{T_{i,\ell}\}_{\ell\geq 1, i\in\mathcal{V}}$ and the counting process $\boldsymbol{N}(t)$. Section 4.1.3 derives the probability distribution of $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ (in Theorem 1) and presents related properties (in Lemmas 5–7). By translating the results of $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ into the analogues of $\boldsymbol{N}(t)$, Section 4.1.4 demonstrates the bounded variance and finiteness properties (in Theorem 2) for our counting process $\boldsymbol{N}(t)$.

4.1.1 Piecewise-constant $\lambda_i(t \mid \mathscr{F}_t)$

Recall that from (3.1), (3.5), and (3.6), the conditional intensity function at each node $i \in \mathcal{V}$ can be rewritten as $\lambda_i(t \mid \mathscr{F}_t) = \exp\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \beta_{j,i}\, g(N_j((t-\phi,t])/\phi)\}$, which is a continuous function of $\{N_j((t-\phi,t])\}_{j\in\mathcal{V}}$, where

$$N_j\big((t-\phi,t]\big) \;=\; N_j(t) - N_j(t-\phi), \quad t \geq 0. \tag{4.1}$$

Thus the smoothness of $\lambda_i(t \mid \mathscr{F}_t)$ directly depends on that of $\{N_j((t-\phi,t])\}_{j\in\mathcal{V}}$.

For each node $j \in \mathcal{V}$ and event time points $\{T_{j,\ell}\}_{\ell\geq 1}$ in (2.1), we define $N_j(\{t\}) = \sum_{\ell\geq 1} \mathrm{I}(T_{j,\ell} = t)$, which represents the jump size of $N_j(\cdot)$ at a single point t. It is clear that $N_j(\{t\}) \in \{0,\, 1\}$, and $N_j(\{t\}) = 1$ is equivalent to $t \in \{T_{j,\ell}\}_{\ell\geq 1}$. Moreover, two properties of $N_j((t-\phi,t])$ can be verified. First, $N_j((t-\phi,t])$ is non-negative, right-continuous, piecewise-constant, but not monotonically increasing in $t \in [0,\infty)$. Accordingly, $\lambda_i(t \mid \mathscr{F}_t)$ is also right-continuous and piecewise-constant. Second, the set of discontinuity points of $N_j((t-\phi,t])$ is given by

$$\big\{t \geq 0 : N_j(\{t\}) - N_j(\{t-\phi\}) = +1\big\} \cup \big\{t \geq 0 : N_j(\{t\}) - N_j(\{t-\phi\}) = -1\big\}, \tag{4.2}$$

where

$$N_j(\{t\}) - N_j(\{t-\phi\}) = \begin{cases} 0, & \text{if } t \notin \{T_{j,\ell}\}_{\ell\geq 1}, \text{and } t \notin \{T_{j,k}+\phi\}_{k\geq 1}, \\ +1, & \text{if } t \in \{T_{j,\ell}\}_{\ell\geq 1}, \text{and } t \notin \{T_{j,k}+\phi\}_{k\geq 1}, \\ -1, & \text{if } t \notin \{T_{j,\ell}\}_{\ell\geq 1}, \text{and } t \in \{T_{j,k}+\phi\}_{k\geq 1}, \\ 0, & \text{if } t \in \{T_{j,\ell}\}_{\ell\geq 1}, \text{and } t \in \{T_{j,k}+\phi\}_{k\geq 1}. \end{cases} \tag{4.3}$$

Following (4.3), we can rewrite the set of discontinuity points in (4.2) as

$$\big\{t \geq 0 : t \in \{T_{j,\ell}\}_{\ell\geq 1} \text{and}\, t \notin \{T_{j,k}+\phi\}_{k\geq 1}\big\} \cup \big\{t \geq 0 : t \notin \{T_{j,\ell}\}_{\ell\geq 1} \text{and}\, t \in \{T_{j,k}+\phi\}_{k\geq 1}\big\},$$

which belongs to the set

$$\{T_{j,\ell}\}_{\ell\geq 1} \,\cup\, \{T_{j,k}+\phi\}_{k\geq 1}.$$

Utilizing [15] (Theorem 2.4.7, p. 84) and the intensity function $\lambda_j(t \mid \mathscr{F}_t) < \infty$ in (3.1), the event time points $\{T_{j,\ell}\}_{\ell\geq 1}$ are *totally inaccessible stopping times*, implying that $\mathrm{P}(\cup_{\ell\geq 1}\cup_{k\geq 1} \{T_{j,\ell} = T_{j,k}+\phi\}) = 0$. Thus, the right-continuous $\lambda_i(t \mid \mathscr{F}_t)$ is piecewise-constant in $t \in [0,\infty)$, with the set of discontinuity points specified in Lemma 2.

Lemma 2 (Piecewise-constant $\lambda_i(t \mid \mathscr{F}_t)$ and its discontinuity points) *Assume conditions* A1 *and* A2 *in Appendix* B. *For each $i \in \mathcal{V}$, let $\mathrm{Pa}(i) = \{j \in \mathcal{V} \setminus i : \beta_{j,i} \neq 0\}$ denote the set of parent nodes for node i. If $\mathrm{Pa}(i) \neq \varnothing$, then $\lambda_i(t \mid \mathscr{F}_t)$ is a piecewise-constant function of $t \in [0, \infty)$, with all its discontinuity points listed in the set*

$$\cup_{j \in \mathrm{Pa}(i)} \big\{\{T_{j,\ell}\}_{\ell \geq 1} \cup \{T_{j,k} + \phi\}_{k \geq 1}\big\}.$$

If $\mathrm{Pa}(i) = \varnothing$, then $\lambda_i(t \mid \mathscr{F}_t) \equiv \exp(\beta_{0;i})$ is a constant, and $\{T_{i,\ell}\}_{\ell \geq 1}$ reduces to a homogeneous Poisson process.

An illustration of $N_j(t), N_j((t-\phi, t])$ and $\lambda_i(t \mid \mathscr{F}_t)$ is given in Figure 2. By aggregating the discontinuity points of all intensity functions $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i \in \mathcal{V}}$, we directly obtain the following Lemma 3.

Lemma 3 (Discontinuity points of all intensities $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i \in \mathcal{V}}$) *Assume conditions* A1 *and* A2 *in Appendix* B. *Then we have the following results:*

(i) *The discontinuity points of all intensity functions $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i \in \mathcal{V}}$ are listed in the set*

$$\mathcal{T}^* = \cup_{i \in \mathcal{V}} \cup_{j \in \mathrm{Pa}(i)} \big\{\{T_{j,\ell}\}_{\ell \geq 1} \cup \{T_{j,k} + \phi\}_{k \geq 1}\big\}. \tag{4.4}$$

(ii) *For each $i \in \mathcal{V}$, let $\mathrm{Ch}(i) = \{j \in \mathcal{V} \setminus i : \beta_{i,j} \neq 0\}$ denote the set of child nodes for node i. Define the sequence of time points*

$$\{\breve{T}_1, \breve{T}_2, \ldots\} = \cup_{j \in \mathcal{V}} \big\{\{T_{j,\ell}\}_{\ell \geq 1} \cup \{T_{j,k} + \phi\}_{k \geq 1}\big\}, \tag{4.5}$$

with $0 < \breve{T}_1 < \breve{T}_2 < \cdots$ arranged in increasing order. Then $\mathcal{T}^*$ is a subset of $\{\breve{T}_\ell\}_{\ell \geq 1}$. Moreover, if $\mathrm{Ch}(i) \neq \varnothing$ for all $i \in \mathcal{V}$, then we have $\mathcal{T}^* = \{\breve{T}_\ell\}_{\ell \geq 1}$.

Remark 2 *Lemmas* 2 *and* 3 *demonstrate the close relationship between the fundamental characteristics of intensity processes $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i \in \mathcal{V}}$ in model* (3.1) *and the properties of the network structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in* (3.2). *For instance, if a node i has a parent node in the network $\mathcal{G}$, then $\lambda_i(t \mid \mathscr{F}_t)$ is non-constant and the point process on node i is not reduced to the trivial case of a homogeneous Poisson process. Additionally, if each node in the network $\mathcal{G}$ has at least one child node, then the set of all intensity discontinuities $\mathcal{T}^*$ in* (4.4) *is identical to the set of time points $\{\breve{T}_\ell\}_{\ell \geq 1}$ in* (4.5). *Since $\{\breve{T}_\ell\}_{\ell \geq 1}$ contains all the discontinuity points in $\mathcal{T}^*$ and has a simpler form than $\mathcal{T}^*$, we will focus our remaining analysis on $\{\breve{T}_\ell\}_{\ell \geq 1}$ and refer to it as "the set of intensity discontinuities" with a slight abuse of terminology.*

4.1.2 Marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for studying discontinuity points of $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$

To investigate $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$, we next introduce the notion of "marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I}) = (\{\breve{T}_\ell\}_{\ell\geq 1}, \{I_\ell\}_{\ell\geq 1})$ for intensity discontinuities" in Definition 2 below. A general marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ is a double sequence, where $\{\breve{T}_\ell\}_{\ell\geq 1}$ is a point process, and each $\breve{T}_\ell$ is associated with a *mark* I_ℓ, usually representing some additional features (such as labels or locations) related to the time point $\breve{T}_\ell$; refer to [11] and the references therein for further details.

Definition 2 (Marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for discontinuity points of $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$) *Assume conditions* A1 *and* A2 *in Appendix* B. *For the strictly increasing time points* $\{\breve{T}_1, \breve{T}_2, \ldots\}$ *defined in* (4.5) *and integers* $\ell \geq 1$, *let* $I_\ell \in \mathcal{V} \cup \{0\}$ *be the mark corresponding to* $\breve{T}_\ell$, *defined by:*

$$I_\ell = \begin{cases} i, & \text{if } \breve{T}_\ell \in \{T_{i,k}\}_{k\geq 1} \text{ for some node } i \in \mathcal{V}, \\ 0, & \text{if } \breve{T}_\ell \in \{T_{i,k} + \phi\}_{k\geq 1} \text{ for some node } i \in \mathcal{V}. \end{cases} \tag{4.6}$$

We refer to the double sequence $(\breve{\boldsymbol{T}}, \boldsymbol{I}) = (\{\breve{T}_\ell\}_{\ell\geq 1}, \{I_\ell\}_{\ell\geq 1})$ *as the "marked point process for intensity discontinuities".*

The mark I_ℓ in (4.6) indicates the identity of $\breve{T}_\ell$: if the discontinuity point $\breve{T}_\ell$ of $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ is due to an event occurrence from some node i at that time point, then I_ℓ represents the index i of that node; otherwise, we set $I_\ell = 0$. Lemma 4 below guarantees the uniqueness of the mark I_ℓ defined in (4.6) for each $\breve{T}_\ell$.

Lemma 4 (Uniqueness of the mark I_ℓ corresponding to $\breve{T}_\ell$) *Assume conditions* A1 *and* A2 *in Appendix* B. *Then, the mark* I_ℓ *in* (4.6), *corresponding to the discontinuity point* $\breve{T}_\ell$, *is uniquely defined* a.s., *i.e.,*

(i) *For any distinct* $i, j \in \mathcal{V}$, $\mathrm{P}(I_\ell = i,\, I_\ell = j) = \mathrm{P}(\breve{T}_\ell \in \{T_{i,k}\}_{k\geq 1}\,,\, \breve{T}_\ell \in \{T_{j,k}\}_{k\geq 1}) = 0$.

(ii) *For any* $i \in \mathcal{V}$, $\mathrm{P}(I_\ell = i,\, I_\ell = 0) = \mathrm{P}(\breve{T}_\ell \in \{T_{i,k}\}_{k\geq 1}\,,\, \breve{T}_\ell \in \{T_{j,k} + \phi\}_{j\in\mathcal{V},\, k\geq 1}) = 0$.

As stated by Definition 2 and Lemma 4, a multivariate point process $\{T_{i,\ell}\}_{\ell\geq 1,\, i\in\mathcal{V}}$ uniquely defines a marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$. Conversely, $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ uniquely yields a

multivariate point process $\{T_{i,\ell}\}_{\ell\geq 1,\, i\in\mathcal{V}}$, and accordingly, a multivariate counting process $\{\boldsymbol{N}(t)\}_{t\geq 0}$, due to the identities:

$$T_{i,\ell} = \inf\{\breve{T}_k > T_{i,\ell-1} : I_k = i,\, k\geq 1\}, \qquad \ell\geq 1, \quad i\in\mathcal{V}, \tag{4.7}$$

$$N_i(t) = \sum_{k\geq 1} \mathrm{I}(\breve{T}_k \leq t,\, I_k = i), \qquad t\geq 0, \quad i\in\mathcal{V}, \tag{4.8}$$

where $T_{i,0}=0$. Hence, the point process $\{T_{i,\ell}\}_{\ell\geq 1,\, i\in\mathcal{V}}$, the counting process $\{\boldsymbol{N}(t)\}_{t\geq 0}$, and the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ can be deduced from each other. As shown in Theorem 1 below, the probability distribution of the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ has a closed-form expression, making $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ more convenient to analyze than $\{T_{i,\ell}\}_{\ell\geq 1,\, i\in\mathcal{V}}$ and $\{\boldsymbol{N}(t)\}_{t\geq 0}$.

4.1.3 Probabilistic properties of $(\breve{\boldsymbol{T}}, \boldsymbol{I})$

For each integer $\ell\geq 1$, let $\mathcal{F}_{\breve{T}_\ell} = \{A\in\mathscr{F} : A\cap\{\breve{T}_\ell\leq t\}\in\mathscr{F}_t \text{ for every } t>0\}$ be the stopping-time σ-algebra (defined as in [17]) with respect to $\breve{T}_\ell$, i.e., generated by the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ up to time $\breve{T}_\ell$. For $\ell=0$, define $\mathcal{F}_{\breve{T}_0} = \mathscr{F}_0 = \{\Omega, \varnothing\}$, $\breve{T}_0 = 0$, and $I_0 = 0$. Theorem 1 presents the probability distribution of the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ conditional on the filtration $\{\mathcal{F}_{\breve{T}_\ell}\}_{\ell\geq 0}$ (i.e., $\mathcal{F}_{\breve{T}_0}\subseteq\mathcal{F}_{\breve{T}_1}\subseteq\cdots$). For a σ-field $\mathscr{F}$ and a random variable X, denote $\sigma(X)$ as the σ-field generated by X, and $\sigma(\mathscr{F}, X)$ as the smallest σ-field that contains all the events belonging to $\mathscr{F}\cup\sigma(X)$.

Theorem 1 (Conditional distributions of $\breve{T}_{\ell+1}$ and $I_{\ell+1}$ given $\mathcal{F}_{\breve{T}_\ell}$) *Assume conditions* A1*,* A2*,* A3*,* A4*, and* A5 *in Appendix* B*. For each integer $\ell\geq 0$, define the set of event time points in the interval $(\breve{T}_\ell - \phi, \breve{T}_\ell]$ as:*

$$\mathcal{T}_\ell = \bigcup_{i\in\mathcal{V}} \big\{t\in(\breve{T}_\ell-\phi, \breve{T}_\ell] : N_i(\{t\}) = 1\big\}. \tag{4.9}$$

Define the $\mathcal{F}_{\breve{T}_\ell}$-measurable random variable:

$$T_\ell^* = \begin{cases} \min(\mathcal{T}_\ell)+\phi, & \text{if } \mathcal{T}_\ell\neq\varnothing, \\ \infty, & \text{if } \mathcal{T}_\ell = \varnothing. \end{cases} \tag{4.10}$$

We have the following results:

(i) (*Support of* $\breve{T}_{\ell+1}$) $\mathrm{P}(\breve{T}_\ell < \breve{T}_{\ell+1}\leq T_\ell^*) = 1$.

(ii) (*Conditional distribution of* $\breve{T}_{\ell+1}$) *If* $T_\ell^* < \infty$, *then* $\breve{T}_{\ell+1}$, *conditional on* $\mathcal{F}_{\breve{T}_\ell}$, *has a mixed-type probability distribution with probability mass function* (p.m.f.) *at the point* T_ℓ^*:

$$\mathrm{P}\big(\breve{T}_{\ell+1} = T_\ell^* \mid \mathcal{F}_{\breve{T}_\ell}\big) = \exp\{-\lambda^{\mathrm{sum}}(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}) \cdot (T_\ell^* - \breve{T}_\ell)\}, \tag{4.11}$$

and the probability density function (p.d.f.),

$$f_{\breve{T}_{\ell+1}|\mathcal{F}_{\breve{T}_\ell}}(t \mid \breve{T}_\ell) = \lambda^{\mathrm{sum}}(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}) \exp\{-\lambda^{\mathrm{sum}}(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}) \cdot (t - \breve{T}_\ell)\}, \quad \textit{for } t \in (\breve{T}_\ell, T_\ell^*), \tag{4.12}$$

where $\lambda^{\mathrm{sum}}(t \mid \mathscr{F}_t) = \sum_{i=1}^V \lambda_i(t \mid \mathscr{F}_t)$. *If* $T_\ell^* = \infty$, *then* (4.11) *and* (4.12) *reduce to* $(\breve{T}_{\ell+1} - \breve{T}_\ell) \mid \mathcal{F}_{\breve{T}_\ell} \sim \mathrm{Exp}(\lambda^{\mathrm{sum}}(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}))$.

(iii) (*Conditional distribution of* $I_{\ell+1}$) *If* $T_\ell^* < \infty$, *then for* $i \in \mathcal{V}$,

$$\mathrm{P}\big(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1})\big) = \begin{cases} 0, & \text{if } \breve{T}_{\ell+1} = T_\ell^*, \\ \lambda_i(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}) / \lambda^{\mathrm{sum}}(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}), & \text{if } \breve{T}_{\ell+1} \in (\breve{T}_\ell, T_\ell^*). \end{cases} \tag{4.13}$$

If $T_\ell^* = \infty$, *then* (4.13) *reduces to* $\mathrm{P}(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1})) = \lambda_i(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}) / \lambda^{\mathrm{sum}}(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell})$, *for* $i \in \mathcal{V}$ *and* $\breve{T}_{\ell+1} \in (\breve{T}_\ell, \infty)$.

The derivation of Theorem 1 primarily relies on the fact that the intensity functions $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ are constant within each interval $[\breve{T}_\ell, \breve{T}_{\ell+1})$. For instance, if $T_\ell^* < \infty$, (4.12) indicates that, conditional on $\mathcal{F}_{\breve{T}_\ell}$, the duration $\breve{T}_{\ell+1} - \breve{T}_\ell$ follows an exponential distribution with a rate $\lambda^{\mathrm{sum}}(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell})$ before $\breve{T}_{\ell+1}$ reaches T_ℓ^*. Furthermore, $\breve{T}_{\ell+1} = T_\ell^*$ implies that $\breve{T}_{\ell+1} \in \{T_{i,k} + \phi\}_{i\in\mathcal{V}, k\geq 1}$, while $\breve{T}_{\ell+1} < T_\ell^*$ indicates that the probability of the event $\{\breve{T}_{\ell+1} \in \{T_{i,k}\}_{k\geq 1}\}$, conditional on $\sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1})$, is proportional to the corresponding intensity $\lambda_i(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell})$ at node i. It is important to note that the V-dimensional intensity process $\boldsymbol{\lambda}(t \mid \mathscr{F}_t) = (\lambda_1(t \mid \mathscr{F}_t), \ldots, \lambda_V(t \mid \mathscr{F}_t))^\top$ in model (3.1) is not a piecewise deterministic Markov process (PDMP) [2], and thus the general results for PDMP do not apply to the derivation of Theorem 1.

Theorem 1 has two important applications. Firstly, it provides a simulation algorithm for generating synthetic point process data $\{T_{i,\ell}\}_{i\in\mathcal{V},\ell\geq 1}$ with intensities modeled by (3.1). By using the conditional probability distributions of $(\breve{T}_{\ell+1}, I_{\ell+1})$ given in (4.11)–(4.13), one can sequentially generate the marked time points $(\breve{T}_{\ell+1}, I_{\ell+1})$ for each $\ell \geq 0$, and then convert them into $\{T_{i,\ell}\}_{i\in\mathcal{V},\ell\geq 1}$ using (4.7). Secondly, Theorem 1 further leads to probabilistic results of $(\breve{\boldsymbol{T}}, \boldsymbol{I})$, as presented in Lemmas 5, 6, and 7, which are used to prove Theorem 2. For the sake of clarity, the following notations are required:

Duration τ_ℓ between two consecutive discontinuity time points $\breve{T}_\ell$:

$$\tau_\ell = \breve{T}_\ell - \breve{T}_{\ell-1}, \quad \ell \geq 1. \tag{4.14}$$

Event counts $M_{i,\ell}$ at node $i \in \mathcal{V}$:

$$M_{i,0} = 0, \quad M_{i,\ell} = \sum_{k=1}^{\ell} \mathrm{I}(I_k = i), \quad \ell \geq 1. \tag{4.15}$$

Piecewise-constant intensity at node $i \in \mathcal{V}$ within the time interval $[\breve{T}_\ell, \breve{T}_{\ell+1})$:

$$\lambda_{i,\ell} = \lambda_i(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}), \quad \ell \geq 0. \tag{4.16}$$

Lemma 5 (Expectation and Variance related to $(\breve{\boldsymbol{T}}, \boldsymbol{I})$) *Assume conditions* A1, A2, A3, A4, *and* A5 *in Appendix* B. *Then for each integer* $k \geq 1$, *we have:*

$$\begin{aligned} \mathrm{E}\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k \mid \mathcal{F}_{\breve{T}_{k-1}}\big\} &= 0, \\ \mathrm{var}\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k \mid \mathcal{F}_{\breve{T}_{k-1}}\big\} &= \mathrm{E}\big\{\mathrm{I}(I_k = i) \mid \mathcal{F}_{\breve{T}_{k-1}}\big\}, \end{aligned} \tag{4.17}$$

where $\lambda_{i,0} = \lambda_i(0)$. *Furthermore, for each integer* $\ell \geq 1$,

$$\mathrm{E}\Big(M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\,\tau_k\Big) = 0, \qquad \mathrm{var}\Big(M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\,\tau_k\Big) = \mathrm{E}(M_{i,\ell}).$$

Lemma 6 (Martingale property related to $(\breve{\boldsymbol{T}}, \boldsymbol{I})$) *Assume conditions* A1, A2, A3, A4, *and* A5 *in Appendix* B. *Then the random process* $\{M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\,\tau_k\}_{\ell \geq 1}$ *is a martingale with respect to* $\{\mathcal{F}_{\breve{T}_\ell}\}_{\ell \geq 1}$.

Lemma 7 (Upper bound for variance related to t-truncated $(\breve{\boldsymbol{T}}, \boldsymbol{I})$) *Assume conditions* A1, A2, A3, A4, *and* A5 *in Appendix* B. *For a given deterministic time point* $t \in (0, \infty)$, *let*

$$L_t = \sum_{\ell=1}^{\infty} \mathrm{I}(\breve{T}_\ell \leq t) \tag{4.18}$$

count the number of discontinuity points $\{\breve{T}_\ell\}_{\ell \geq 1}$ *that occur up to* t. *For integers* $\ell \geq 1$, *let*

$$\tau_\ell^{[t]} = \breve{T}_\ell \wedge t - \breve{T}_{\ell-1} \wedge t \tag{4.19}$$

be the duration between t*-truncated* $\breve{T}_\ell$ *and* $\breve{T}_{\ell-1}$, *where* $a \wedge b = \min(a, b)$. *Let* $\{X_\ell\}_{\ell \geq 0}$ *be a sequence of random variables such that* $X_\ell \geq 0$ *is measurable with respect to* $\mathcal{F}_{\breve{T}_\ell}$ *for each* $\ell \geq 0$, *and* $\sup_{\ell \geq 0} X_\ell \leq c_1$ a.s. *for a constant* $c_1 \in (0, \infty)$. *Then, for each* $i \in \mathcal{V}$, *we have*

$$\mathrm{E}\Big\{\sum_{k=1}^{L_t} X_{k-1}\,\mathrm{I}(I_k = i) - \sum_{k=1}^{L_t+1} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big\} = 0, \tag{4.20}$$

and

$$\operatorname{var}\Big\{ \sum_{k=1}^{L_t} X_{k-1}\, \mathrm{I}(I_k = i) - \sum_{k=1}^{L_t+1} X_{k-1}\, \lambda_{i,k-1}\, \tau_k^{[t]} \Big\} = \mathrm{E}\Big\{ \sum_{k=1}^{L_t} X_{k-1}^2\, \mathrm{I}(I_k = i) \Big\} \leq C_i\, c_1^2\, t, \tag{4.21}$$

where the constant $C_i = \exp(\beta_{0;i} + C_0 \cdot \sum_{j\in\mathcal{V}} \beta_{j,i})$*, with* $C_0 \in (0,\infty)$ *provided in Condition* A5*.*

Remark 3 *Derivations of Lemmas* 5–7 *are outlined as follows. Lemma* 5 *is obtained from direct calculations based on the probability distribution of* $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ *in Theorem* 1*. Lemma* 6 *follows from* (4.17) *in Lemma* 5*. Lemma* 7 *is a non-trivial extension of Lemma* 5*, where a non-random index* ℓ *is replaced with a random index* L_t*; Lemma* 7 *aims to study the properties of the marked point process* $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ *when truncated by a fixed time point* $t \in (0,\infty)$*, which is further used to translate these results into the forms of the counting process* $\boldsymbol{N}(t)$*.*

4.1.4 Translating results of $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ into results of $\boldsymbol{N}(t)$

The equivalence verified in (4.8), between the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ and the counting process $\boldsymbol{N}(t)$, enables us to translate the results of Lemmas 5–7 into the counterparts of $\boldsymbol{N}(t)$ and directly obtain Theorem 2 below, which describes some useful properties of $\boldsymbol{N}(t)$.

Theorem 2 (Upper bounds for variances related to $\boldsymbol{N}(t)$; finiteness of $\boldsymbol{N}(t)$) *Assume conditions* A1*,* A2*,* A3*,* A4*, and* A5 *in Appendix* B*. Then, there exists a constant* $c_1 \in (0,\infty)$ *such that for any* $i \in \mathcal{V}$ *and any* $t \in (0,\infty)$*, we have*

$$\operatorname{var}\Big\{ N_i(t) - \int_0^t \lambda_i(u \mid \mathscr{F}_u)\, \mathrm{d}u \Big\} = \mathrm{E}\{N_i(t)\} \leq c_1\, t, \tag{4.22}$$

which implies that the counting process $N_i(t)$ *is finite* a.s.*, i.e.,*

$$\mathrm{P}(N_i(t) < \infty) = 1, \quad i \in \mathcal{V}. \tag{4.23}$$

Furthermore, for a random process $\{x(t)\}_{t\geq 0}$ *such that* $x(t)$ *is* $\mathscr{F}_t$*-measurable,* $0 \leq \inf_{t\geq 0} x(t) \leq \sup_{t\geq 0} x(t) \leq c_2$ a.s. *for a constant* $c_2 \in (0,\infty)$*, and* $x(t)$ *is constant in the interval* $[\breve{T}_\ell, \breve{T}_{\ell+1})$ *for each integer* $\ell \geq 0$*, it follows that for any* $t \in (0,\infty)$*,*

$$\operatorname{var}\Big[\int_0^t \big\{ x(u-)\, \mathrm{d}N_i(u) - x(u)\, \lambda_i(u \mid \mathscr{F}_u)\, \mathrm{d}u \big\} \Big] = \mathrm{E}\Big\{ \int_0^t x^2(u)\, \lambda_i(u \mid \mathscr{F}_u)\, \mathrm{d}u \Big\} \leq c_1\, c_2^2\, t, \tag{4.24}$$

where $x(u-) = \lim_{t\uparrow u} x(t)$ *denotes the left limit.*

By considering the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$, we obtain Theorem 2, which ensures certain fundamental probabilistic properties of our counting process $\boldsymbol{N}(t)$. In the subsequent discussions in Section 5, we will demonstrate the significance of these results in deriving the associated statistical properties, as presented in Theorems 5–7 and Corollary 1.

4.2 Cyclicity and asymptotic mean stationarity of $\boldsymbol{N}(t)$

A counting process $\{N(t)\}_{t\geq 0}$ is said to be *strict-sense stationary* if, for any time point $s \in [0,\infty)$, $N(t+s) - N(s) \overset{\mathrm{D}}{=} N(t)$ for each $t \geq 0$, where $X_1 \overset{\mathrm{D}}{=} X_2$ denotes that random quantities X_1 and X_2 have identical distributions (see [11] and references therein). Throughout this paper, the term "stationarity" refers to *strict-sense stationarity*, while non-stationarity refers to other cases. A strict-sense stationary counting process $\{N(t)\}_{t\geq 0}$ exhibits various well-known properties, some of which are listed as follows:

(P1) Invariant distribution of the conditional intensity function: The probability distribution of the conditional intensity function $\lambda(t \mid \mathscr{F}_t)$ defined in (2.6) and (2.7) remains invariant for any $t \in [0,\infty)$.

(P2) Constant mean intensity: For any $t \in [0,\infty)$, the mean intensity function satisfies $\mathrm{E}\{\lambda(t \mid \mathscr{F}_t)\} \equiv \lambda_0$ for some constant $\lambda_0 \in (0,\infty)$.

(P3) Expectation of increments: For any $t \in (0,\infty)$ and $s \in (0,\infty)$, $\mathrm{E}\{N(t+s) - N(s)\} = \lambda_0 \times t$. Furthermore, if $N(t)$ is ergodic, then $\lim_{t\to\infty} N(t)/t = \lambda_0$ a.s..

(P4) Finiteness of $N(t)$: for any $t \in (0,\infty)$, $\mathrm{P}(N(t) < \infty) = 1$.

These features resulting from the stationarity assumption significantly facilitate theoretical analysis. Therefore, the stationarity assumption is widely imposed in the relevant literature, e.g., [20, 22, 36]. We refer to a multivariate counting process $\{\boldsymbol{N}(t)\}_{t\geq 0}$ as *strict-sense stationary* if $\{N_i(t)\}_{t\geq 0}$ is strict-sense stationary for each $i \in \mathcal{V}$.

However, the counting process $\boldsymbol{N}(t)$ associated with the conditional intensity functions $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ in model (3.1) is not strict-sense stationary. Lemma B.9 in Appendix B justifies that $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ in (3.1) violates property (P1) of stationarity. Without possessing properties (P1)–(P4) listed above, a non-stationary point process poses significant challenges to theoretical analysis. Therefore, it becomes necessary to explore alternative properties for non-stationary point processes using a new approach.

Recall that a Poisson process assumes the independent increment property, which leads to the memoryless property [24]. In other words, for any $s \in (0, \infty)$, the time-shifted counting process $\{N(t+s) - N(s)\}_{t \geq 0}$ is independent of the history up to the time point s. In Theorem 3, we will demonstrate that our study establishes a relaxed version of this memoryless property for our $\boldsymbol{N}(t)$. Specifically, the counting process $\{\boldsymbol{N}(t + R_\ell) - \boldsymbol{N}(R_\ell)\}_{t \geq 0}$, at random time points R_ℓ, is independent of the σ-field $\mathcal{F}_{R_\ell}$, where R_ℓ and $\mathcal{F}_{R_\ell}$ are introduced in Definition 3.

Definition 3 (Recurrence time points R_ℓ, recurrence cycle of $\boldsymbol{N}(t)$, and $\mathcal{F}_{R_\ell}$) *Let $R_0 = 0$. For each integer $\ell \geq 1$, let R_ℓ be the first time point, after $R_{\ell-1} + \phi$, such that no events occur at any node in the time interval $(R_\ell - \phi, R_\ell]$, i.e.,*

$$R_\ell \;\; = \;\; \min\{t \geq R_{\ell-1} + \phi : \boldsymbol{N}((t-\phi, t]) = \boldsymbol{0}\}. \tag{4.25}$$

We call R_ℓ the ℓth recurrence time point and the interval $(R_{\ell-1}, R_\ell]$ the ℓth recurrence cycle. Denote by $\mathcal{F}_{R_\ell} = \{A \in \mathscr{F} : A \cap \{R_\ell \leq t\} \in \mathscr{F}_t \text{ for every } t > 0\}$ the stopping-time σ-algebra with respect to R_ℓ.

Figure 3 illustrates the recurrence time points R_ℓ. For our $\boldsymbol{N}(t)$, the existence of R_ℓ is verified by Lemma 8.

Lemma 8 (Existence of R_ℓ) *Assume conditions* A1*,* A2*,* A3*,* A4*, and* A5 *in Appendix* B*. For each integer $\ell \geq 1$, the recurrence time point R_ℓ in Definition* 3 *exists with probability one.*

The memoryless property induced by R_ℓ can be intuitively explained as follows. In our model (3.1), the intensity functions $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i \in \mathcal{V}}$ at the current time t primarily depend on the historical event counts $\boldsymbol{N}((t-\phi, t])$ in the lag window $(t-\phi, t]$ of a fixed length ϕ. Once the counting process $\boldsymbol{N}(t)$ reaches a recurrence time point $t = R_\ell$, all event counts in the lag window $(t-\phi, t]$ become empty, which separates the dependence of the future intensity processes $\{\lambda_i(t \mid \mathscr{F}_t) : t \geq R_\ell, i \in \mathcal{V}\}$ on the past event history up to that time point R_ℓ. At $t = R_\ell$, both the intensity processes and the counting process "reset," becoming independent of $\mathcal{F}_{R_\ell}$, and initiating a renewed "cyclic" process. Based on these considerations, we establish a new cyclicity property of $\boldsymbol{N}(t)$, formally presented in Theorem 3.

Theorem 3 (Cyclicity of $\boldsymbol{N}(t)$ driven by R_ℓ) *Assume conditions* A1, A2, A3, A4, *and* A5 *in Appendix* B. *Let* $\boldsymbol{N}(t)$ *be the counting process with the intensity processes* $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ *in* (3.1). *Then, for each recurrence time point* R_ℓ *in* (4.25) *with* $\ell \geq 1$,

(i) *both* $\{\lambda_i(t + R_\ell \mid \mathcal{F}_{t+R_\ell})\}_{i\in\mathcal{V}}$ *and* $\boldsymbol{N}(t + R_\ell) - \boldsymbol{N}(R_\ell)$ *are independent of* $\mathcal{F}_{R_\ell}$, *with* $\boldsymbol{N}(t + R_\ell) - \boldsymbol{N}(R_\ell) \overset{\mathrm{D}}{=} \boldsymbol{N}(t)$ *for each* $t \geq 0$.

(ii) $\{\boldsymbol{N}((R_{\ell-1}, R_\ell]) : \ell \geq 1\}$ *is a sequence of* i.i.d. *random vectors.*

(iii) $\{R_\ell - R_{\ell-1} : \ell \geq 1\}$ *is a sequence of* i.i.d. *random variables with finite second moment.*

This cyclicity property of $\boldsymbol{N}(t)$ will be used to derive Theorem 4 below, as well as Theorem 5 in Section 5.1. In comparison, our cyclicity property is analogous to the renewal property of the non-linear Hawkes process [10] or queuing models [23, 27]. However, tools for deriving the renewal property are not directly applicable to model (3.1), as it violates some basic assumptions underlying the non-linear Hawkes process and queuing models. For example, our point process (when $\mathcal{E} \neq \varnothing$) does not meet the assumption of a deterministic arrival rate required in $\mathrm{M}_t/\mathrm{G}/\infty$ queues, and the non-linear Hawkes process does not allow for the general type of shape function $g(\cdot)$ in model (3.1).

Theorem 4 (Asymptotic mean stationarity of $\boldsymbol{N}(t)$) *Assume conditions* A1, A2, A3, A4, *and* A5 *in Appendix* B. *Then, there exists a constant vector* $\boldsymbol{c}_0 \in (0, \infty)^V$ *such that the counting process* $\boldsymbol{N}(t)$ *associated with the intensity processes* $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ *in* (3.1) *satisfies*

$$\boldsymbol{N}(t)/t \xrightarrow{\mathrm{P}} \boldsymbol{c}_0, \quad \text{as } t \to \infty. \tag{4.26}$$

Theorem 4 verifies that the vector $\boldsymbol{N}(t)/t$ of average counts converges in probability to a constant vector as t approaches infinity. For a non-stationary counting process $\boldsymbol{N}(t)$, this type of property is called *asymptotic mean stationarity* (a notion used in [31]). It is noted that the constant $\boldsymbol{c}_0$ in (4.26) deterministically depends on the network parameters $\{\beta_{0;i}\}$ and $\{\beta_{j,i}\}$ in model (3.1), but a closed-form formulation of this dependence is not available due to the non-linearities of both the exponential link function and the shape function $g(\cdot)$ in model (3.1). This is in contrast with the case of a linear Hawkes process [4], for which a closed-form moment equation (Equation (3) in [4]) could be constructed to relate the mean intensity with the network parameters. Nevertheless, without knowing the explicit value of $\boldsymbol{c}_0$, Theorem 4 suffices to assist in proving further useful statistical convergence properties, as will be shown in Section 5.

In summary, Lemma B.9 states the fact that our counting process $\boldsymbol{N}(t)$ is not *strict-sense stationary*; nevertheless, we have verified that $\boldsymbol{N}(t)$ possesses some desirable properties similar to stationary processes. For example, Theorem 4 is similar to the ergodicity in property (P3); Theorem 2 verifies property (P4); and Theorem 3(i) indicates a feature similar to the shift invariance property of stationarity. Theorems 3 and 4 are crucial for deriving the related statistical asymptotic properties (in Theorems 5–7) in Section 5. In comparison with existing results, technical tools we have developed are easier to interpret and utilize.

5 Parameter estimation via penalized M-estimation

Our primary interest is to learn the network structure from the observed data $\{\boldsymbol{T}_i\}_{i\in\mathcal{V}}$ in (1.1) of the multivariate point process in the time interval $[0, \mathrm{T}]$, where $\mathrm{T} \in (0, \infty)$ is the total time length of the experiment. We denote the true values of the conditional intensity function (3.1) as

$$\lambda_i^*(t \mid \mathscr{F}_t) \;=\; \exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^*\big\}, \tag{5.1}$$

where $\widetilde{\boldsymbol{\beta}}_i^* = (\beta_{0;i}^*, \boldsymbol{\beta}_i^{*\top})^\top = (\beta_{0;i}^*, \beta_{1,i}^*, \ldots, \beta_{i-1,i}^*, \beta_{i+1,i}^*, \ldots, \beta_{V,i}^*)^\top \in \mathbb{R}^V$ is the vector of true parameters, and $\widetilde{\boldsymbol{x}}_i(t) = (1, \boldsymbol{x}_i(t)^\top)^\top = (1, x_1(t), \ldots, x_{i-1}(t), x_{i+1}(t), \ldots, x_V(t))^\top \in \mathbb{R}^V$ is the vector of regression covariates. Our statistical learning aims to estimate $\widetilde{\boldsymbol{\beta}}_i^*$ in (5.1) and recover the true network structure $\mathcal{G}^* = \{\mathcal{V}, \mathcal{E}^*\}$, where the true edge set $\mathcal{E}^* = \mathcal{E}_+^* \cup \mathcal{E}_-^*$ corresponds to $\mathcal{E} = \mathcal{E}_+ \cup \mathcal{E}_-$ in (3.2) with parameters $\beta_{j,i}$ replaced by $\beta_{j,i}^*$.

The existing parameter estimation methods can be categorized into two categories: (i) moment or correlation-based approaches [4, 25]; and (ii) intensity-based approaches [20, 42]. The moment or correlation-based approaches are typically applied to the linear models of $\lambda_i(t \mid \mathscr{F}_t)$ and are not suitable for our non-linear model (3.1). Therefore, we adopt the intensity-based approach, where parameter estimation is achieved through the minimization of a suitable loss function that measures the discrepancy between the true and estimated intensity processes.

5.1 Loss function

In the existing literature, there are different loss functions used for estimating parameters in a generic counting process $N(t)$ associated with an intensity process $\lambda(t \mid \mathscr{F}_t)$, including the negative log-likelihood function [38, 42]:

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\log\{\lambda(t- \mid \mathscr{F}_{t-})\} \, \mathrm{d}N(t) - \lambda(t \mid \mathscr{F}_t) \, \mathrm{d}t \Big], \tag{5.2}$$

and the squared loss: [20, 36]

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big\{ \lambda^2(t \mid \mathscr{F}_t) \, \mathrm{d}t - 2\lambda(t- \mid \mathscr{F}_{t-}) \, \mathrm{d}N(t) \Big\}. \tag{5.3}$$

where $\lambda(t- \mid \mathscr{F}_{t-}) = \lim_{u \uparrow t} \lambda(u \mid \mathscr{F}_u)$ denotes the left limit.

The squared loss (5.3) is more suitable for linear models of $\lambda(t \mid \mathscr{F}_t)$, such as the linear Hawkes process [36], while the negative log-likelihood function (5.2) is typically used for non-linear cases, such as when using an exponential link function in model (3.1). Therefore, we will focus our discussion on the use of (5.2). In our multi-dimensional setting, we choose to estimate $\widetilde{\boldsymbol{\beta}}_i^*$ at individual nodes i, and recover the network structure by aggregating estimators of $\{\widetilde{\boldsymbol{\beta}}_i^*\}_{i \in \mathcal{V}}$ using the following loss function:

$$\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) = -\frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\widetilde{\boldsymbol{x}}_i(t-)^\top \widetilde{\boldsymbol{\beta}}_i \, \mathrm{d}N_i(t) - \exp\big\{ \widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i \big\} \, \mathrm{d}t \Big], \tag{5.4}$$

where $\widetilde{\boldsymbol{\beta}}_i = (\beta_{0;i}, \boldsymbol{\beta}_i^\top)^\top = (\beta_{0;i}, \beta_{1,i}, \ldots, \beta_{i-1,i}, \beta_{i+1,i}, \ldots, \beta_{V,i})^\top \in \mathbb{R}^V$ represents a vector of generic parameters.

In many application fields [13, 8, 34, 43], the number of recorded event time points could be large, often in the order of millions or more. This motivates us to study the behavior of our estimation approach for a large number $N_i(\mathrm{T})$ of event time points, or equivalently, a long total time length T. Theorem 5 presents the asymptotic convergence results for the gradient vector and the Hessian matrix of $\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)$ as T approaches infinity. These results will be used to derive parameter estimation consistency (Theorems 6 and 7) in Section 5.3.

Theorem 5 (Asymptotic convergence related to loss function $\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ in (5.4)) *Assume conditions* A1*,* A2*,* A3*,* A4*,* A5*, and* A6 *in Appendix* B*. For each $i \in \mathcal{V}$, denote $\nabla \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ and $\nabla^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ as the gradient vector and Hessian matrix of $\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ in* (5.4) *respectively. Then, we have the following results as $\mathrm{T} \to \infty$:*

(i) *$\nabla \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)$ converges to $\mathbf{0}$ in probability at a square-root rate, i.e.,*

$$\nabla \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\widetilde{\boldsymbol{x}}_i(t) \, \exp\big\{ \widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^* \big\} \, \mathrm{d}t - \widetilde{\boldsymbol{x}}_i(t-) \, \mathrm{d}N_i(t) \Big] = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}}). \tag{5.5}$$

(ii) *There exists a constant matrix* $\mathbf{C}_i$ *such that*

$$\nabla^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \widetilde{\boldsymbol{x}}_i(t)\, \widetilde{\boldsymbol{x}}_i(t)^\top \exp\left\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^*\right\} \mathrm{d}t \xrightarrow{\mathrm{P}} \mathbf{C}_i. \tag{5.6}$$

Furthermore, the matrix $\mathbf{C}_i$ *is positive definite with all entries positive.*

Theorem 5 is derived from the probabilistic results of $\boldsymbol{N}(t)$ in Section 4. Specifically, (5.5) is obtained from the bounded variance property (4.24) of $\boldsymbol{N}(t)$ in Theorem 2, which is derived from the properties of the marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ (in Theorem 1 and Lemmas 5–7). (5.6) applies the cyclicity property of $\boldsymbol{N}(t)$ in Theorem 3 and the asymptotic mean stationarity of $\boldsymbol{N}(t)$ in Theorem 4.

Remark 4 *Conventional tools for asymptotic results, such as the law of large numbers or central limit theorems, are not directly applicable to Theorem* 5 *due to the distinctive features of the stochastic processes* $\boldsymbol{N}(t)$ *and* $\widetilde{\boldsymbol{x}}_i(t)$ *in* (5.5) *and* (5.6)*. Specifically, the non-stationary counting process* $\boldsymbol{N}(t)$ *is closely linked with the historical events up to* t *(via its associated intensity processes* $\{\lambda_i^*(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ *modeled by* (5.1)*), resulting in a complicated dependence structure of* $\boldsymbol{N}(t)$ *across time* t*. Additionally, the stochastic process* $\widetilde{\boldsymbol{x}}_i(t)$*, defined as* $\widetilde{\boldsymbol{x}}_i(t) = (1, x_1(t), \ldots, x_{i-1}(t), x_{i+1}(t), \ldots, x_V(t))^\top$*, relies on the special type of stochastic process* $N_j((t-\phi, t])$ *(see* (3.5)*,* (3.6) *and* (4.1)*), for which probabilistic properties are not available in the existing literature. The use of Theorems* 1–4 *enables us to prove Theorem* 5*, justifying the importance of our results in Section* 4*.*

5.2 Penalized estimation of parameters

Sparsity assumptions are commonly imposed on the true network structure in various real-world applications (e.g., [20, 44, 46]). To promote a sparse network structure with the most significant interactions, we employ the weighted L_1-penalty:

$$\mathcal{P}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) = \sum_{j\in\mathcal{V}\setminus i} w_{j,i,\mathrm{T}}\, |\beta_{j,i}|, \tag{5.7}$$

where $\{w_{j,i,\mathrm{T}} : j \in \mathcal{V} \setminus i\}$ represent non-negative weights. We estimate the true parameter vector $\widetilde{\boldsymbol{\beta}}_i^*$ using the *penalized M-estimator*, which minimizes the sum of the loss function (5.4) and the penalty function (5.7):

$$\begin{aligned} \widehat{\widetilde{\boldsymbol{\beta}}}_i &= \arg\min_{\widetilde{\boldsymbol{\beta}}_i\in\mathbb{R}^V} \left\{\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) + \mathcal{P}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)\right\} \\ &= \arg\min_{\widetilde{\boldsymbol{\beta}}_i\in\mathbb{R}^V} \left\{ \tfrac{1}{\mathrm{T}} \textstyle\int_0^{\mathrm{T}} \left[\exp\left\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i\right\} \mathrm{d}t - \widetilde{\boldsymbol{x}}_i(t-)^\top \widetilde{\boldsymbol{\beta}}_i\, \mathrm{d}N_i(t) \right] + \sum_{j\in\mathcal{V}\setminus i} w_{j,i,\mathrm{T}}\, |\beta_{j,i}| \right\}, \end{aligned} \tag{5.8}$$

where the vector $\widehat{\widetilde{\boldsymbol{\beta}}}_i = (\widehat{\beta}_{0;i}, \widehat{\beta}_{1,i}, \ldots, \widehat{\beta}_{i-1,i}, \widehat{\beta}_{i+1,i}, \ldots, \widehat{\beta}_{V,i})^\top$ collects $\widehat{\beta}_{0;i}$ and all $\{\widehat{\beta}_{j,i} : j \in \mathcal{V} \setminus i\}$. The estimated network is obtained as follows:

$$\widehat{\mathcal{E}} = \{(j,i) \in \mathcal{V} \times \mathcal{V} : \widehat{\beta}_{j,i} \neq 0,\, j \neq i\}.$$

Furthermore, considering that the sign of an estimator indicates the type of effect, we estimate the sets of excitatory and inhibitory effects separately:

$$\widehat{\mathcal{E}}_+ = \{(j,i) \in \mathcal{V} \times \mathcal{V} : \widehat{\beta}_{j,i} > 0,\, j \neq i\}, \tag{5.9}$$

$$\widehat{\mathcal{E}}_- = \{(j,i) \in \mathcal{V} \times \mathcal{V} : \widehat{\beta}_{j,i} < 0,\, j \neq i\}. \tag{5.10}$$

5.3 Asymptotic results for structure learning

For continuous-time point process data, the total time length T is roughly proportional to the number of the observed data points and typically serves as the sample size (e.g., in [20, 39]). Therefore, in this section, we establish the asymptotic properties of the penalized M-estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ in (5.8) with respect to T approaching infinity. To establish estimation consistency, we first provide the following conditions for the weights $w_{j,i,\mathrm{T}}$ in (5.7):

$$\max_{j \in \mathrm{Pa}^*(i)} w_{j,i,\mathrm{T}} = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}}); \tag{5.11}$$

$$\max_{j \in \mathrm{Pa}^*(i)} w_{j,i,\mathrm{T}} = o_{\mathrm{P}}(\sqrt{1/\mathrm{T}}); \tag{5.12}$$

$$\min_{j \in \mathcal{V} \setminus \{\mathrm{Pa}^*(i) \cup i\}} \sqrt{\mathrm{T}}\, w_{j,i,\mathrm{T}} \xrightarrow{\mathrm{P}} \infty, \quad \text{as } \mathrm{T} \to \infty. \tag{5.13}$$

Here, $\mathrm{Pa}^*(i) = \{j \in \mathcal{V} \setminus i : \beta^*_{j,i} \neq 0\}$ denotes the nodes that have a true non-zero effect on node i; condition (5.12) employed in Theorem 7 and Corollary 1 is stronger than condition (5.11) used in Theorem 6. An example of weights $\{w_{j,i,\mathrm{T}}\}$ that satisfy (5.12) and (5.13) is the adaptive lasso penalty [47], in which $w_{j,i,\mathrm{T}} = \eta_{\mathrm{T}}\, |\breve{\beta}_{j,i}|^\gamma$, with $\eta_{\mathrm{T}} = O(1/\mathrm{T}^a)$ for $1/2 < a < 3/2$, $\gamma = -2$, and $\breve{\widetilde{\boldsymbol{\beta}}}_i = (\breve{\beta}_{0;i}, \breve{\beta}_{1,i}, \ldots, \breve{\beta}_{i-1,i}, \breve{\beta}_{i+1,i}, \ldots, \breve{\beta}_{V,i})^\top$ denoting the minimizer of $\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$.

Theorem 6 guarantees the existence of a $\sqrt{1/\mathrm{T}}$-consistent estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ in (5.8).

Theorem 6 (Existence of a consistent penalized M-estimator) *Assume conditions* A1, A2, A3, A4, A5, A6, *and* A7 *in Appendix* B. *Assume* (5.11) *for the weights* $w_{j,i,\mathrm{T}}$. *Then, there exists a local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ *in* (5.8) *such that* $\|\widehat{\widetilde{\boldsymbol{\beta}}}_i - \widetilde{\boldsymbol{\beta}}^*_i\| = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}})$, *as* $\mathrm{T} \to \infty$.

Following Theorem 6, the sparsistency of the penalized M-estimator is given in Theorem 7 below. Before stating it, we introduce some notations. We partition the true parameter vector as $\widetilde{\boldsymbol{\beta}}_i^* = (\beta_{0;i}^*, \boldsymbol{\beta}_i^{*\top})^\top = (\beta_{0;i}^*, \boldsymbol{\beta}_i^{*(\mathrm{I})\top}, \boldsymbol{\beta}_i^{*(\mathrm{II})\top})^\top = (\widetilde{\boldsymbol{\beta}}_i^{*(\mathrm{I})\top}, \boldsymbol{\beta}_i^{*(\mathrm{II})\top})^\top$, where $\boldsymbol{\beta}_i^{*(\mathrm{II})} = \mathbf{0}$, and $\boldsymbol{\beta}_i^{*(\mathrm{I})}$ collects all the non-zero components in $\boldsymbol{\beta}_i^*$. Similarly, for the estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_i$, we adopt the partition $\widehat{\widetilde{\boldsymbol{\beta}}}_i = (\widehat{\beta}_{0;i}, \widehat{\boldsymbol{\beta}}_i^{(\mathrm{I})\top}, \widehat{\boldsymbol{\beta}}_i^{(\mathrm{II})\top})^\top = (\widehat{\widetilde{\boldsymbol{\beta}}}_i^{(\mathrm{I})\top}, \widehat{\boldsymbol{\beta}}_i^{(\mathrm{II})\top})^\top$, with index sets I and II corresponding to those of $\boldsymbol{\beta}_i^{*(\mathrm{I})}$ and $\boldsymbol{\beta}_i^{*(\mathrm{II})}$, respectively.

Theorem 7 (Sparsistency of the penalized M-estimator) *Assume conditions* A1*,* A2*,* A3*,* A4*,* A5*,* A6*, and* A7 *in Appendix* B*. Assume that the weights $w_{j,i,\mathrm{T}}$ satisfy* (5.12) *and* (5.13)*. Then, any $\sqrt{1/\mathrm{T}}$-consistent local minimizer $\widehat{\widetilde{\boldsymbol{\beta}}}_i = (\widehat{\widetilde{\boldsymbol{\beta}}}_i^{(\mathrm{I})\top}, \widehat{\boldsymbol{\beta}}_i^{(\mathrm{II})\top})^\top$ in* (5.8) *satisfies*

$$\mathrm{P}\big(\widehat{\boldsymbol{\beta}}_i^{(\mathrm{II})} = \mathbf{0}\big) \to 1, \quad \text{as } \mathrm{T} \to \infty. \tag{5.14}$$

The sparsistency result in (5.14) immediately yields the network recovery consistency stated in Corollary 1.

Corollary 1 (Network recovery consistency) *Assume the same conditions as in Theorem* 7*. Then, the network structure estimators $\widehat{\mathcal{E}}_+$ in* (5.9) *and $\widehat{\mathcal{E}}_-$ in* (5.10)*, based on $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ in* (5.8)*, are consistent with the true edges $\mathcal{E}_+^*$ and $\mathcal{E}_-^*$, respectively. In other words, $\mathrm{P}(\widehat{\mathcal{E}}_+ = \mathcal{E}_+^*, \widehat{\mathcal{E}}_- = \mathcal{E}_-^*) \to 1$ as $\mathrm{T} \to \infty$.*

Corollary 1 demonstrates that our method can consistently recover the true network structure as the total time length T increases. This provides theoretical support for the utility of our proposed statistical learning procedure.

Remark 5 *Standard results of parameter estimation consistency for general M-estimators have been well established in existing statistical literature, as discussed in Chapter* 5 *of* [*41*]*. However, the general theory is not directly applicable to prove the consistency results presented in Theorems* 6 *and* 7 *in our context. This is due to the fact that our loss function* (5.4) *is based on a complicated stochastic integral, while the loss function in* [*41*] *is typically restricted to simple sum statistics form. The proofs of Theorems* 6 *and* 7 *rely on the asymptotic convergence of $\nabla \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)$ and $\nabla^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)$ in Theorem* 5*, which is derived from the probabilistic results of $\boldsymbol{N}(t)$ in Theorems* 1–4*.*

6 Simulation study

In this section, we conduct numerical experiments to demonstrate the practical utility of our continuous-time modeling approach and estimation procedure.

6.1 Types of network structures

The simulation studies consider three simulated networks, as depicted in Figure 4, representing networks with varying degrees of complexity. Network-1 is a simple network of 10 nodes, including 6 excitatory and 4 inhibitory effects. It is designed to resemble a directed acyclic graph, aiming to capture the information flow from sensory neurons to motor neurons. Network-2, adapted from [46], is a moderately complex network comprising 20 nodes, with 12 excitatory and 8 inhibitory effects. This network is intended to mimic the potential *hub* and *leaf* structures observed in neuron ensembles. Specifically, nodes 1, 6, 11, and 16 are hub nodes with a degree of 6, while the remaining nodes are leaves with a degree of 1. Network-3 is a complex network consisting of 50 nodes, with 30 excitatory and 30 inhibitory effects. It shares the same design motivation as Network-1. The synthetic multivariate point process data were generated using the simulation algorithm induced by Theorem 1. The conditional intensity functions in model (3.1) employ

$$x_j(t) = g\big(r_{j,\phi}(t)\big),\ j \in \mathcal{V},\ t \in [0, \mathrm{T}],\ \text{where } \phi = 1,\ \text{and } g(x) = \log(1 + x \wedge 10). \qquad (6.1)$$

6.2 Comparison of methods

We compare the following estimation procedures:

(i) Continuous-time modeling (our proposed method): This method estimates the parameters using the penalized M-estimator in (5.8) with two scenarios for the penalty (5.7): the L_1-penalty $\mathcal{P}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) = \eta \sum_{j \in \mathcal{V}\setminus i} |\beta_{j,i}|$; the weighted-$L_1$ penalty $\mathcal{P}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) = \sum_{j \in \mathcal{V}\setminus i} \eta_{\mathrm{T}} |\breve{\beta}_{j,i}|^{\gamma} \cdot |\beta_{j,i}|$, with $\gamma = -2$, and the M-estimator $\breve{\beta}_{j,i}$ of $\beta^*_{j,i}$, where the tuning parameters η and η_{T} are selected using the Bayesian Information Criterion (BIC) [32].

(ii) Discrete-time approximation modeling: Method (ii) utilizes the discrete-time approximation. The entire time interval $[0, \mathrm{T}]$ is divided into n equally-spaced time bins $\{(t_{k-1}, t_k] : k = 1, \ldots, n\}$, each of length T/n. The observed point process $\{\boldsymbol{T}_i\}_{i \in \mathcal{V}}$ is transformed into sequences of bin counts $\{N_{i,k}\}_{i \in \mathcal{V}; k = 1, \ldots, n}$. The interaction parameters are estimated using a penalized M-estimation similar to (5.8). However, in this case, a Poisson distribution with rate $\lambda_i(t_{k-1})\,\mathrm{T}/n$ is assumed for $N_{i,k}$ at node i.

(iii) Discrete-time modeling with groups of connection parameters in [46]: Method (iii) is similar to method (ii), with the difference that the effect from node j to i is modeled by a group of parameters $\{\beta_{j,i,q} : q = 1, \ldots, Q\}$ instead of a single parameter $\beta_{j,i}$. Here, Q is determined by $Q = [\phi/(\mathrm{T}/n)]$.

(iv) SIE-GLM method in [44]: Method (iv) is an extension of method (iii) that incorporates structural information in the parameter space. It employs the sparse group lasso penalty for parameter estimation.

(v) Bayesian method in [35]: Method (v) is a continuous-time modeling approach that uses a default shape-function $\log(1 + x)$ and the same loss function as our proposed method (i). This method explores all subsets of components in $\widetilde{\boldsymbol{\beta}}_i$ and selects the best subset with the maximum Bayesian posterior density. To ensure a fair comparison, method (v) assumes a uniform prior (i.e., no prior information) for $\widetilde{\boldsymbol{\beta}}_i$.

All methods which involve ϕ and $g(\cdot)$ for parameter estimation adopt our empirical choices: $\phi = 1$ and $g(x) = \log(1 + x \wedge c)$, where the data-driven choice c is given in (A.2), unless stated otherwise. The coordinate descent algorithm [18] is utilized to solve (5.8). To aid in further discussion, we categorize all methods in Table 1.

6.3 Simulation results

We consider three different total time lengths: $\mathrm{T} \in \{500, 1000, 2000\}$. For each $i \in \mathcal{V}$, the true baseline intensity parameter is $\beta^*_{0;i} = -0.8$, resulting in a base rate of approximately $\exp(-0.8) \approx 0.45$. The true connection strength parameters $\{\beta^*_{j,i} : i, j \in \mathcal{V}, j \neq i\}$ are set as follows: $\beta^*_{j,i}$ is β for the excitatory effect, $-\beta$ for the inhibitory effect, and 0 for no effect from node j to node i. Here $\beta \in \{0.4, 0.5\}$ reflects the magnitude of the connection strength.

The performance of each method is evaluated using the following criterion measures: Corret_All (correctly detected number of excitatory and inhibitory effects), Detected_A (correctly detected number of excitatory effects), Detected_B (correctly detected number of inhibitory effects), and Correct_NC (correctly detected number of non-effects). For comparison, Corret_All, Detected_A, and Detected_B reflect the sensitivity level, which is defined as the percentage of correctly identified effects. It measures how sensitive each method is in detecting excitatory or inhibitory effects. Additionally, Correct_NC indicates the specificity level, defined as the percentage of correctly identified non-effects. It represents the ability of the method to correctly identify the absence of an effect.

6.3.1 Complex network

For complex network Network-3, we first compare the performance of each method under different connection strengths $\beta \in \{0.4, 0.5\}$ in Table 2. Most methods are successful in detecting the sparse structure of the network, correctly identifying most true non-effects and achieving a good level of specificity. However, the sensitivity results are relatively worse compared to specificity. All methods with $\beta = 0.5$ exhibit better sensitivity results compared to $\beta = 0.4$. This is expected since a larger connection strength parameter implies stronger interaction between nodes, making detection easier. In both strength parameter settings, continuous-time methods (Continuous_L1 and Continuous_wL1) outperform the discrete-time approximation methods (Discrete_L1, Discrete_wL1, bin = 0.5, 0.25, 0.1) in terms of sensitivity. It is worth noting that for Discrete_L1 and Discrete_wL1, a smaller bin width yields better results but does not surpass the corresponding continuous-time methods Continuous_L1 and Continuous_wL1. This observation suggests that continuous-time modeling can be considered as a limiting case of discrete-time modeling when the bin width approaches zero, thus providing the most accurate results. Regarding the penalty choices in methods (i) and (ii), consistently using the weighted-L_1 penalty yields better results than using the L_1-penalty when the same loss function is employed. Methods (iii) (Zhao_2012, bin = 0.5, 0.25) and (iv) (SIE-GLM, bin = 0.5, 0.25) exhibit relatively reduced sensitivity performance compared to other methods, with Correct_All beling less than 38 out of 60. As for method (v) (Raj_2005, parent = 2), to reduce the computational cost of searching all possible subsets of parents for each node, only subsets with a maximum size of 2 are considered. Since the true network is sparse with a degree no greater than 2 for each node, this setting is most favorable for method (v). Nevertheless, method (v) only performs well in terms of sensitivity and significantly underperforms in terms of specificity compared to other methods. In summary, our proposed continuous-time method (Continuous_wL1) with the weighted-L_1 penalty demonstrates the best overall performance across $\beta \in \{0.4, 0.5\}$.

We next present Table 3 to compare the results using different values of the total time length T $\in \{1000, 2000\}$. It is evident that T = 2000 ourperforms T = 1000 for all methods. This aligns with expectations since larger datasets provide more information and lead to more accurate estimations. This finding is also consistent with our theoretical result of network recovery consistency stated in Corollary 1 of Section 5.3, which indicates that the detected network becomes closer to the true network as the time length T increases.

Under each T setting, the pattern of results is similar to that in Table 2. Continuous_wL1 maintains the best overall performance.

To investigate the robustness of our method to misspecified time-lags for the true time-lag ϕ (equal to 1), we use specified time-lag $\phi_a \in \{0.5, 1, 1.5\}$ in the estimation procedure. The results are provided in Table 4. As anticipated, $\phi_a = 1$ exhibits the best performance. The misspecified ϕ_a values of $\{0.5, 1.5\}$ do not significantly impact specificity but do reduce sensitivity for most of the listed methods. Among all the listed methods, the continuous-time methods (Continuous_L1 and Continuous_wL1) still demonstrate the best overall performance. Specifically, the sensitivity of Continuous_L1 decreases by less than 15% under both of both misspecified time-lags, which supports the robustness of our method to some extent against this misspecification.

To assess the robustness of our methods against misspecified intensity models for $\lambda_i(t \mid \mathscr{F}_t)$, we conducted a separate simulation study on data generated from the non-linear Hawkes model, with the true intensity function:

$$\lambda_i^*(t \mid \mathscr{F}_t) \;=\; \exp\Big\{\beta_{0;i}^* + \sum_{j\in\mathcal{V}} \int_{-\infty}^{t} \beta_{j,i}^* \,\mathrm{I}(0 \le t-u < 1)\,\mathrm{d}N_j(u)\Big\}, \quad i \in \mathcal{V}. \tag{6.2}$$

In this model, we set $\beta_{0;i}^* = -0.8$, $\beta_{j,i}^* = 3$ for the excitatory effect, $\beta_{j,i}^* = -3$ for the inhibitory effect, and $\beta_{j,i}^* = 0$ for no effect from node j to node i. The results in Table 5 indicate that the performances of each method largely agrees with the results reflected in Tables 2, 3, and 4. Our proposed method, Continuous_wL1, continues to exhibit the best overall performance. In summary, this simulation result demonstrates the robustness of our estimation method against model misspecification. Our method performs well even when the shape-function g is unbounded in the true model. This indicates that our estimation method and theoretical results are applicable to a broader range of models beyond the non-linear Hawkes process.

6.3.2 Simple and medium-complex networks

For Network-1 and Network-2, we conducted the same simulation evaluation as for Network-3. The results of the two networks, comparing connection strength, time length and time-lag width, resemble those obtained for Network-3 and have been omitted for brevity. Among all the methods, Continuous_wL1 consistently exhibits the best overall performance in each setting. This finding indicates that our conclusions are consistent across different types of

true networks.

In summary, all of these simulation results confirm the superiority of our proposed continuous-time method over the other methods, regardless of the complexity level of the true network structure.

7 Real data analysis

In this section, we apply our method to real-world multivariate point process data. We analyze the prefrontal cortex spike train dataset pfc-6 on CRCNS, available at `https://crcns.org/data-sets/pfc/pfc-6/about-pfc-6`. This dataset comprises neuronal ensemble recordings from the medial prefrontal cortex, primarily the prelimbic cortex, of freely moving rats using tetrodes. The data were collected during the rats' performance of a behavioral contingency task, as well as during sleep before and after the task. This dataset consists of 90 sessions, each corresponding to an experiment. For our real data experiment, we select the session folder 181020. In the chosen session, we have spike train data from 55 neurons recorded over a period of 6500 seconds. This data is stored in the file '181020_SpikeData.dat,', containing a total of $1,309,619$ spikes from the 55 neurons.

We apply our continuous-time modeling method, Continuous_wL1, to this dataset, with the tuning parameter selected using the BIC. Similar to the simulation studies, our estimation procedure adopts the empirical choices of $\phi = 1$ and $g(x) = \log(1 + x \wedge c)$, where the data-driven choice for c is given in (A.2). Previous studies in neuroscience [7, 40] have indicated that a neuron's spiking activity may influence other neurons primarily within a short period, often less than 1 second, known as the refractory-recovery period. Taking this into account, we empirically choose $\phi = 1$ to capture short-term interactions among neurons while considering the refractory-recovery period. The estimated network structure is presented in Figure 5 (left panel). We identify a total of 579 connections, including 352 excitatory effects and 227 inhibitory effects. Several interesting findings emerge from this study. For instance, pairs of neurons $\{6, 7\}$, $\{24, 34\}$, $\{38, 42\}$, $\{25, 27\}$, $\{21, 23\}$ demonstrate strong mutual excitatory effects, suggesting close functional connectivity and similarity within these pairs. Neuron 13 exhibits 34 excitatory effects on other neurons, which is significantly higher than any other neuron, while it does not impose any inhibitory effect. This suggests that neuron 13 may potentially serve as a hub neuron, playing a cru-

cial role in triggering the activities of the entire neuron ensemble. To compare with BIC, we also incorporate the Generalized Information Criterion (GIC) [32] with a penalty perm $a_{\mathrm{T}} = V \log(\mathrm{T})$ to select the tuning parameter. The resulting network, shown in Figure 5 (right panel), is sparser and includes a number of isolated neurons that are disconnected with others. In this regard, GIC fails to capture all potential interactions compared to BIC in our experiment. It is important to note that the recovered connections, obtained through either the BIC or GIC method, represent the estimated statistical dependencies between neurons. However, these estimated connections do not necessarily imply the existence of real neuronal connections in the brain. Nevertheless, our results are valuable in assisting further neurological research.

We also apply two other modeling methods to this dataset: (a) the Discrete_wL1 method with BIC criterion and a bin size of $= 0.25$; and (b) the linear Hawkes process (HK) modeling method [42]. The estimated networks are displayed in Figure 6. The Discrete_wL1 method identifies a total of 479 effects, among which, 448 are also detected by the Continuous_wL1 method with BIC. This indicates a significant overlap between the estimated network obtained by Discrete_wL1 and Continuous_wL1 methods, which is expected since the Discrete_wL1 method approximates the Continuous_wL1 method when the bin size is sufficiently small. However, we believe that the Continuous_wL1 method is more accurate than the Discrete_wL1 method as long as the real physical intensity of neuronal spikes evolves in continuous-time, as supported by our simulation results in Section 6. Regarding the HK method, it only detected 353 excitatory effects and no inhibitory effects. This limitation arises from the nature of the linear Hawkes process, which is inherently self-exciting and does not allow negative parameterizations in its kernel function. In contrast, our Continuous_wL1 method is capable of detecting both excitatory and inhibitory interactions between neurons, providing a more comprehensive estimation of the potential network of functional connectivities among this group of neurons.

8 Discussion

Motivated by the crucial task of inferring neural connectivity from ensemble neural spike train data in neuroscience research, this paper aims to uncover the network-structured dependence underlying a class of non-stationary multivariate point process models. To achieve this goal, we propose a novel continuous-time stochastic model for the intensity

processes. We formulate the associated theoretical framework and derive probabilistic properties that are essential for learning the statistical properties of the proposed penalized M-estimator for graph parameters. These parameters are crucial for identifying the causal relationships among nodes in the network. In our approach, we develop new technical tools, including the marked point process with explicit conditional distributions, recurrence time points, and cyclicity property. These tools prove instrumental in analyzing the probabilistic properties of a wide range of continuous-time models for point processes. Furthermore, they play a central role in the statistical learning of network structure.

Our proposed framework extends beyond the learning of interaction effects among nodes. It has the flexibility to incorporate other factors, such as autoregressive effects, experimental units, and other extrinsic conditions, into model (3.1). Furthermore, acknowledging potential variations in interaction time-lags among nodes, we can also allow the lag-width ϕ in the covariate $x_j(t)$, as illustrated in (3.5)–(3.6), to vary according to the node j. While these additional extensions hold potential for enhancing our understanding of complex systems, a comprehensive exploration of these aspects is beyond the scope of this paper. However, investigating these factors in future research would be valuable and could provide further insights into the dynamics of network structures.

References

[1] Aalen, O. O. (1987). Dynamic modelling and causality. *Scandinavian Actuarial Journal*, 1987(3-4), 177-190.

[2] Azais, R., Bardet, J. B., Génadot, A., Krell, N, and Zitt, P. A. (2014). Piecewise deterministic Markov process — recent results. *ESAIM: Proceedings*, vol. 44, 276–290.

[3] Andersen P. K., Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11, 91–115.

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

[5] Bremaud, P. and Massoulie, L. (1996). Stability of nonlinear hawkes processes. *Annals of Probability*, 24, 1563–1588.

[6] Brillinger, D. R. and Villa, A. E. P. (1994). Examples of the investigation of neural information processing by point process analysis. *Advanced Methods of Physiological System Modelling*, 3, 111–127.

[7] Brown, E. N., Kass, R. E. and Mitra, P. P. (2004). Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 7, 456–461.

[8] Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology*, 38 (1), 155–200.

[9] Chen, S., Witten, D., and Shojaie, A. (2017). Nearly assumptionless screening for the mutually-exciting multivariate Hawkes process. *Electronic journal of statistics*, 11(1), 1207.

[10] Costa, M., Graham, C., Marsalle, L., and Tran, V. C. (2020). Renewal in Hawkes processes with self-excitation and inhibition. *Advances in Applied Probability*, 52(3), 879–915.

[11] Daley, D. J. and Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes, Volume II: General Theory and Structure*, 2nd ed., Springer, New York.

[12] Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, 5:142–150.

[13] Dempsey, W., Oselio, B., Hero, A. (2021). Hierarchical network models for exchangeable structured interaction processes. *Journal of the American Statistical Association*, Vol. 00, No. 0, pp 1–18.

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

[15] Dolivo, F. B. (1974). Counting processes and integrated conditional rates: a martingale approach with application to detection. PhD thesis, University of Michigan.

[16] Embrechts, P., and Kirchner, M. (2018). Hawkes graphs. *Theory of Probability & Its Applications*, 62(1), 132-156.

[17] Fischer, T. (2013). On simple representations of stopping times and stopping time sigma-algebras. *Statistics and Probability Letters*, 83, 345–349.

[18] Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, 1, 302–332.

[19] Gerstein, G. L. and Perkel, D. H. (1969). Simultaneously recorded trains of action potentials: analysis and functional interpretation. *Science*, 164, 828–830.

[20] Hansen, N., Reynaud-Bouret, P. and Rivoirard, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21, 83–143.

[21] Harris, K. D., Csicsvari, J., Hirase, H., Dragoi, G. and Buzsaki, G. (2003). Organization of cell assemblies in the hippocampus. *Nature*, 424, 552–556.

[22] Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11, 493–503.

[23] Kalashnikov, V. V. (1993). *Mathematical Methods in Queuing Theory*, Netherlands: Springer Netherlands.

[24] Kass, R. E., Brown, E. N. and Eden, U. (2014). *Analysis of Neural Data*, Springer, New York.

[25] Krumin, M., Reutsky, I. and Shoham, S. (2010). Correlation-based analysis and generation of multiple spike trains using Hawkes models with an exogenous input. *Frontiers in Computational Neuroscience*, 4, 147.

[26] Lukasik, M., Srijith, P. K., Cohn, T., and Bontcheva, K. (2015). Modeling tweet arrival times using log-Gaussian Cox processes. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 250–255.

[27] Meyn, S., and Tweedie, R. (1993). *Markov Chains and Stochastic Stability*, Springer Verlag, New York.

[28] Murphy, K. P. (2001). Dynamic Bayesian networks: representation, inference and learning. PhD thesis, UC Berkeley, Computer Science Division.

[29] Neuts, M. F., (1979). A versatile Markovian point process. *Journal of Applied Probability*, 16(4), pp.764–779.

[30] Nicolis, O., Riquelme Quezada, L. M., and Ibacache-Pulgar, G. (2022). Temporal Cox process with folded normal intensity. *Axioms*, 11(10), 513.

[31] Nieuwenhuis, G. (2013). Asymptotic mean stationarity and absolute continuity of point process distributions. *Bernoulli*, 19, 1612–1636.

[32] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, 12, 758–765.

[33] Perkel, D. H., Gerstein, G. L. and Moore, G. P. (1967). Neuronal spike trains and stochastic point processes: II. Simultaneous spike trains. *Biophysical Journal*, 7, 419–440.

[34] Perry, P. O. and Wolfe, P.J. (2013). Point process modelling for directed interaction networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol*, 75(5), 821-849.

[35] Rajaram, S., Graepel, T. and Herbrich, R. (2005). Poisson-networks: a model for structured point processes. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics* (*AISTAT*).

[36] Reynaud-Bouret, P. and Schbath, S. (2010). Adaptive estiamtion for Hawkes process; application to genome analysis. *Annals of Statistics*, 38, 2781–2822.

[37] Ross, S. M. (1996). *Stochastic Processes*. John Wiley & Sons.

[38] Rubin, I. (1972). Regular point process and their detection. *IEEE Transactions on Information Theory*, 18, 547–557.

[39] Tang, X. and Li, L. (2021). Multivariate temporal point process regression. *Journal of the American Statistical Association*, pp.1-16.

[40] Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P. and Brown, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, 93, 1074–1089.

[41] Van der Vaart, A. W. (2000). *Asymptotic Statistics*, Vol. 3, Cambridge university press.

[42] Xu, H., Farajtabar, M. and Zha, H. (2016). Learning granger causality for hawkes processes. In *Proceedings of The 33rd International Conference on Machine Learning* (*ICML*), pp. 1717–1726.

[43] Yuan, B., Li, H., Bertozzi, A. L., Brantingham, P. J., and Porter, M. A. (2019). Multivariate spatiotemporal Hawkes processes and network reconstruction. *SIAM Journal on Mathematics of Data Science*, 1 (2), 356–382.

[44] Zhang, C. M., Chai, Y., Guo, X., Gao, M., Devilbiss, D. M. and Zhang, Z. (2016). Statistical learning of neuronal functional connectivity. *Technometrics*, 58, 350–359.

[45] Zhang, C. M., Jiang, Y. and Chai, Y. (2010). Penalized Bregman divergence for large dimensional regression and classification. *Biometrika*, 97, 551–566.

[46] Zhao, M., Batista, A., Cunningham, J. P., Chestek, C., Rivera-Alvidrez, Z., Kalmar, R., Ryu, S., Shenoy, K. and Iyengar, S. (2012). An L_1-regularized logistic model for detecting short-term neuronal interactions. *Journal of Computational Neuroscience*, 32, 479–497.

[47] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

Appendices

A Practical issue; figures and tables in the paper

A.1 Practical issues on selecting ϕ and $g(\cdot)$ in (3.5) and (3.6)

In practice, there are various ways to choose the time-lag ϕ and shape-function $g(\cdot)$, via either prior knowledge or data-driven methods. Below, we provide some suggestions.

Selection of ϕ. The time-lag ϕ can be chosen in line with the number n of time bins with the bin-width T/n. Alternatively, our empirical choices fix $\phi = 1$ (a unit of time, also used in [35]). Besides, ϕ could also be selected by a data-driven algorithm as outlined below.

(1). Choose a sufficiently large $\phi_{\max}$, guided by domain knowledge or prior information.

(2). Iterate for each $k = 0, 1, 2, \ldots, k_{\max}$ using $\phi_k = \phi_{\max}\zeta^k$, where $\zeta \in (0, 1)$ is a step length, and $k_{\max} > 1$ is the maximum iteration number. Obtain penalized M-estimators $\{\widehat{\widetilde{\boldsymbol{\beta}}}_i^{(\phi_k)}\}_{i\in\mathcal{V}}$ by minimizing (5.8) with ϕ replaced by ϕ_k.

(3). Compute the joint negative log-likelihood $L_k = \sum_{i\in\mathcal{V}} \boldsymbol{\mathcal{L}}_{i,\mathrm{T}}^{(k)}(\widehat{\widetilde{\boldsymbol{\beta}}}_i^{(\phi_k)})$, where $\boldsymbol{\mathcal{L}}_{i,\mathrm{T}}^{(k)}(\cdot)$ resembles $\boldsymbol{\mathcal{L}}_{i,\mathrm{T}}(\cdot)$ in (5.4), replacing ϕ with ϕ_k.

(4). For $k = 0, 1, \ldots, k_{\max} - 1$, identify the first $k = \widehat{k}$ where $L_{\widehat{k}+1} > L_{\widehat{k}}$. Terminate the algorithm upon finding $\widehat{k}$. If $\widehat{k}$ is nonexistent, let $\widehat{k} = k_{\max}$. Our selected lag-width is $\phi_{\widehat{k}}$, and the corresponding estimators are $\{\widehat{\widetilde{\boldsymbol{\beta}}}_i^{(\phi_{\widehat{k}})}\}_{i\in\mathcal{V}}$.

The primary strategy of the aforementioned algorithm is to backtrack and locate ϕ, aiming to achieve the highest likelihood value among all ϕ's for the corresponding estimators $\{\widehat{\widetilde{\boldsymbol{\beta}}}_i^{(\phi)}\}_{i\in\mathcal{V}}$. Table 6 presents simulation results on Network 1 employing data-driven $\phi_{\widehat{k}}$ for estimation. It's evident that our methods Continuous_L1 and Continuous_wL1 consistently outperform other methods, even without precise knowledge of the true time-lag ϕ. Additionally, a simulation scenario has been added where both ϕ and $g(\cdot)$ are misspecified; the results are presented in Table 8. These results demonstrate that our proposed methods Continuous_L1 and Continuous_wL1 consistently outperform other approaches, showcasing a certain level of robustness against misspecified ϕ and $g(\cdot)$.

Selection of $g(\cdot)$. We specify $g(\cdot)$ to be some bounded functions, e.g.,

$$g(x) = \log(1 + x \wedge c), \quad \text{or} \quad g(x) = x \wedge c, \tag{A.1}$$

for some constant $c \in (0, \infty)$; for practical applications, we suggest the data-driven choice of c by

$$c = \text{the 90th percentile of } \Big\{ \max_{t\in[0,\mathrm{T}]} \{N_j((t-\phi, t])/\phi\} : j \in \mathcal{V} \Big\}, \tag{A.2}$$

which ensures that the covariates $\{x_j(t)\}_{j\in\mathcal{V};t\in[0,\mathrm{T}]}$ embrace the empirical rates as closely as possible with guaranteed numerical stability, without increasing the computational costs.

Regarding practical applications, we've included a simulation scenario in Table 7 using the unbounded function $g(x) = \log(1+x)$ for both generating synthetic data and estimating model parameters. These results demonstrate that even with an unbounded g, the proposed network modeling and recovery method can still be effective.

A.2 Explicit procedure for implementing BIC criterion in simulation

In Method (i) of simulation, we choose the tuning parameter η (or η_{T}) by minimizing the BIC function

$$\mathrm{BIC}(\widehat{\widetilde{\boldsymbol{\beta}}}_i) = 2\mathcal{L}_{i,\mathrm{T}}(\widehat{\widetilde{\boldsymbol{\beta}}}_i) + \mathrm{df}(\widehat{\boldsymbol{\beta}}_i) \cdot \log(\mathrm{T})/\mathrm{T}, \tag{A.3}$$

where $\mathcal{L}_{i,\mathrm{T}}(\cdot)$ is defined in (5.4), and $\mathrm{df}(\widehat{\boldsymbol{\beta}}_i) = \sum_{j\in\mathcal{V},j\neq i} \mathrm{I}(\widehat{\beta}_{j,i} \neq 0)$ is the number of non-zero elements in $\widehat{\boldsymbol{\beta}}_i$. Here, $\mathrm{BIC}(\widehat{\widetilde{\boldsymbol{\beta}}}_i)$ is seen as a function of η, since $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ depends on η, i.e., $\widehat{\widetilde{\boldsymbol{\beta}}}_i = \widehat{\widetilde{\boldsymbol{\beta}}}_i^{(\eta)}$. The minimizer η of $\mathrm{BIC}(\widehat{\widetilde{\boldsymbol{\beta}}}_i)$ is found by searching from the set of grid points $\{\eta_{\max} h^k : k = 0, 1, \ldots, 12\}$, where $\eta_{\max} = \sup\{\eta : \mathrm{df}(\widehat{\widetilde{\boldsymbol{\beta}}}_i^{(\eta)}) > 0\}$ and $h \in (0, 1)$ is a constant. Specifically, we use $h = 0.7$ in all our numerical experiments.

A.3 Figures and tables

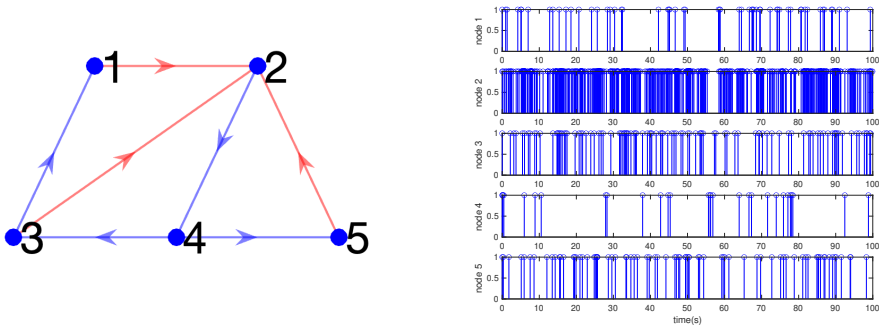

Figure 1: *Each node of the network graph in the left panel corresponds to a point process in the right panel. Arrows indicate interactions (red for excitatory and blue for inhibitory effects).*

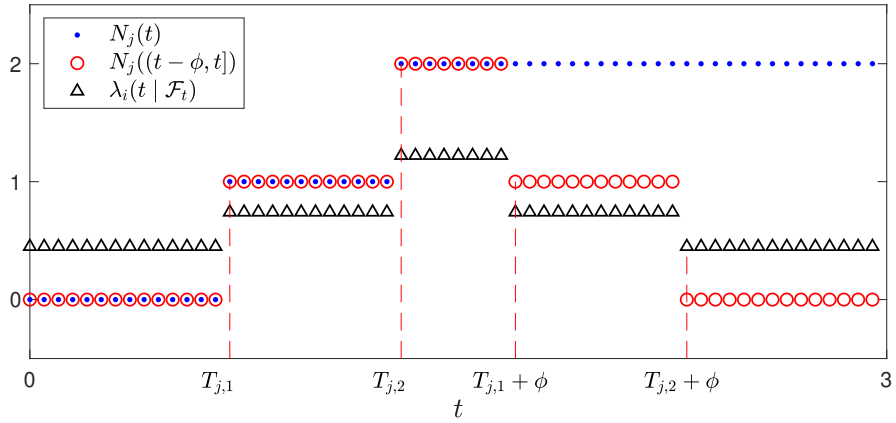

Figure 2: *Illustrative plot showing sample paths of stochastic processes $N_j(t)$ in (2.2), $N_j((t-\phi,t])$ in (4.1), and $\lambda_i(t \mid \mathscr{F}_t) = \exp\{-0.8 + 0.5 \cdot N_j((t-\phi,t])\}$ in (3.1), with $\mathcal{V} = \{1, 2\}$, $i = 1$, $j = 2$, and time-lag $\phi = 1$. Notice the overlap between $N_j(t)$ and $N_j((t-\phi,t])$ within the time interval $[0, 1.7)$. $\lambda_i(t \mid \mathscr{F}_t)$ is a piecewise-constant function with discontinuities identical to those of $N_j((t-\phi,t])$.*

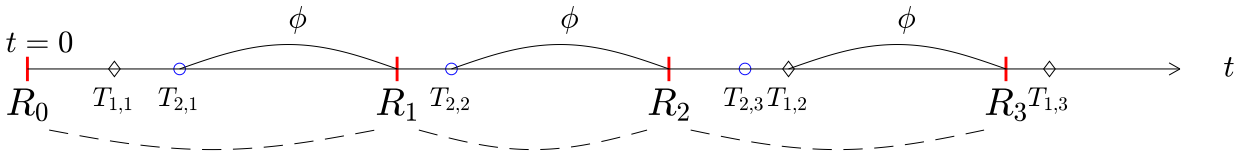

Figure 3: *Illustrative plot depicting recurrence time points $R_0, R_1, \ldots$, and event time points $\{T_{i,\ell}\}_{\ell \geq 1}$ of nodes $i \in \mathcal{V}$, where $\mathcal{V} = \{1, 2\}$. The cyclicity property in Theorem 3 denote that after reaching each recurrence time point R_ℓ, $\boldsymbol{N}(t)$ enters a recurrence cycle $(R_\ell, R_{\ell+1}]$. Within this cycle, $\lambda_i(R_\ell \mid \mathcal{F}_{R_\ell}) = \lambda_i(0)$, initiating a renewed process $\{\boldsymbol{N}(t + R_\ell) - \boldsymbol{N}(R_\ell)\}_{t \geq 0}$, independent of $\mathcal{F}_{R_\ell}$.*

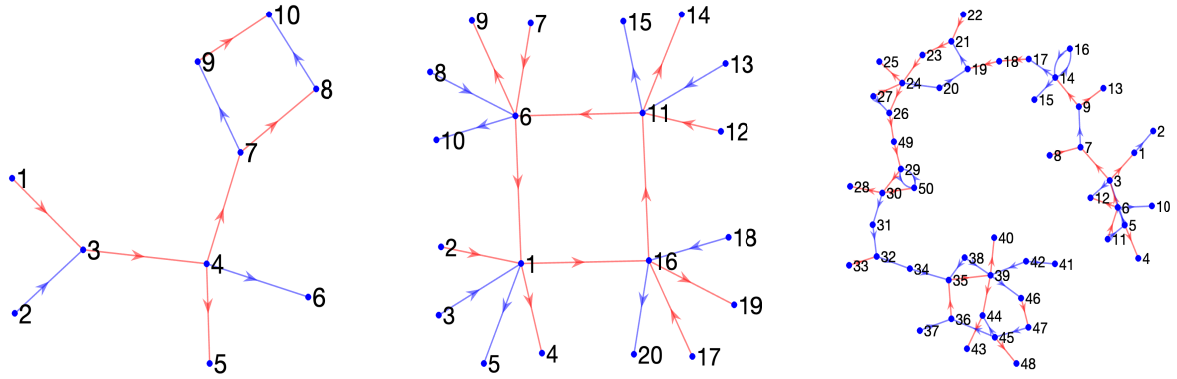

Figure 4: **(Simulation Study: True** Network-1**,** Network-2**, and** Network-3**)** *The left panel:* Network-1*, a simple network with* 10 *nodes; the middle panel:* Network-2*, a medium-complexity network with* 20 *nodes; and the right panel:* Network-3*, a complex network with* 50 *nodes. Red arrows indicate excitatory effects, while blue arrows indicate inhibitory effects.*

Table 1: *Method Descriptions for Simulation Studies.*

abbreviation of method		description
Discrete_L1	bin=0.5	method (ii) with L_1-penalty, and bin width = 0.5.
	bin=0.25	method (ii) with L_1-penalty, and bin width = 0.25.
	bin=0.1	method (ii) with L_1-penalty, and bin width = 0.1.
Continuous_L1		method (i) with L_1-penalty.
Discrete_wL1	bin=0.5	method (ii) with weighted-L_1 penalty, and bin width = 0.5.
	bin=0.25	method (ii) with weighted-L_1 penalty, and bin width = 0.25.
	bin=0.1	method (ii) with weighted-L_1 penalty and bin width = 0.1.
Continuous_wL1		method (i) with weighted-L_1 penalty.
Zhao_2012	bin=0.5	method (iii) with bin width = 0.5.
	bin=0.25	method (iii) with bin width = 0.25.
	bin=0.1	method (iii) with bin width = 0.1.
SIE-GLM	bin=0.5	method (iv) with bin width = 0.5.
	bin=0.25	method (iv) with bin width = 0.25.
	bin=0.1	method (iv) with bin width = 0.1.
Raj_2005	parent=3	method (v) with maximum parent number = 3.
	parent=2	method (v) with maximum parent number = 2.

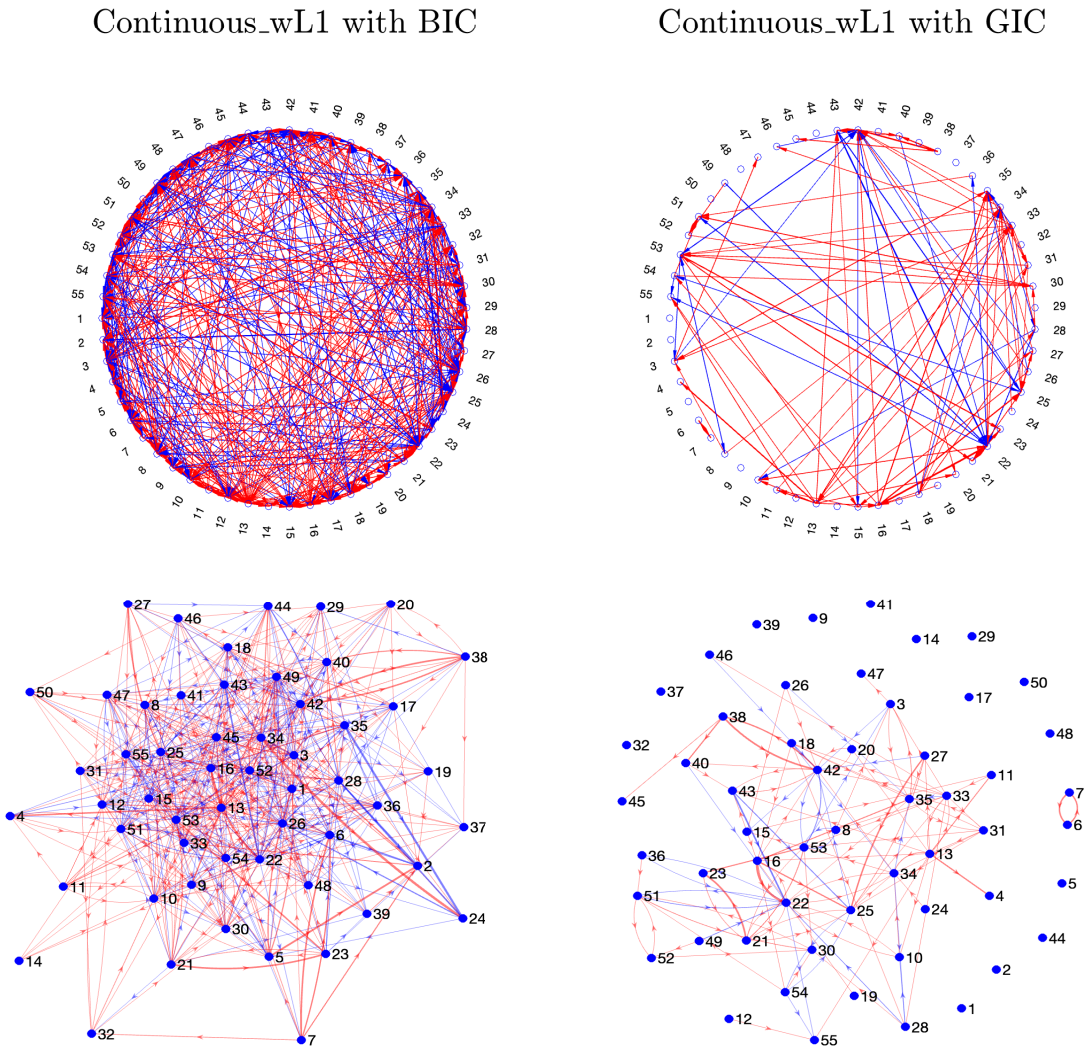

Figure 5: **(Real Data: Estimated Networks Using the Continuous-Time Modeling Method with Weighted-L_1 Penalty)** *Red arrows denote excitatory effects, while blue arrows indicate inhibitory effects. Thicker arrows represent stronger interactions. Top-left panel:* BIC *criterion (in circular layout); bottom-left panel:* BIC *criterion (in equilibrium layout); top-right panel:* GIC *criterion (in circular layout); bottom-right panel:* GIC *criterion (in equilibrium layout).*

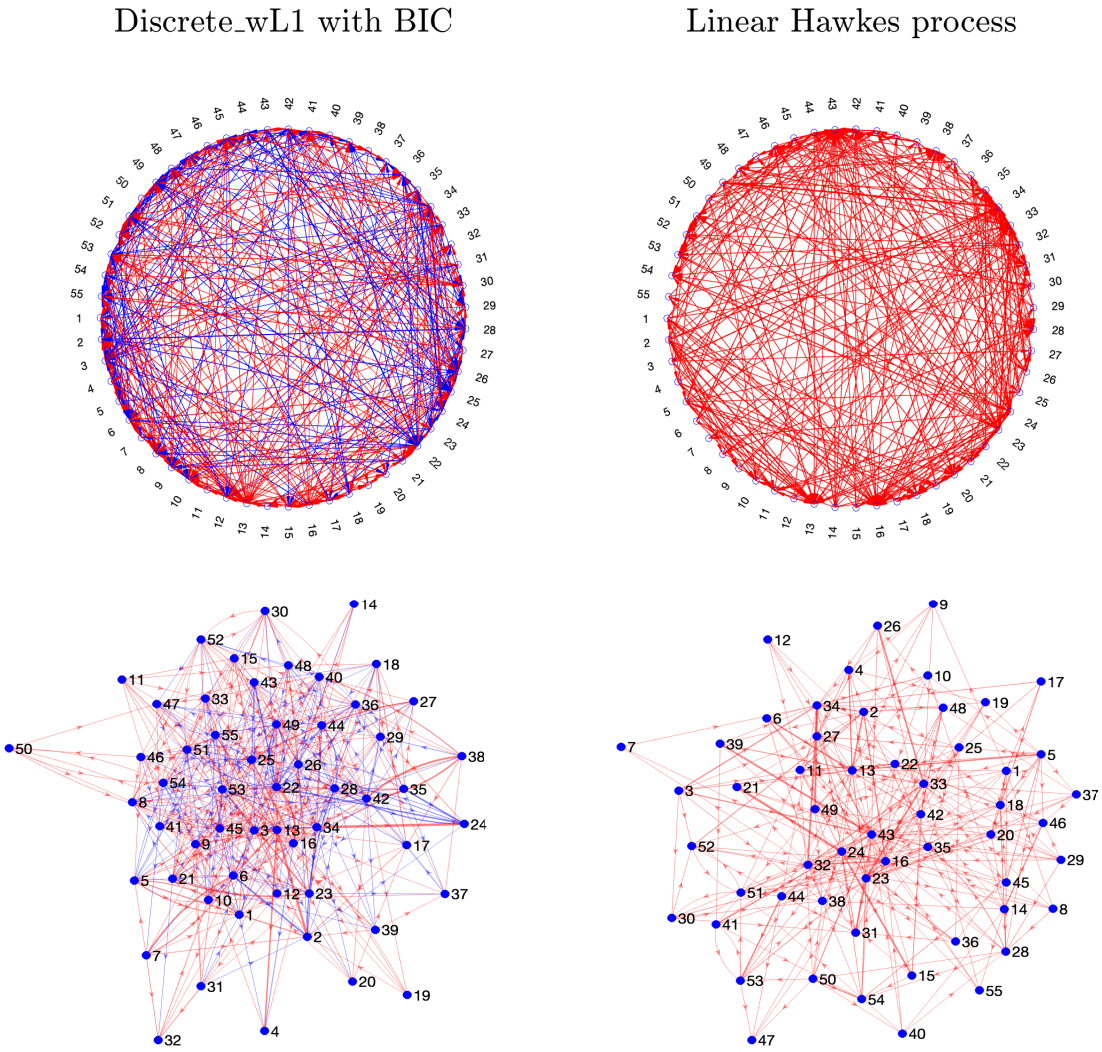

Figure 6: **(Real Data: Estimated Networks Using Discrete-Time Modeling and Linear Hawkes Process Modeling Methods)** *Red arrows represent excitatory effects, while blue arrows signify inhibitory effects; thicker arrows indicate stronger interactions. Top-left panel: Discrete-time modeling (plot in circular layout); bottom-left panel: Discrete-time modeling (plot in equilibrium layout); top-right panel: Linear Hawkes process (plot in circular layout); bottom-right panel: Linear Hawkes process (plot in equilibrium layout).*

Table 2: **(Simulation Study:** Network-3 **with Connection Strength** $\beta \in \{0.4, 0.5\}$**)** *The time length is* $\mathrm{T} = 2000$*. Results are averaged over* 100 *replications, with standard errors denoted in parentheses.*

		Correct_All		Detected_A		Detected_B		Correct_NC	
strength $\beta =$		0.4	0.5	0.4	0.5	0.4	0.5	0.4	0.5
Discrete_L1	bin=0.5	21.18 (0.42)	42.55 (0.33)	13.71 (0.26)	24.92 (0.17)	7.47 (0.25)	17.63 (0.26)	2385.87 (0.26)	2379.92 (0.33)
	bin=0.25	36.21 (0.45)	53.39 (0.25)	21.29 (0.25)	28.83 (0.10)	14.92 (0.30)	24.56 (0.21)	2382.13 (0.33)	2376.73 (0.42)
	bin=0.1	44.04 (0.41)	56.95 (0.17)	24.76 (0.18)	29.67 (0.05)	19.28 (0.29)	27.28 (0.15)	2379.40 (0.38)	2375.46 (0.38)
Continuous_L1		51.86 (0.33)	58.44 (0.14)	27.33 (0.15)	29.87 (0.03)	24.53 (0.26)	28.57 (0.12)	2368.23 (0.60)	2367.52 (0.43)
Discrete_wL1	bin=0.5	40.02 (0.39)	54.05 (0.25)	22.69 (0.22)	28.86 (0.11)	17.32 (0.29)	25.19 (0.20)	2380.52 (0.39)	2379.92 (0.36)
	bin=0.25	50.60 (0.28)	58.55 (0.11)	27.05 (0.17)	29.83 (0.03)	23.55 (0.21)	28.72 (0.09)	2380.32 (0.39)	2381.44 (0.34)
	bin=0.1	54.81 (0.22)	59.44 (0.07)	28.51 (0.12)	29.97 (0.01)	26.30 (0.18)	29.47 (0.06)	2380.07 (0.36)	2382.73 (0.30)
Continuous_wL1		56.80 (0.21)	59.72 (0.05)	29.08 (0.11)	30.00 (0.00)	27.72 (0.17)	29.72 (0.05)	2380.71 (0.34)	2383.44 (0.25)
Zhao_2012	bin=0.5	10.08 (0.29)	26.42 (0.37)	6.36 (0.20)	16.46 (0.24)	3.72 (0.16)	9.96 (0.19)	2388.66 (0.13)	2385.15 (0.29)
	bin=0.25	3.62 (0.17)	9.41 (0.26)	3.28 (0.16)	8.08 (0.22)	0.34 (0.06)	1.33 (0.11)	2389.66 (0.05)	2388.86 (0.14)
SIE-GLM	bin=0.5	19.88 (0.40)	39.28 (0.33)	14.00 (0.26)	24.54 (0.18)	5.88 (0.22)	14.74 (0.25)	2387.55 (0.18)	2384.09 (0.28)
	bin=0.25	17.43 (0.35)	37.24 (0.31)	13.29 (0.23)	24.28 (0.20)	4.13 (0.20)	12.96 (0.22)	2388.23 (0.13)	2384.84 (0.25)
Raj_2005	parent=2	57.34 (0.08)	57.91 (0.02)	29.37 (0.07)	29.45 (0.06)	27.97 (0.09)	28.46 (0.06)	2347.34 (0.08)	2347.91 (0.02)
true		60		30		30		2390	

Table 3: **(Simulation Study:** Network-3 **with Time Length** $\mathrm{T} \in \{1000, 2000\}$**)** *The connection strength is* $\beta = 0.5$*. Results are averaged over* 100 *replications, with standard errors denoted in parentheses.*

		Correct_All		Detected_A		Detected_B		Correct_NC	
time length T =		1000	2000	1000	2000	1000	2000	1000	2000
Discrete_L1	bin=0.5	16.75 (0.39)	42.55 (0.33)	11.74 (0.26)	24.92 (0.17)	5.01 (0.22)	17.63 (0.26)	2385.84 (0.23)	2379.92 (0.33)
	bin=0.25	28.24 (0.40)	53.39 (0.25)	18.19 (0.24)	28.83 (0.10)	10.05 (0.25)	24.56 (0.21)	2383.26 (0.30)	2376.73 (0.42)
	bin=0.1	35.75 (0.41)	56.95 (0.17)	21.91 (0.24)	29.67 (0.05)	13.84 (0.26)	27.28 (0.15)	2379.78 (0.34)	2375.46 (0.38)
Continuous_L1		48.79 (0.33)	58.44 (0.14)	27.84 (0.12)	29.87 (0.03)	20.95 (0.28)	28.57 (0.12)	2359.78 (0.81)	2367.52 (0.43)
Discrete_wL1	bin=0.5	33.18 (0.45)	54.05 (0.25)	19.69 (0.26)	28.86 (0.11)	13.49 (0.30)	25.19 (0.20)	2376.73 (0.42)	2379.92 (0.36)
	bin=0.25	44.57 (0.34)	58.55 (0.11)	25.04 (0.19)	29.83 (0.03)	19.53 (0.26)	28.72 (0.09)	2377.19 (0.42)	2381.44 (0.34)
	bin=0.1	49.88 (0.30)	59.44 (0.07)	27.24 (0.15)	29.97 (0.01)	22.64 (0.25)	29.47 (0.06)	2376.29 (0.41)	2382.73 (0.30)
Continuous_wL1		52.61 (0.26)	59.72 (0.05)	28.06 (0.12)	30.00 (0.00)	24.55 (0.21)	29.72 (0.05)	2375.96 (0.42)	2383.44 (0.25)
Zhao_2012	bin=0.5	6.15 (0.25)	26.42 (0.37)	4.33 (0.20)	16.46 (0.24)	1.82 (0.13)	9.96 (0.19)	2388.78 (0.12)	2385.15 (0.29)
	bin=0.25	2.84 (0.17)	9.41 (0.26)	2.60 (0.16)	8.08 (0.22)	0.24 (0.04)	1.33 (0.11)	2389.59 (0.08)	2388.86 (0.14)
SIE-GLM	bin=0.5	15.07 (0.36)	39.28 (0.33)	11.77 (0.27)	24.54 (0.18)	3.30 (0.18)	14.74 (0.25)	2387.59 (0.18)	2384.09 (0.28)
	bin=0.25	12.24 (0.36)	37.24 (0.31)	10.30 (0.29)	24.28 (0.20)	1.94 (0.15)	12.96 (0.22)	2388.46 (0.12)	2384.84 (0.25)
Raj_2005	parent=2	55.44 (0.16)	57.91 (0.02)	28.96 (0.09)	29.45 (0.06)	26.48 (0.13)	28.46 (0.06)	2345.44 (0.16)	2347.91 (0.02)
true		60		30		30		2390	

Table 4: **(Simulation Study:** Network-3 **Parameter Estimation Using Specified Time-Lags** $\phi_a \in \{0.5, 1, 1.5\}$ **for the True Time-Lag** $\phi = 1$**)** *The connection strength is* $\beta = 0.5$*, and the time length is* $\mathrm{T} = 2000$*. Results are averaged over* 100 *replications, with standard errors denoted in parentheses.*

		Correct_All			Detected_A			Detected_B			Correct_NC		
time-lag ϕ_a =		0.5	1	1.5	0.5	1	1.5	0.5	1	1.5	0.5	1	1.5
Discrete_L1	bin=0.5	37.10 (0.36)	42.55 (0.33)	23.29 (0.45)	22.57 (0.19)	24.92 (0.17)	15.42 (0.28)	14.53 (0.25)	17.63 (0.26)	7.87 (0.27)	2382.26 (0.31)	2379.92 (0.33)	2384.75 (0.25)
	bin=0.25	37.11 (0.42)	53.39 (0.25)	37.15 (0.43)	22.70 (0.25)	28.83 (0.10)	22.39 (0.25)	14.41 (0.25)	24.56 (0.21)	14.76 (0.27)	2381.75 (0.30)	2376.73 (0.42)	2381.04 (0.32)
	bin=0.1	36.77 (0.42)	56.95 (0.17)	44.89 (0.38)	22.26 (0.21)	29.67 (0.05)	25.88 (0.19)	14.51 (0.29)	27.28 (0.15)	19.01 (0.28)	2382.06 (0.29)	2375.46 (0.38)	2379.39 (0.34)
Continuous_L1		43.90 (0.32)	58.44 (0.14)	54.26 (0.26)	25.70 (0.17)	29.87 (0.03)	28.06 (0.14)	18.20 (0.25)	28.57 (0.12)	26.20 (0.19)	2362.46 (0.60)	2367.52 (0.43)	2369.17 (0.60)
Discrete_wL1	bin=0.5	51.83 (0.25)	54.05 (0.25)	40.63 (0.36)	27.70 (0.13)	28.86 (0.11)	23.83 (0.21)	24.13 (0.19)	25.19 (0.20)	16.80 (0.27)	2380.07 (0.42)	2379.92 (0.36)	2380.05 (0.36)
	bin=0.25	51.52 (0.29)	58.55 (0.11)	50.86 (0.28)	27.60 (0.15)	29.83 (0.03)	27.85 (0.11)	23.92 (0.24)	28.72 (0.09)	23.01 (0.23)	2379.61 (0.34)	2381.44 (0.34)	2380.12 (0.40)
	bin=0.1	51.44 (0.28)	59.44 (0.07)	54.72 (0.23)	27.56 (0.13)	29.97 (0.01)	29.10 (0.08)	23.88 (0.22)	29.47 (0.06)	25.62 (0.20)	2380.21 (0.34)	2382.73 (0.30)	2380.28 (0.35)
Continuous_wL1		51.05 (0.28)	59.72 (0.05)	56.49 (0.17)	27.58 (0.15)	30.00 (0.00)	29.43 (0.07)	23.47 (0.21)	29.72 (0.05)	27.06 (0.15)	2380.23 (0.36)	2383.44 (0.25)	2381.25 (0.35)
Zhao_2012	bin=0.5	34.61 (0.35)	26.42 (0.37)	23.51 (0.37)	20.62 (0.20)	16.46 (0.24)	14.86 (0.25)	14.00 (0.23)	9.96 (0.19)	8.65 (0.18)	2383.59 (0.33)	2385.15 (0.29)	2386.07 (0.21)
	bin=0.25	10.59 (0.31)	9.41 (0.26)	8.53 (0.26)	8.76 (0.25)	8.08 (0.22)	7.41 (0.23)	1.83 (0.13)	1.33 (0.11)	1.12 (0.11)	2388.87 (0.12)	2388.86 (0.14)	2388.94 (0.12)
SIE-GLM	bin=0.5	36.04 (0.32)	39.28 (0.33)	35.00 (0.32)	22.71 (0.18)	24.54 (0.18)	22.95 (0.20)	13.33 (0.22)	14.74 (0.25)	12.05 (0.23)	2385.73 (0.21)	2384.09 (0.28)	2384.92 (0.25)
	bin=0.25	25.84 (0.39)	37.24 (0.31)	29.83 (0.35)	18.60 (0.25)	24.28 (0.20)	21.09 (0.22)	7.24 (0.22)	12.96 (0.22)	8.74 (0.21)	2387.31 (0.17)	2384.84 (0.25)	2386.44 (0.21)
Raj_2005	parent=2	55.34 (0.14)	57.91 (0.02)	57.16 (0.08)	28.99 (0.09)	29.45 (0.06)	29.44 (0.06)	26.35 (0.15)	28.46 (0.06)	27.72 (0.09)	2345.34 (0.14)	2347.91 (0.02)	2347.16 (0.08)
true		60			30			30			2390		

Table 5: **(Simulation Study:** Network-3 **for Data from a Non-Linear Hawkes Model with Intensity Function in** (6.2)**)** *The time length is* $T = 1000$. *Results are averaged over* 100 *replications, with standard errors denoted in parentheses.*

		Correct_All	Detected_A	Detected_B	Correct_NC
Discrete_L1	bin=0.5	19.32 (0.37)	14.27 (0.28)	5.05 (0.21)	2385.11 (0.27)
	bin=0.25	29.47 (0.41)	20.03 (0.25)	9.44 (0.28)	2382.94 (0.36)
	bin=0.1	36.04 (0.37)	23.17 (0.21)	12.87 (0.26)	2380.29 (0.33)
Continuous_L1		48.70 (0.28)	28.13 (0.11)	20.57 (0.25)	2359.48 (0.88)
Discrete_wL1	bin=0.5	35.11 (0.37)	22.21 (0.22)	12.91 (0.28)	2376.48 (0.43)
	bin=0.25	45.04 (0.34)	26.27 (0.17)	18.77 (0.27)	2375.78 (0.45)
	bin=0.1	49.84 (0.27)	27.96 (0.13)	21.88 (0.24)	2376.30 (0.44)
Continuous_wL1		52.58 (0.23)	28.70 (0.10)	23.88 (0.22)	2376.53 (0.42)
Zhao_2012	bin=0.5	9.41 (0.32)	7.27 (0.24)	2.14 (0.14)	2388.17 (0.16)
	bin=0.25	4.95 (0.20)	4.69 (0.19)	0.26 (0.05)	2389.42 (0.07)
SIE-GLM	bin=0.5	19.03 (0.31)	15.65 (0.25)	3.38 (0.15)	2387.44 (0.20)
	bin=0.25	17.70 (0.33)	15.43 (0.28)	2.27 (0.13)	2387.73 (0.18)
Raj_2005	parent=2	55.33 (0.15)	29.19 (0.08)	26.14 (0.15)	2345.33 (0.15)
true		60	30	30	2390

Table 6: **(Simulation Study:** Network-1 **with Data-Driven** $\phi_{\widehat{k}}$**)** *The connection strength is* $\beta = 0.5$*. The* Continuous_L1 *and* Continuous_wL1 *methods use the data-driven time-lag* $\phi_{\widehat{k}}$ *following the algorithm in Appendix* A.1*, with* $\phi_{\max} = 3$*,* $\zeta = 0.7$*, and* $k_{\max} = 8$*. Parameter estimation involves* $g(x) = \log(1 + x \wedge c)$*, with data-driven* c *from* (A.2)*. Results are averaged over* 100 *replications, with standard errors denoted in parentheses.*

		Correct_All		Detected_A		Detected_B		Correct_NC	
time length T =		500	1000	500	1000	500	1000	500	1000
Discrete_L1	bin=0.5	2.39 (0.15)	5.69 (0.16)	1.81 (0.12)	3.88 (0.11)	0.57 (0.07)	1.81 (0.08)	79.48 (0.08)	78.94 (0.11)
	bin=0.25	3.63 (0.16)	7.46 (0.14)	2.69 (0.11)	4.88 (0.09)	0.94 (0.08)	2.58 (0.09)	79.18 (0.10)	78.48 (0.14)
	bin=0.1	4.63 (0.17)	8.40 (0.12)	3.32 (0.11)	5.37 (0.07)	1.31 (0.09)	3.03 (0.08)	78.91 (0.14)	78.47 (0.13)
Continuous_L1		5.85 (0.19)	9.41 (0.08)	3.80 (0.13)	5.77 (0.05)	2.04 (0.10)	3.64 (0.06)	75.87 (0.24)	76.41 (0.21)
Discrete_wL1	bin=0.5	3.92 (0.15)	7.34 (0.14)	2.75 (0.12)	4.76 (0.10)	1.17 (0.08)	2.58 (0.09)	78.98 (0.10)	78.95 (0.11)
	bin=0.25	5.34 (0.16)	8.69 (0.11)	3.65 (0.11)	5.51 (0.06)	1.69 (0.09)	3.18 (0.08)	78.91 (0.12)	79.11 (0.10)
	bin=0.1	6.51 (0.15)	9.25 (0.08)	4.34 (0.10)	5.76 (0.04)	2.17 (0.10)	3.49 (0.07)	78.80 (0.12)	79.18 (0.09)
Continuous_wL1		6.50 (0.22)	9.38 (0.08)	4.18 (0.14)	5.82 (0.03)	2.31 (0.11)	3.56 (0.07)	78.19 (0.15)	79.01 (0.11)
Zhao_2012	bin=0.5	0.83 (0.08)	2.84 (0.16)	0.67 (0.07)	1.92 (0.12)	0.16 (0.03)	0.92 (0.07)	79.83 (0.04)	79.52 (0.07)
	bin=0.25	0.63 (0.06)	1.19 (0.09)	0.60 (0.06)	1.06 (0.08)	0.03 (0.01)	0.13 (0.03)	79.95 (0.01)	79.91 (0.03)
	bin=0.1	0.20 (0.04)	0.32 (0.05)	0.20 (0.04)	0.32 (0.05)	0.00 (0.00)	0.00 (0.00)	79.95 (0.01)	79.95 (0.02)
SIE-GLM	bin=0.5	2.00 (0.13)	5.12 (0.17)	1.63 (0.11)	3.67 (0.11)	0.38 (0.05)	1.45 (0.09)	79.69 (0.06)	79.20 (0.11)
	bin=0.25	1.81 (0.11)	4.63 (0.15)	1.61 (0.10)	3.51 (0.10)	0.20 (0.04)	1.12 (0.09)	79.68 (0.06)	79.39 (0.08)
	bin=0.1	0.82 (0.08)	2.50 (0.13)	0.80 (0.08)	2.27 (0.11)	0.02 (0.01)	0.22 (0.04)	79.94 (0.02)	79.68 (0.10)
Raj_2005	parent=3	9.68 (0.04)	9.97 (0.01)	5.85 (0.03)	5.99 (0.00)	3.83 (0.04)	3.98 (0.01)	60.15 (0.08)	63.78 (0.20)
true		10		6		4		80	

Table 7: **(Simulation Study:** Network-1 **with Unbounded** $g(\cdot)$**)** *The connection strength is* $\beta = 0.5$*. We use* $g(x) = \log(1 + x)$ *in both synthetic data generation and parameter estimation. Results are averaged over* 100 *replications, with standard errors denoted in parentheses.*

		Correct_All		Detected_A		Detected_B		Correct_NC	
time length T =		500	1000	500	1000	500	1000	500	1000
Discrete_L1	bin=0.5	2.39 (0.15)	5.69 (0.16)	1.81 (0.12)	3.88 (0.11)	0.57 (0.07)	1.81 (0.08)	79.48 (0.08)	78.94 (0.11)
	bin=0.25	3.63 (0.16)	7.46 (0.14)	2.69 (0.11)	4.88 (0.09)	0.94 (0.08)	2.58 (0.09)	79.18 (0.10)	78.48 (0.14)
	bin=0.1	4.63 (0.17)	8.40 (0.12)	3.32 (0.11)	5.37 (0.07)	1.31 (0.09)	3.03 (0.08)	78.91 (0.14)	78.47 (0.13)
Continuous_L1		6.56 (0.14)	9.44 (0.07)	4.24 (0.09)	5.80 (0.04)	2.31 (0.09)	3.65 (0.05)	76.58 (0.21)	76.48 (0.21)
Discrete_wL1	bin=0.5	3.92 (0.15)	7.34 (0.14)	2.75 (0.12)	4.76 (0.10)	1.17 (0.08)	2.58 (0.09)	78.98 (0.10)	78.95 (0.11)
	bin=0.25	5.34 (0.16)	8.69 (0.11)	3.65 (0.11)	5.51 (0.06)	1.69 (0.09)	3.18 (0.08)	78.91 (0.12)	79.11 (0.10)
	bin=0.1	6.51 (0.15)	9.25 (0.08)	4.34 (0.10)	5.76 (0.04)	2.17 (0.10)	3.49 (0.07)	78.80 (0.12)	79.18 (0.09)
Continuous_wL1		7.44 (0.14)	9.52 (0.06)	4.80 (0.09)	5.90 (0.03)	2.64 (0.08)	3.62 (0.06)	78.67 (0.11)	79.03 (0.10)
Zhao_2012	bin=0.5	0.83 (0.08)	2.84 (0.16)	0.67 (0.07)	1.92 (0.12)	0.16 (0.03)	0.92 (0.07)	79.83 (0.04)	79.52 (0.07)
	bin=0.25	0.63 (0.06)	1.19 (0.09)	0.60 (0.06)	1.06 (0.08)	0.03 (0.01)	0.13 (0.03)	79.95 (0.01)	79.91 (0.03)
	bin=0.1	0.20 (0.04)	0.32 (0.05)	0.20 (0.04)	0.32 (0.05)	0.00 (0.00)	0.00 (0.00)	79.95 (0.01)	79.95 (0.02)
SIE-GLM	bin=0.5	2.00 (0.13)	5.12 (0.17)	1.63 (0.11)	3.67 (0.11)	0.38 (0.05)	1.45 (0.09)	79.69 (0.06)	79.20 (0.11)
	bin=0.25	1.81 (0.11)	4.63 (0.15)	1.61 (0.10)	3.51 (0.10)	0.20 (0.04)	1.12 (0.09)	79.68 (0.06)	79.39 (0.08)
	bin=0.1	0.82 (0.08)	2.50 (0.13)	0.80 (0.08)	2.27 (0.11)	0.02 (0.01)	0.22 (0.04)	79.94 (0.02)	79.68 (0.10)
Raj_2005	parent=3	9.68 (0.04)	9.97 (0.01)	5.85 (0.03)	5.99 (0.00)	3.83 (0.04)	3.98 (0.01)	60.15 (0.08)	63.78 (0.20)
true		10		6		4		80	

Table 8: **(Simulation Study:** Network-1 **with Misspecified ϕ and $g(\cdot)$)** *The connection strength is $\beta = 0.5$. In the true model, the lag-width is $\phi = 1$, and the shape function is $g(x) = \log(1 + x \wedge 10)$. In the estimation process, a misspecified $\phi_a = 0.5$ is used alongside $g_a(x) = x \wedge c$, incorporating the data-driven c from* (A.2)*. Results are averaged over* 100 *replications, with standard errors denoted in parentheses.*

		Correct_All		Detected_A		Detected_B		Correct_NC	
time length T =		500	1000	500	1000	500	1000	500	1000
Discrete_L1	bin=0.5	2.12 (0.13)	4.68 (0.17)	1.67 (0.11)	3.40 (0.12)	0.45 (0.06)	1.28 (0.09)	79.19 (0.11)	79.59 (0.07)
	bin=0.25	2.00 (0.12)	4.73 (0.16)	1.57 (0.10)	3.40 (0.11)	0.44 (0.05)	1.33 (0.09)	79.36 (0.09)	79.67 (0.05)
	bin=0.1	2.08 (0.12)	4.97 (0.17)	1.60 (0.10)	3.54 (0.12)	0.48 (0.05)	1.43 (0.09)	79.25 (0.10)	79.63 (0.06)
Continuous_L1		4.30 (0.15)	6.75 (0.17)	2.99 (0.11)	4.22 (0.12)	1.32 (0.09)	2.52 (0.09)	76.90 (0.21)	75.56 (0.21)
Discrete_wL1	bin=0.5	3.43 (0.15)	6.47 (0.15)	2.44 (0.11)	4.28 (0.11)	0.99 (0.08)	2.19 (0.09)	78.98 (0.10)	79.19 (0.09)
	bin=0.25	3.28 (0.16)	6.40 (0.15)	2.24 (0.12)	4.23 (0.11)	1.04 (0.08)	2.17 (0.10)	79.09 (0.10)	79.06 (0.09)
	bin=0.1	3.37 (0.15)	6.43 (0.16)	2.29 (0.11)	4.21 (0.10)	1.08 (0.08)	2.22 (0.10)	79.23 (0.08)	78.93 (0.11)
Continuous_wL1		3.37 (0.16)	6.37 (0.16)	2.20 (0.12)	4.16 (0.11)	1.17 (0.07)	2.21 (0.09)	79.06 (0.09)	78.91 (0.11)
Zhao_2012	bin=0.5	1.50 (0.12)	3.99 (0.18)	1.12 (0.10)	2.73 (0.13)	0.38 (0.05)	1.26 (0.09)	79.28 (0.09)	79.70 (0.06)
	bin=0.25	0.66 (0.07)	1.35 (0.09)	0.61 (0.07)	1.11 (0.08)	0.05 (0.02)	0.24 (0.04)	79.92 (0.03)	79.94 (0.02)
	bin=0.1	0.23 (0.04)	0.27 (0.04)	0.23 (0.04)	0.27 (0.04)	0.00 (0.00)	0.00 (0.00)	79.95 (0.02)	79.92 (0.03)
SIE-GLM	bin=0.5	1.89 (0.13)	4.47 (0.18)	1.51 (0.11)	3.25 (0.13)	0.38 (0.05)	1.22 (0.09)	79.55 (0.07)	79.78 (0.04)
	bin=0.25	1.22 (0.10)	3.22 (0.14)	1.09 (0.09)	2.54 (0.11)	0.13 (0.03)	0.68 (0.07)	79.75 (0.06)	79.84 (0.04)
	bin=0.1	0.53 (0.06)	1.49 (0.09)	0.53 (0.06)	1.38 (0.09)	0.00 (0.00)	0.11 (0.03)	79.84 (0.03)	79.89 (0.03)
Raj_2005	parent=3	8.39 (0.11)	9.68 (0.05)	5.20 (0.07)	5.84 (0.03)	3.19 (0.07)	3.84 (0.04)	64.48 (0.21)	59.41 (0.15)
true		10		6		4		80	

B Proofs of main results

B.1 Notations in the proof

For an event A in the sample space Ω, the event $\overline{A}$ denotes the complement of A. For two events A and B, the event $A \setminus B$ denotes $A \cap \overline{B}$. For an event A, we write $\sigma(\mathscr{F}, A) = \sigma(\mathscr{F}, \mathrm{I}(A))$. Let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Let $\mathbf{C} \succ \mathbf{0}$ denote a positive definite matrix $\mathbf{C}$.

B.2 Conditions

The conditions are not the weakest possible, but facilitate the derivations.

A1. The number of nodes $V \geq 2$ is a fixed integer. In the multivariate point process, event time points $\{T_{i,\ell}\}_{i\in\mathcal{V},\ell\geq 1}$ satisfy $0 < T_{i,1} < T_{i,2} < \cdots$ for each $i \in \mathcal{V}$.

A2. The multivariate counting process satisfies $\lim_{\Delta\downarrow 0} \Delta^{-1}\mathrm{P}(N_i(t+\Delta) = N_i(t)+1 \mid \mathscr{F}_t) = \lim_{\Delta\downarrow 0} \Delta^{-1}\mathrm{P}(N_i(t+\Delta) \neq N_i(t) \mid \mathscr{F}_t)$ a.s. for every $i \in \mathcal{V}$ and $t \geq 0$.

A3. The multivariate counting process $\boldsymbol{N}(t)$ satisfies the OM condition in Definition 1.

A4. There exists a random variable $Z > 0$ with $\mathrm{E}(Z) < \infty$, such that for any $\Delta \in (0, c_0)$ with a constant $c_0 \in (0, 1)$ and any $t \geq 0$, $\mathrm{P}(\boldsymbol{N}(t+\Delta) \neq \boldsymbol{N}(t) \mid \mathscr{F}_t)/\Delta \leq Z$, a.s..

A5. In (3.5), the shape-function $g(\cdot) : [0, \infty) \to [0, \infty)$ is continuous, non-negative, monotonically increasing, and bounded above, with $g(0) = 0$, and $\sup_{x\in[0,\infty)} g(x) \leq C_0$ for some constant $C_0 \in (0, \infty)$.

A6. For all $i \in \mathcal{V}$, the true self-effect parameter $\beta^*_{i,i} = 0$.

A7. The true edge set $\mathcal{E}^* \neq \varnothing$.

A8. The edge set $\mathcal{E}$ in (3.2) satisfies $\mathcal{E} \neq \varnothing$.

Condition A1 relates to the basic definition of a multivariate point process. Condition A2 is related to the regular point process as defined in [38] and is explicitly discussed in our Remark 1. Condition A3 is explicitly presented in our Definition 1. Condition A4 bears resemblance to conditions (2)–(3) in [38], which ensure the applicability of the dominated convergence theorem. Condition A5 guarantees the boundedness property of the conditional intensity processes $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ in our model (3.1). Condition A6 excludes self-effects in model (3.1), thereby preventing the presence of "self-loop" in the corresponding network structure $\mathcal{G}$. Conditions A7 and A8 are imposed to ensure that our multivariate point process does not reduce to the trivial case of a homogeneous Poisson process.

B.3 Proof of the statement in Remark 1

We aim to prove the statement: any *multivariate regular point process* also has identical limits (2.8) and (2.9).

From the definition of a multivariate regular point process $\boldsymbol{N}(t)$, we have $\lim_{\Delta\downarrow 0}\Delta^{-1}\mathrm{P}(N_i(t+\Delta)=N_i(t)+1\mid\mathscr{F}_t)=\lim_{\Delta\downarrow 0}\Delta^{-1}\mathrm{P}(N_i(t+\Delta)\neq N_i(t)\mid\mathscr{F}_t)$, a.s., for any $i\in\mathcal{V}$ and $t\geq 0$. Since $\{N_i(t+\Delta)=N_i(t)+1\}\subseteq\{N_i(t+\Delta)\neq N_i(t)\}$, we further get

$$\lim_{\Delta\downarrow 0}\Delta^{-1}\mathrm{P}\big(\{N_i(t+\Delta)\neq N_i(t)\}\setminus\{N_i(t+\Delta)=N_i(t)+1\}\,\big|\,\mathscr{F}_t\big)=0,\quad\text{a.s..}\tag{B.1}$$

Thus,

$$\begin{aligned}0 &\leq \lim_{\Delta\downarrow 0}\Delta^{-1}\Big[\mathrm{P}\big(\boldsymbol{N}(t+\Delta)\neq\boldsymbol{N}(t)\mid\mathscr{F}_t\big)-\mathrm{P}\Big(\bigcup_{i\in\mathcal{V}}\{N_i(t+\Delta)=N_i(t)+1\}\,\Big|\,\mathscr{F}_t\Big)\Big]\\ &\leq \sum_{i\in\mathcal{V}}\lim_{\Delta\downarrow 0}\Delta^{-1}\mathrm{P}\Big(\{N_i(t+\Delta)\neq N_i(t)\}\setminus\{N_i(t+\Delta)=N_i(t)+1\}\,\Big|\,\mathscr{F}_t\Big)\\ &= 0,\quad\text{a.s.},\end{aligned}$$

where the last equality is from (B.1). Hence, we obtain $\lim_{\Delta\downarrow 0}\Delta^{-1}\mathrm{P}(\cup_{i\in\mathcal{V}}\{N_i(t+\Delta)=N_i(t)+1\}\mid\mathscr{F}_t)=\lim_{\Delta\downarrow 0}\Delta^{-1}\mathrm{P}(\boldsymbol{N}(t+\Delta)\neq\boldsymbol{N}(t)\mid\mathscr{F}_t)$, a.s.. This completes the proof. ■

B.4 Proof of Lemma 1

Define the events $A_{i,\Delta}=\{N_i(t+\Delta)=N_i(t)+1\}$. Then

$$\mathrm{P}\Big(\bigcup_{i\in\mathcal{V}}\{N_i(t+\Delta)=N_i(t)+1\}\,\Big|\,\mathscr{F}_t\Big)\;=\;\mathrm{P}\Big(\bigcup_{i\in\mathcal{V}}A_{i,\Delta}\,\Big|\,\mathscr{F}_t\Big).\tag{B.2}$$

By the inclusion-exclusion formula, we have that

$$\begin{aligned}&\mathrm{P}\Big(\bigcup_{i\in\mathcal{V}}A_{i,\Delta}\,\Big|\,\mathscr{F}_t\Big)\\ =\;&\sum_{k=1}^{V}(-1)^{k+1}\sum_{\{i_1,\ldots,i_k\}\subseteq\mathcal{V}}\mathrm{P}\Big(\bigcap_{j\in\{i_1,\ldots,i_k\}}A_{j,\Delta}\,\Big|\,\mathscr{F}_t\Big)\\ =\;&\sum_{i=1}^{V}\mathrm{P}(A_{i,\Delta}\mid\mathscr{F}_t)-\sum_{k=2}^{V}(-1)^{k}\sum_{\{i_1,\ldots,i_k\}\subseteq\mathcal{V}}\mathrm{P}\Big(\bigcap_{j\in\{i_1,\ldots,i_k\}}A_{j,\Delta}\,\Big|\,\mathscr{F}_t\Big).\end{aligned}\tag{B.3}$$

For mutually distinct $\{i_1,\ldots,i_k\}$ with $k\geq 2$, the OM condition (2.10) implies that

$$\mathrm{P}\Big(\bigcap_{j\in\{i_1,\ldots,i_k\}}A_{j,\Delta}\,\Big|\,\mathscr{F}_t\Big)\;\leq\;\mathrm{P}(A_{i_1,\Delta}\cap A_{i_2,\Delta}\mid\mathscr{F}_t)$$

$$= \Delta^2\{\lambda_{i_1}(t \mid \mathscr{F}_t)\,\lambda_{i_2}(t \mid \mathscr{F}_t) + o(1)\}, \quad \text{a.s.} \tag{B.4}$$

as $\Delta \downarrow 0$. Plugging (B.4) into (B.3), we obtain

$$\begin{aligned} &\frac{1}{\Delta}\mathrm{P}\Big(\bigcup_{i\in\mathcal{V}} A_{i,\Delta}\,\Big|\,\mathscr{F}_t\Big) \\ = \;& \frac{1}{\Delta}\sum_{i=1}^{V}\mathrm{P}(A_{i,\Delta} \mid \mathscr{F}_t) + O(\Delta) = \sum_{i=1}^{V}\lambda_i(t \mid \mathscr{F}_t) + o(1), \quad \text{a.s.} \end{aligned}$$

as $\Delta \downarrow 0$. It follows that $\lambda^{\mathrm{sum}}(t \mid \mathscr{F}_t) = \lim_{\Delta\downarrow 0}\Delta^{-1}\mathrm{P}(\cup_{i\in\mathcal{V}}A_{i,\Delta} \mid \mathscr{F}_t) = \sum_{i=1}^{V}\lambda_i(t \mid \mathscr{F}_t)$. This completes the proof. ■

B.5 Proof of Lemma 4

We first prove part (i). By utilizing [15] (Theorem 2.4.7, p. 84) and the fact that $\lambda_j(t \mid \mathscr{F}_t)$ in (3.1) is finite, our event time points $\{T_{j,\ell}\}_{j\in\mathcal{V},\ell\geq 1}$ are totally inaccessible stopping times. This combined with [15] (Proposition 2.4.6, p. 83) gives that $\mathrm{P}(T_{i,k} = T_{j,r}) = 0$, for any distinct nodes $i, j \in \mathcal{V}$, and any integers $k \geq 1$ and $r \geq 1$. It follows that

$$\begin{aligned} &\mathrm{P}\big(\breve{T}_\ell \in \{T_{i,k}\}_{k\geq 1}\,,\,\breve{T}_\ell \in \{T_{j,k}\}_{k\geq 1}\big) \\ \leq\;& \mathrm{P}\big(\cup_{k\geq 1}\cup_{r\geq 1}\{T_{i,k} = T_{j,r}\}\big) \;=\; 0. \end{aligned} \tag{B.5}$$

Next we prove part (ii). Using the similar proof of (B.5), for any $i \in \mathcal{V}$, we have

$$\begin{aligned} &\mathrm{P}\big(\breve{T}_\ell \in \{T_{i,k}\}_{k\geq 1}\,,\,\breve{T}_\ell \in \{T_{j,k}+\phi\}_{k\geq 1}\big) \\ \leq\;& \mathrm{P}\big(\cup_{k\geq 1}\cup_{r\geq 1}\{T_{i,k} = T_{j,r}+\phi\}\big) \;=\; 0 \end{aligned} \tag{B.6}$$

for any $j \in \mathcal{V}$, and thus

$$\begin{aligned} &\mathrm{P}\big(\breve{T}_\ell \in \{T_{i,k}\}_{k\geq 1}\,,\,\breve{T}_\ell \in \{T_{j,k}+\phi\}_{j\in\mathcal{V},\,k\geq 1}\big) \\ \leq\;& \sum_{j\in\mathcal{V}}\mathrm{P}\big(\breve{T}_\ell \in \{T_{i,k}\}_{k\geq 1}\,,\,\breve{T}_\ell \in \{T_{j,k}+\phi\}_{k\geq 1}\big) = 0. \end{aligned} \tag{B.7}$$

The proof is completed. ■

B.6 Proof of Theorem 1

Before proving Theorem 1, we first show Lemmas B.1 and B.2 following Definition 4.

Definition 4 ($\mathrm{E}(X \parallel \mathscr{F}, A)$) *Let $(\Omega, \mathscr{G}, \mathrm{P})$ be a probability space. Let $\mathscr{F} \subseteq \mathscr{G}$ be a sub σ-field, and $A \in \mathscr{G}$ be an event such that $A \notin \mathscr{F}$ and $\mathrm{P}(A \mid \mathscr{F}) > 0$ almost surely. For any random variable X in $(\Omega, \mathscr{G}, \mathrm{P})$, define*

$$\mathrm{E}(X \parallel \mathscr{F}, A) \;=\; \mathrm{E}\{X\,\mathrm{I}(A) \mid \mathscr{F}\}/\mathrm{P}(A \mid \mathscr{F}). \tag{B.8}$$

(*Remark: when $\mathscr{F} = \{\Omega, \varnothing\}$, $\mathrm{E}(X \parallel \mathscr{F}, A)$ in* (B.8) *reduces to $\mathrm{E}(X \mid A) = \mathrm{E}\{X\,\mathrm{I}(A)\}/\mathrm{P}(A)$; when X is independent of $\mathscr{F}$ and A, $\mathrm{E}(X \parallel \mathscr{F}, A)$ in* (B.8) *reduces to $\mathrm{E}(X)$.*)

Lemma B.1 (Conditional probability $\mathrm{P}(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s)$**)** *Assume conditions* A1, A2, A3, A4 *and* A5 *in Appendix* B. *Then for $t \geq s \geq 0$,*

$$\mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big) \;=\; \exp\Big\{ -\int_s^t \lambda^{\mathrm{sum}}(u; \mathscr{F}_s)\,\mathrm{d}u \Big\}, \tag{B.9}$$

with

$$\lambda^{\mathrm{sum}}(t; \mathscr{F}_s) \;=\; \mathrm{E}\{\lambda^{\mathrm{sum}}(t \mid \mathscr{F}_t) \parallel \mathscr{F}_s, \boldsymbol{N}(t) = \boldsymbol{N}(s)\}, \tag{B.10}$$

where $\lambda^{\mathrm{sum}}(t \mid \mathscr{F}_t)$ is the total intensity in (2.8) *and* (2.9). (*Remark: $\lambda^{\mathrm{sum}}(t; \mathscr{F}_s)$ in* (B.10) *is motivated by the definition $\lambda_{N(t)}(t, \mathscr{B}_s)$ in Lemma* 1 *of* [38], *and* (B.9) *is motivated by Corollary* 1 *in* [38].)

Proof: For $t \geq s \geq 0$, and $\Delta > 0$, note that

$$\{\boldsymbol{N}(t+\Delta) = \boldsymbol{N}(s)\} = \{\boldsymbol{N}((s, t+\Delta]) = \boldsymbol{0}\} \;\subseteq\; \{\boldsymbol{N}((s,t]) = \boldsymbol{0}\} = \{\boldsymbol{N}(t) = \boldsymbol{N}(s)\},$$

which implies that

$$\begin{aligned} & \mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big) - \mathrm{P}\big(\boldsymbol{N}(t+\Delta) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big) \\ = \;& \mathrm{P}\big(\boldsymbol{N}(t+\Delta) \neq \boldsymbol{N}(t),\, \boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big). \end{aligned}$$

Combining this with the fact that $\mathscr{F}_s \subseteq \mathscr{F}_t$, and (2.9), (B.8), (B.10), we obtain

$$\begin{aligned} & \frac{\partial\, \mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big)}{\partial\, t} \\ = \;& -\lim_{\Delta \downarrow 0} \Delta^{-1} \mathrm{E}\big\{ \mathrm{P}\big(\boldsymbol{N}(t+\Delta) \neq \boldsymbol{N}(t) \mid \mathscr{F}_t\big) \cdot \mathrm{I}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s)\big) \mid \mathscr{F}_s \big\} \\ = \;& -\mathrm{E}\Big\{ \lim_{\Delta \downarrow 0} \Delta^{-1} \mathrm{P}\big(\boldsymbol{N}(t+\Delta) \neq \boldsymbol{N}(t) \mid \mathscr{F}_t\big) \cdot \mathrm{I}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s)\big) \mid \mathscr{F}_s \Big\} \qquad \text{(B.11)} \\ = \;& -\lambda^{\mathrm{sum}}(t; \mathscr{F}_s) \cdot \mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big), \end{aligned}$$

where the interchange of limit and expectation in (B.11) follows by the dominated convergence theorem and condition A4. Solving the above differential equation completes the proof of (B.9). ■

Lemma B.2 ($\mathrm{E}\{\lambda_i(S \mid \mathcal{F}_S) \parallel \mathscr{F}_s, \boldsymbol{N}(S) = \boldsymbol{N}(s)\}$) *Assume conditions* A1, A2, A3, A4 *and* A5 *in Appendix* B. *Define* $\mathcal{T}_s = \cup_{j\in\mathcal{V}}\{t \in (s-\phi, s] : N_j(\{t\}) = 1\}$, *and*

$$T_s^{\mathrm{o}} = \begin{cases} \min(\mathcal{T}_s) + \phi, & \text{if } \mathcal{T}_s \neq \varnothing, \\ \infty, & \text{if } \mathcal{T}_s = \varnothing, \end{cases} \tag{B.12}$$

for a fixed time point $s \in [0, \infty)$. *Then for any random variable* S *satisfying* $s \leq S < T_s^{\mathrm{o}}$, *we have*

$$\mathrm{E}\big\{\lambda_i(S \mid \mathcal{F}_S) \parallel \mathscr{F}_s, \boldsymbol{N}(S) = \boldsymbol{N}(s)\big\} = \lambda_i\big(s \mid \mathscr{F}_s\big), \text{ for all } i \in \mathcal{V}. \tag{B.13}$$

Proof: For $S = s$, (B.13) obviously holds. It suffices to prove (B.13) for $s < S < T_s^{\mathrm{o}}$. First, we show the following statement:

$$\begin{gathered} \text{for } s < S < T_s^{\mathrm{o}},\ \boldsymbol{N}(S) = \boldsymbol{N}(s) \text{ implies } \cup_{j\in\mathcal{V}} \big\{t \in (s, S] : N_j(\{t\}) = 1\big\} = \varnothing \\ \text{and } \cup_{j\in\mathcal{V}} \big\{t \in (s, S] : N_j(\{t-\phi\}) = 1\big\} = \varnothing. \end{gathered} \tag{B.14}$$

Let $B_1 = \cup_{j\in\mathcal{V}}\{t \in (s, S] : N_j(\{t\}) = 1\}$ and $B_2 = \cup_{j\in\mathcal{V}}\{t \in (s, S] : N_j(\{t-\phi\}) = 1\}$. Note that $\boldsymbol{N}(S) = \boldsymbol{N}(s)$ directly implies $B_1 = \varnothing$. To prove (B.14), it suffices to show $B_2 = \varnothing$, whose proof is given according to whether $\mathcal{T}_s \neq \varnothing$ or not.

If $\mathcal{T}_s \neq \varnothing$, then (B.12) yields that $s - \phi < \min(\mathcal{T}_s)$, implying $\cup_{j\in\mathcal{V}}\{t \in (s, \min(\mathcal{T}_s) + \phi) : N_j(\{t-\phi\}) = 1\} - \phi = \cup_{j\in\mathcal{V}}\{t \in (s-\phi, \min(\mathcal{T}_s)) : N_j(\{t\}) = 1\} = \varnothing$. This, together with $s < S < \min(\mathcal{T}_s) + \phi$, gives $B_2 = \varnothing$.

If $\mathcal{T}_s = \varnothing$, we obtain

$$\begin{aligned} B_2 &\subseteq \big\{ \cup_{j\in\mathcal{V}} \{t \in (s, s+\phi] : N_j(\{t-\phi\}) = 1\}\big\} \\ &\quad \bigcup \big\{ \cup_{j\in\mathcal{V}} \{t \in ((s+\phi)\wedge S, S] : N_j(\{t-\phi\}) = 1\}\big\} \\ &\subseteq \{\mathcal{T}_s + \phi\} \cup \{B_1 + \phi\} \\ &= \varnothing, \end{aligned}$$

where $\mathcal{T}_s + \phi = \{t + \phi : t \in \mathcal{T}_s\}$.

Combining the above two cases, we verified (B.14).

Next, we prove Lemma B.2. If $\boldsymbol{N}(S) = \boldsymbol{N}(s)$, then (B.14) indicates that the intensity functions $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ are continuous in $(s, S]$. For piecewise-constant functions $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$, it follows that if $\boldsymbol{N}(S) = \boldsymbol{N}(s)$, then $\lambda_i(S \mid \mathcal{F}_S) = \lambda_i(s \mid \mathscr{F}_s)$, i.e.,

$$\lambda_i(S \mid \mathcal{F}_S) \cdot \mathrm{I}(A) = \lambda_i(s \mid \mathscr{F}_s) \cdot \mathrm{I}(A), \quad i \in \mathcal{V},$$

where A denotes the event $\{\boldsymbol{N}(S) = \boldsymbol{N}(s)\}$. Combining this with (B.8), we obtain

$$\mathrm{E}\big\{\lambda_i(S \mid \mathcal{F}_S) \parallel \mathscr{F}_s,\ A\big\} = \frac{\mathrm{E}\{\lambda_i(S \mid \mathcal{F}_S) \cdot \mathrm{I}(A) \mid \mathscr{F}_s\}}{\mathrm{P}(A \mid \mathscr{F}_s)} = \frac{\mathrm{E}\{\lambda_i(s \mid \mathscr{F}_s) \cdot \mathrm{I}(A) \mid \mathscr{F}_s\}}{\mathrm{P}(A \mid \mathscr{F}_s)} = \lambda_i(s \mid \mathscr{F}_s).$$

This completes the proof. ■

Now we prove Theorem 1. We first show part (i). By the definition in (4.5), $\breve{T}_\ell < \breve{T}_{\ell+1}$ holds for any integer $\ell \geq 1$. It suffices to show $\breve{T}_{\ell+1} \leq T_\ell^*$. If $\mathcal{T}_\ell = \varnothing$, then $T_\ell^* = \infty$ in (4.10) completes the proof. If $\mathcal{T}_\ell \neq \varnothing$, then any $t_\ell \in \mathcal{T}_\ell$ in (4.9) indicates that $t_\ell = T_{i,k}$ for some integers $i \in \mathcal{V}$ and $k \geq 1$, and $\breve{T}_\ell < t_\ell + \phi \equiv T_{i,k} + \phi \leq \breve{T}_\ell + \phi$. Also, $t_\ell + \phi \equiv T_{i,k} + \phi \in \{\breve{T}_1, \breve{T}_2, \ldots\}$. This combined with $\breve{T}_\ell < \breve{T}_{\ell+1}$ implies $\breve{T}_{\ell+1} \leq t_\ell + \phi$. Thus $\breve{T}_{\ell+1} \leq \min\{t_\ell : t_\ell \in \mathcal{T}_\ell\} + \phi = \min(\mathcal{T}_\ell) + \phi = T_\ell^*$.

Before proving parts (ii) and (iii), preparations (a), (b), (c) and (d) are made below.

(a) Since $\mathcal{F}_{\breve{T}_\ell} = \sigma\{(\breve{T}_0, I_0), \ldots, (\breve{T}_\ell, I_\ell)\}$, it suffices to show that (4.11)–(4.13) hold conditional on each realization

$$\Big\{\{(\breve{T}_0, I_0), \ldots, (\breve{T}_\ell, I_\ell)\} = \{(\breve{t}_0, i_0), \ldots, (\breve{t}_\ell, i_\ell)\}\Big\} = \bullet \in \mathcal{F}_{\breve{T}_\ell}. \tag{B.15}$$

Note that the realization $\bullet$ in (B.15) is known from the history up to time $\breve{t}_\ell$. Following the notation $\mathscr{F}_t$ in (2.4), we also have $\bullet \in \mathscr{F}_{\breve{t}_\ell}$, and thus for a random variable $X : \Omega \to \mathbb{R}$, which is measurable with respect to either $\mathcal{F}_{\breve{T}_\ell}$ or $\mathscr{F}_{\breve{t}_\ell}$, denote by $X(\bullet)$ the value of X at the realization $\bullet$. For example, we can write $\breve{T}_\ell(\bullet) = \breve{t}_\ell$ and $I_\ell(\bullet) = i_\ell$.

(b) Comparing the random variables T_ℓ^* in (4.10) and $T_{\breve{t}_\ell}^{\mathrm{o}}$ in (B.12) (with $s = \breve{t}_\ell$), we observe that they have the same value t_ℓ^* at the realization $\bullet$, i.e.,

$$t_\ell^* = T_\ell^*(\bullet) = T_{\breve{t}_\ell}^{\mathrm{o}}(\bullet), \qquad \text{and} \quad \breve{t}_\ell < t_\ell^*.$$

(c) Also, we verify the following equation for $t \in (\breve{t}_\ell, t_\ell^*)$:

$$\mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \mid \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) = \exp\{-\lambda^{\mathrm{sum}}(\breve{t}_\ell \mid \mathscr{F}_{\breve{t}_\ell})(\bullet) \cdot (t - \breve{t}_\ell)\}. \tag{B.16}$$

Using (B.9) (with $s = \breve{t}_\ell$) yields that

$$\mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \mid \mathscr{F}_{\breve{t}_\ell}\big) = \exp\Big\{-\int_{\breve{t}_\ell}^t \lambda^{\mathrm{sum}}\big(u; \mathscr{F}_{\breve{t}_\ell}\big)\,\mathrm{d}u\Big\}, \text{ for } t > \breve{t}_\ell. \tag{B.17}$$

Since both sides of (B.17) are $\mathscr{F}_{\breve{t}_\ell}$-measurable random variables, it follows that

$$\mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \mid \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) = \exp\Big\{-\int_{\breve{t}_\ell}^t \lambda^{\mathrm{sum}}\big(u; \mathscr{F}_{\breve{t}_\ell}\big)(\bullet)\,\mathrm{d}u\Big\}, \text{ for } t > \breve{t}_\ell. \tag{B.18}$$

For $t \in (\breve{t}_\ell, t_\ell^*)$, define a random variable S to satisfy $S(\bullet) = t$ and $\breve{t}_\ell \leq S < T_{\breve{t}_\ell}^{\mathrm{o}}$ (e.g., if $T_\ell^* < \infty$, then let $S = \eta\,\breve{t}_\ell + (1-\eta)T_{\breve{t}_\ell}^{\mathrm{o}}$, with $\eta = (t_\ell^* - t)/(t_\ell^* - \breve{t}_\ell) \in (0,1)$; if $T_\ell^* = \infty$,

then let $S = \breve{t}_\ell \cdot \mathrm{I}(T^{\mathrm{o}}_{\breve{t}_\ell} < \infty) + t \cdot \mathrm{I}(T^{\mathrm{o}}_{\breve{t}_\ell} = \infty)$, where $S(\bullet) = t$ holds due to $T^{\mathrm{o}}_{\breve{t}_\ell}(\bullet) = t^*_\ell$, and $\breve{t}_\ell \le S < T^{\mathrm{o}}_{\breve{t}_\ell}$ is valid due to $\breve{t}_\ell < T^{\mathrm{o}}_{\breve{t}_\ell}$). Applying Lemma B.2 (with $s = \breve{t}_\ell$), we have

$$\mathrm{E}\big\{\lambda_i(S \mid \mathcal{F}_S) \parallel \mathscr{F}_{\breve{t}_\ell}, \boldsymbol{N}(S) = \boldsymbol{N}(\breve{t}_\ell)\big\} = \lambda_i\big(\breve{t}_\ell \mid \mathscr{F}_{\breve{t}_\ell}\big), \text{ for all } i \in \mathcal{V}. \tag{B.19}$$

Similarly to (B.10), for $i \in \mathcal{V}$ and $t \ge s \ge 0$, define

$$\lambda_i(t; \mathscr{F}_s) \;=\; \mathrm{E}\{\lambda_i(t \mid \mathscr{F}_t) \parallel \mathscr{F}_s, \boldsymbol{N}(t) = \boldsymbol{N}(s)\}. \tag{B.20}$$

For $t \in (\breve{t}_\ell, t^*_\ell)$, using the definition (B.20) (with $s = \breve{t}_\ell$), the fact $S(\bullet) = t$ and (B.19), we obtain

$$\begin{aligned}\lambda_i\big(t; \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) &= \mathrm{E}\big\{\lambda_i(t \mid \mathscr{F}_t) \parallel \mathscr{F}_{\breve{t}_\ell}, \boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell)\big\}(\bullet) \\ &= \mathrm{E}\big\{\lambda_i(S \mid \mathcal{F}_S) \parallel \mathscr{F}_{\breve{t}_\ell}, \boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell)\big\}(\bullet) \;=\; \lambda_i\big(\breve{t}_\ell \mid \mathscr{F}_{\breve{t}_\ell}\big)(\bullet), \text{ for all } i \in \mathcal{V}. \end{aligned} \tag{B.21}$$

Summing over $i \in \mathcal{V}$ on both sides of (B.21) gives that

$$\lambda^{\mathrm{sum}}\big(t; \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) = \lambda^{\mathrm{sum}}\big(\breve{t}_\ell \mid \mathscr{F}_{\breve{t}_\ell}\big)(\bullet). \tag{B.22}$$

By (B.22), we simplify (B.18) to be

$$\begin{aligned}\mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \mid \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) &= \exp\Big\{-\int_{\breve{t}_\ell}^{t} \lambda^{\mathrm{sum}}\big(u \mid \mathscr{F}_u\big)(\bullet)\,\mathrm{d}u\Big\} \\ &= \exp\{-\lambda^{\mathrm{sum}}(\breve{t}_\ell \mid \mathscr{F}_{\breve{t}_\ell})(\bullet) \cdot (t - \breve{t}_\ell)\},\end{aligned}$$

which proves (B.16).

(d) We show the following statement:

$$\text{for } t \in (\breve{T}_\ell, T^*_\ell),\ \{\boldsymbol{N}(t) = \boldsymbol{N}(\breve{T}_\ell)\} \text{ is equivalent to } \{\breve{T}_{\ell+1} > t\}. \tag{B.23}$$

The proof for the sufficiency part is similar to that of (B.14). For the necessity part, if $\boldsymbol{N}(t) \ne \boldsymbol{N}(\breve{T}_\ell)$, then there exists $T_{i,k} \in (\breve{T}_\ell, t]$ for some integers $i \in \mathcal{V}$ and $k \ge 1$. Also, $\breve{T}_\ell$ and $\breve{T}_{\ell+1}$ are two consecutive discontinuity points in the set (4.5), which implies that $\cup_{i\in\mathcal{V}} \cup_{k\ge1} \{T_{i,k} : \breve{T}_\ell < T_{i,k} < \breve{T}_{\ell+1}\} = \varnothing$. Combining this with the fact that $T_{i,k} \in (\breve{T}_\ell, t]$, we obtain $\breve{T}_\ell < \breve{T}_{\ell+1} \le T_{i,k} \le t$, and thus $\breve{T}_{\ell+1} \le t$.

We next prove parts (ii) and (iii) of Theorem 1. **Proof of part** (ii). For $T^*_\ell < \infty$, using the similar proof as that of (B.23), we have that $\boldsymbol{N}((\breve{T}_\ell, T^*_\ell)) = \mathbf{0}$ is equivalent to $\breve{T}_{\ell+1} \ge T^*_\ell$. Also, the result $\breve{T}_{\ell+1} \in (\breve{T}_\ell, T^*_\ell]$ in part (i) implies that $\breve{T}_{\ell+1} \ge T^*_\ell$ is equivalent to $\breve{T}_{\ell+1} = T^*_\ell$. It follows that

$$\mathrm{P}\big(\breve{T}_{\ell+1} = T^*_\ell \,\big|\, \mathcal{F}_{\breve{T}_\ell}\big) = \mathrm{P}\big(\breve{T}_{\ell+1} \ge T^*_\ell \,\big|\, \mathcal{F}_{\breve{T}_\ell}\big) = \mathrm{P}\big(\boldsymbol{N}\big((\breve{T}_\ell,\, T^*_\ell)\big) = \mathbf{0} \mid \mathcal{F}_{\breve{T}_\ell}\big). \tag{B.24}$$

Evaluating both sides of (B.24) at the realization $\bullet$ and using $t_\ell^* = T_\ell^*(\bullet)$, we obtain

$$\begin{aligned} \mathrm{P}\big(\breve{T}_{\ell+1} = T_\ell^* \mid \mathcal{F}_{\breve{T}_\ell}\big)(\bullet) &= \lim_{t \uparrow t_\ell^*} \mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \mid \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) \\ &= \lim_{t \uparrow t_\ell^*} \exp\{-\lambda^{\mathrm{sum}}(\breve{t}_\ell \mid \mathscr{F}_{\breve{t}_\ell})(\bullet) \cdot (t - \breve{t}_\ell)\} \qquad \text{(B.25)} \\ &= \exp\big\{ -\lambda^{\mathrm{sum}}(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}) \cdot (T_\ell^* - \breve{T}_\ell)\big\}(\bullet), \end{aligned}$$

where (B.25) is derived from (B.16) with $t \in (\breve{t}_\ell, t_\ell^*)$. This proves (4.11). By using $\breve{t}_\ell = \breve{T}_\ell(\bullet)$, (B.23) and (B.16), for $t \in (\breve{T}_\ell, T_\ell^*)$, we have

$$\begin{aligned} f_{\breve{T}_{\ell+1} \mid \mathcal{F}_{\breve{T}_\ell}}(t)(\bullet) &= \frac{-\partial\, \mathrm{P}(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell) \mid \mathscr{F}_{\breve{t}_\ell})(\bullet)}{\partial\, t} \\ &= \frac{-\partial \exp\{-\lambda^{\mathrm{sum}}\big(\breve{t}_\ell \mid \mathscr{F}_{\breve{t}_\ell}\big)(\bullet) \cdot (t - \breve{t}_\ell)\}}{\partial\, t} \\ &= \lambda^{\mathrm{sum}}\big(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}\big)(\bullet) \cdot \exp\{-\lambda^{\mathrm{sum}}\big(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}\big) \cdot (t - \breve{T}_\ell)\}(\bullet), \qquad \text{(B.26)} \end{aligned}$$

which verifies (4.12). For $T_\ell^* = \infty$, following the same proof as that of (B.26), we have that $(\breve{T}_{\ell+1} - \breve{T}_\ell) \mid \mathcal{F}_{\breve{T}_\ell} \sim \mathrm{Exp}(\lambda^{\mathrm{sum}}(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}))$.

Proof of part (iii)**.** For $T_\ell^* < \infty$, if $\breve{T}_{\ell+1} = T_\ell^*$, then (4.9) and (4.10) indicate that $\mathcal{T}_\ell \neq \varnothing$ and $\breve{T}_{\ell+1} - \phi = T_\ell^* - \phi \in \mathcal{T}_\ell$, i.e., $\breve{T}_{\ell+1} = T_{i,k} + \phi$ for some integers $i \in \mathcal{V}$ and $k \geq 1$. This combined with (4.6) gives that $I_{\ell+1} = 0$, and thus $\mathrm{P}(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1})) = 0$ for $i \in \mathcal{V}$. If $\breve{T}_{\ell+1} \in (\breve{T}_\ell, T_\ell^*)$, then for $i \in \mathcal{V}$ and $t \in (\breve{T}_\ell, T_\ell^*)$, using $\breve{t}_\ell = \breve{T}_\ell(\bullet)$, $\mathscr{F}_{\breve{t}_\ell} \subseteq \mathscr{F}_t$, (2.7), (B.20), (B.21), and (B.16), we have

$$\begin{aligned} &\frac{\partial\, \mathrm{P}\big(\breve{T}_{\ell+1} \leq t,\, I_{\ell+1} = i \mid \mathcal{F}_{\breve{T}_\ell}\big)(\bullet)}{\partial\, t} \\ &= \lim_{\Delta \downarrow 0} \Delta^{-1} \mathrm{E}\big\{\mathrm{P}(N_i(t+\Delta) \neq N_i(t) \mid \mathscr{F}_t) \cdot \mathrm{I}(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell)) \mid \mathscr{F}_{\breve{t}_\ell}\big\}(\bullet) \\ &= \mathrm{E}\Big\{ \lim_{\Delta \downarrow 0} \Delta^{-1} \mathrm{P}(N_i(t+\Delta) \neq N_i(t) \mid \mathscr{F}_t) \cdot \mathrm{I}(\boldsymbol{N}(t) = \boldsymbol{N}(\breve{t}_\ell)) \mid \mathscr{F}_{\breve{t}_\ell} \Big\}(\bullet) \qquad \text{(B.27)} \\ &= \lambda_i\big(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}\big)(\bullet) \cdot \exp\{-\lambda^{\mathrm{sum}}\big(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}\big) \cdot (t - \breve{T}_\ell)\}(\bullet), \end{aligned}$$

where the interchange of limit and expectation in (B.27) follows by the dominated convergence theorem and condition A4. This implies that

$$\frac{\partial\, \mathrm{P}\big(\breve{T}_{\ell+1} \leq t,\, I_{\ell+1} = i \mid \mathcal{F}_{\breve{T}_\ell}\big)}{\partial\, t} = \lambda_i\big(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}\big) \exp\{-\lambda^{\mathrm{sum}}\big(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}\big) \cdot (t - \breve{T}_\ell)\}. \qquad \text{(B.28)}$$

Let $t \in (\breve{t}_\ell, t_\ell^*)$, where $t_\ell^* = T_\ell^*(\bullet)$ is defined in preparation (b). Analogously as (B.15), define the realization

$$\circ \;=\; \bullet \cap \{\breve{T}_{\ell+1} = t\}$$

$$= \big\{\{(\breve{T}_0, I_0), \ldots, (\breve{T}_\ell, I_\ell), \breve{T}_{\ell+1}\} = \{(\breve{t}_0, i_0), \ldots, (\breve{t}_\ell, i_\ell), t\}\big\} \in \sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1}).$$

Combining the fact $\breve{T}_{\ell+1}(\circ) = t$, (B.28), and (B.26), for $i \in \mathcal{V}$, we have

$$\begin{aligned}
& \mathrm{P}\big(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1})\big)(\circ) \\
= \; & \frac{\mathrm{P}(I_{\ell+1} = i,\, \breve{T}_{\ell+1} = t \mid \mathcal{F}_{\breve{T}_\ell})(\circ)}{\mathrm{P}(\breve{T}_{\ell+1} = t \mid \mathcal{F}_{\breve{T}_\ell})(\circ)} \\
= \; & \frac{\lambda_i\big(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}\big)(\circ) \cdot \exp\{-\lambda^{\mathrm{sum}}\big(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}\big) \cdot (t - \breve{T}_\ell)\}(\circ)}{\lambda^{\mathrm{sum}}\big(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}\big)(\circ) \cdot \exp\{-\lambda^{\mathrm{sum}}\big(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}\big) \cdot (t - \breve{T}_\ell)\}(\circ)} \\
= \; & \frac{\lambda_i\big(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}\big)(\circ)}{\lambda^{\mathrm{sum}}\big(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}\big)(\circ)},
\end{aligned} \tag{B.29}$$

which proves (4.13). For $T_\ell^* = \infty$, following the same proof as that of (B.29), we have $\mathrm{P}(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{\breve{T}_\ell}, \breve{T}_{\ell+1})) = \lambda_i(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell})/\lambda^{\mathrm{sum}}(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell})$, for $i \in \mathcal{V}$ and $\breve{T}_{\ell+1} \in (\breve{T}_\ell, \infty)$. The proof is completed. ■

B.7 Proofs of Lemmas 5–7, and Theorem 2

The proofs are divided into three parts as follows:

- In Part 1, Definition 5 introduces a class of marked point processes, called *Exponential Marked Point Process (*EMPP*)*, which generalizes the "marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for intensity discontinuities" in Definition 2, and facilitates derivations. Lemmas B.3–B.5 will present the probabilistic properties of the EMPP.
- In Part 2, Definition 6 introduces the notion of *t-truncated* EMPP. Lemmas B.6–B.8 will present the probabilistic properties of the *t-truncated* EMPP.
- In Part 3, by applying the results in Parts 1 and 2 to the "marked point process $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ for intensity discontinuities", we provide the proofs of Lemmas 5–7 and Theorem 2.

B.7.1 Part 1: EMPP and its probabilistic properties

Definition 5 (Exponential Marked Point Process (EMPP**) $(\boldsymbol{T}, \boldsymbol{I})$)** *Let the node set $\mathcal{V} = \{1, 2, \ldots, V\}$ and let the mark set be $\mathcal{V} \cup \{0\}$. Let $T_0 = 0, I_0 = 0$ and $\mathcal{F}_{T_0} = \{\Omega, \varnothing\}$. Let $\{\mathcal{F}_{T_\ell}\}_{\ell \geq 0}$ be the filtration generated by a marked point process $(\boldsymbol{T}, \boldsymbol{I}) = (\{T_\ell\}_{\ell \geq 0}, \{I_\ell\}_{\ell \geq 0}) \in ([0, \infty), \mathcal{V} \cup \{0\})$, where $0 < T_1 < T_2 < \cdots$. We call $(\boldsymbol{T}, \boldsymbol{I})$ an Exponential Marked Point Process* (EMPP)*, if for each integer $\ell \geq 0$, there exist $\mathcal{F}_{T_\ell}$-measurable random variables $\Delta_\ell \in [0, \infty]$ and $\{\lambda_{i,\ell}\}_{i \in \mathcal{V}} \in (0, \infty)$, such that the distributions of $T_{\ell+1}$ and $I_{\ell+1}$ meet the following conditions:*

(i) (*Support of $T_{\ell+1}$*) $\mathrm{P}(T_\ell < T_{\ell+1} \leq T_\ell + \Delta_\ell) = 1$.

(ii) (*Conditional distribution of* $T_{\ell+1}$) *If* $\Delta_\ell < \infty$, *then* $T_{\ell+1}$ *conditional on* $\mathcal{F}_{T_\ell}$ *has a mixed-type distribution, with the* p.m.f.

$$\mathrm{P}(T_{\ell+1} = T_\ell + \Delta_\ell \mid \mathcal{F}_{T_\ell}) \,=\, \exp(-\lambda_\ell^{\mathrm{sum}} \cdot \Delta_\ell) \tag{B.30}$$

at $T_\ell + \Delta_\ell$, *and the* p.d.f.

$$f_{T_{\ell+1}|\mathcal{F}_{T_\ell}}(x \mid \mathcal{F}_{T_\ell}) \,=\, \lambda_\ell^{\mathrm{sum}} \, \exp\{-\lambda_\ell^{\mathrm{sum}} \cdot (x - T_\ell)\}, \tag{B.31}$$

for $x \in (T_\ell, T_\ell + \Delta_\ell)$, *where* $\lambda_\ell^{\mathrm{sum}} = \sum_{i=1}^V \lambda_{i,\ell}$. *If* $\Delta_\ell = \infty$, *then* $(T_{\ell+1} - T_\ell) \mid \mathcal{F}_{T_\ell} \sim \mathrm{Exp}(\lambda_\ell^{\mathrm{sum}})$.

(iii) (*Conditional distribution of* $I_{\ell+1}$) *If* $\Delta_\ell < \infty$, *then* $I_{\ell+1}$ *has the conditional distribution: for* $i \in \mathcal{V}$,

$$\mathrm{P}(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{T_\ell}, T_{\ell+1})) = \begin{cases} 0, & \text{if } T_{\ell+1} - T_\ell = \Delta_\ell, \\ \lambda_{i,\ell}/\lambda_\ell^{\mathrm{sum}}, & \text{if } 0 < T_{\ell+1} - T_\ell < \Delta_\ell. \end{cases} \tag{B.32}$$

If $\Delta_\ell = \infty$, *then* (B.32) *reduces to* $\mathrm{P}(I_{\ell+1} = i \mid \sigma(\mathcal{F}_{T_\ell}, T_{\ell+1})) = \lambda_{i,\ell}/\lambda_\ell^{\mathrm{sum}}$, *for* $i \in \mathcal{V}$ *and* $T_{\ell+1} \in (T_\ell, \infty)$.

Remark 6 *The Exponential Marked Point Process* (EMPP) $(\boldsymbol{T}, \boldsymbol{I})$ *in Definition* 5 *generalizes the class of "marked point process* $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ *for intensity discontinuities" in Definition* 2 *which follow the distribution in Theorem* 1, *according to* $(\boldsymbol{T}, \boldsymbol{I}) = (\breve{\boldsymbol{T}}, \boldsymbol{I})$, $\mathcal{F}_{T_\ell} = \mathcal{F}_{\breve{T}_\ell}$, $\lambda_{i,\ell} = \lambda_i(\breve{T}_\ell \mid \mathcal{F}_{\breve{T}_\ell}) = \exp\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \beta_{j,i}\, x_j(\breve{T}_\ell)\}$, $\Delta_\ell = T_\ell^* - \breve{T}_\ell$. *To prove Lemmas* 5–7, *we will first show probabilistic properties of* EMPP $(\boldsymbol{T}, \boldsymbol{I})$ *which then apply to our* $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ *for intensity discontinuities.*

For ease of exposition, we introduce some notations similar to (4.14) and (4.15).

Duration τ_ℓ between two consecutive time points:

$$\tau_\ell = T_\ell - T_{\ell-1}, \quad \ell \geq 1. \tag{B.33}$$

Event counts $M_{i,\ell}$ at node $i \in \mathcal{V}$:

$$M_{i,0} = 0, \quad M_{i,\ell} = \sum_{k=1}^{\ell} \mathrm{I}(I_k = i), \quad \ell \geq 1. \tag{B.34}$$

Lemma B.3 presents the conditional expectation and variance of τ_k and $\mathrm{I}(I_k = i)$.

Lemma B.3 (Conditional expectation and variance related to an EMPP $(\boldsymbol{T}, \boldsymbol{I})$**)** *Consider an* EMPP $(\boldsymbol{T}, \boldsymbol{I})$ *in Definition* 5. *For integers* $k \geq 1$ *and* $i \in \mathcal{V}$, *we have*

$$\mathrm{var}\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\tau_k \mid \mathcal{F}_{T_{k-1}}\big\} = \mathrm{E}\big\{\mathrm{I}(I_k = i) \mid \mathcal{F}_{T_{k-1}}\big\} = \lambda_{i,k-1}\mathrm{E}(\tau_k \mid \mathcal{F}_{T_{k-1}}). \tag{B.35}$$

Proof: For $\Delta_{k-1} < \infty$, the conditional distribution of τ_k is given by (B.30) and (B.31), and the conditional distribution of $\mathrm{I}(I_k = i)$ is given by (B.32). Direct calculations yield the following (B.36)–(B.38):

$$\begin{aligned}
&\mathrm{E}(\lambda_{i,k-1}\,\tau_k \mid \mathcal{F}_{T_{k-1}}) = \lambda_{i,k-1}\,\mathrm{E}(\tau_k \mid \mathcal{F}_{T_{k-1}}) \\
&\quad = \lambda_{i,k-1}\Big\{\Delta_{k-1}\mathrm{P}(\tau_k = \Delta_{k-1} \mid \mathcal{F}_{T_{k-1}}) + \int_0^{\Delta_{k-1}} x\, f_{\tau_k|\mathcal{F}_{T_{k-1}}}(x \mid \mathcal{F}_{T_{k-1}})\,\mathrm{d}x\Big\} \\
&\quad = \frac{\lambda_{i,k-1}}{\lambda_{k-1}^{\mathrm{sum}}} \cdot (1 - e^{-\lambda_{k-1}^{\mathrm{sum}} \cdot \Delta_{k-1}}) \\
&\quad = \mathrm{E}\big\{\mathrm{I}(I_k = i) \mid \mathcal{F}_{T_{k-1}}\big\}, \qquad \text{(B.36)}
\end{aligned}$$

together with

$$\begin{aligned}
&\mathrm{E}\big\{\lambda_{i,k-1}\,\tau_k\,\mathrm{I}(I_k = i) \mid \mathcal{F}_{T_{k-1}}\big\} \\
= \;& \lambda_{i,k-1}\int_0^{\Delta_{k-1}} \frac{\lambda_{i,k-1}}{\lambda_{k-1}^{\mathrm{sum}}}\, x\, f_{\tau_k|\mathcal{F}_{T_{k-1}}}(x \mid \mathcal{F}_{T_{k-1}})\,\mathrm{d}x \\
= \;& \frac{\lambda_{i,k-1}^2}{(\lambda_{k-1}^{\mathrm{sum}})^2} \cdot \big\{1 - (1 + \lambda_{k-1}^{\mathrm{sum}} \cdot \Delta_{k-1})\, e^{-\lambda_{k-1}^{\mathrm{sum}} \cdot \Delta_{k-1}}\big\}, \qquad \text{(B.37)}
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{E}(\lambda_{i,k-1}^2\,\tau_k^2 \mid \mathcal{F}_{T_{k-1}}) \\
= \;& \lambda_{i,k-1}^2 \cdot \Big\{\Delta_{k-1}^2\, e^{-\lambda_{k-1}^{\mathrm{sum}} \cdot \Delta_{k-1}} + \int_0^{\Delta_{k-1}} x^2\, \lambda_{k-1}^{\mathrm{sum}} \cdot e^{-\lambda_{k-1}^{\mathrm{sum}} \cdot x}\,\mathrm{d}x\Big\} \\
= \;& \frac{\lambda_{i,k-1}^2}{(\lambda_{k-1}^{\mathrm{sum}})^2} \cdot \{2 - (2\,\lambda_{k-1}^{\mathrm{sum}} \cdot \Delta_{k-1} + 2)\, e^{-\lambda_{k-1}^{\mathrm{sum}} \cdot \Delta_{k-1}}\}. \qquad \text{(B.38)}
\end{aligned}$$

Combining (B.37) and (B.38), we have

$$\mathrm{E}\big\{\lambda_{i,k-1}^2\,\tau_k^2 - 2\,\lambda_{i,k-1}\,\tau_k\,\mathrm{I}(I_k = i) \mid \mathcal{F}_{T_{k-1}}\big\} = 0. \qquad \text{(B.39)}$$

For $\Delta_{k-1} = \infty$, the conditional distributions of τ_k and $\mathrm{I}(I_k = i)$ are given in parts (ii) and (iii) of Definition 5. Using this with the similar calculations as in (B.36)–(B.39), we verify that (B.36) and (B.39) also hold for $\Delta_{k-1} = \infty$. By (B.36) and (B.39), we have

$$\begin{aligned}
&\mathrm{var}\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\,\tau_k \mid \mathcal{F}_{T_{k-1}}\big\} \\
= \;& \mathrm{E}\Big[\big\{\mathrm{I}(I_k = i) - \lambda_{i,k-1}\tau_k\big\}^2 \mid \mathcal{F}_{T_{k-1}}\Big] \\
= \;& \mathrm{E}\big\{\mathrm{I}(I_k = i) \mid \mathcal{F}_{T_{k-1}}\big\}. \qquad \text{(B.40)}
\end{aligned}$$

Combining (B.36) and (B.40) completes the proof of (B.35). ■

Lemma B.4 follows from Lemma B.3.

Lemma B.4 (Martingale property for EMPP**)** *In an* EMPP $(\boldsymbol{T}, \boldsymbol{I})$*, for each* $i \in \mathcal{V}$*, the random process*

$$\Big\{ M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\, \tau_k \Big\}_{\ell \geq 1}$$

is a martingale with respect to $\{\mathcal{F}_{T_\ell}\}_{\ell \geq 1}$*.*

Proof: By (B.35), for each integer $k \geq 1$, we have

$$\mathrm{E}\big\{ \mathrm{I}(I_k = i) - \lambda_{i,k-1}\, \tau_k \mid \mathcal{F}_{T_{k-1}} \big\} \;=\; 0. \tag{B.41}$$

This completes the proof. ■

Lemma B.5 derives the variance of the martingale.

Lemma B.5 (Variance of the martingale $\{M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\, \tau_k\}_{\ell \geq 1}$**)** *In an* EMPP $(\boldsymbol{T}, \boldsymbol{I})$*, for integers* $i \in \mathcal{V}$ *and* $\ell \geq 1$*, we have* $\mathrm{var}(M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\, \tau_k) = \mathrm{E}(M_{i,\ell})$*.*

Proof: Using (B.41), for any indices r and k such that $1 \leq r < k$, we have

$$\begin{aligned} & \mathrm{E}\Big[\big\{ \mathrm{I}(I_r = i) - \lambda_{i,r-1}\, \tau_r \big\} \big\{ \mathrm{I}(I_k = i) - \lambda_{i,k-1}\, \tau_k \big\} \Big] \\ = \;& \mathrm{E}\Big[\big\{ \mathrm{I}(I_r = i) - \lambda_{i,r-1}\, \tau_r \big\} \mathrm{E}\big\{ \mathrm{I}(I_k = i) - \lambda_{i,k-1}\, \tau_k \,\big|\, \mathcal{F}_{T_{k-1}} \big\} \Big] \\ = \;& 0. \end{aligned} \tag{B.42}$$

For any index $\ell \geq 1$, uses of (B.42), (B.35), and (B.40) give that

$$\begin{aligned} & \mathrm{E}\Big\{ \Big(M_{i,\ell} - \sum_{k=1}^{\ell} \lambda_{i,k-1}\, \tau_k \Big)^2 \Big\} \\ = \;& \sum_{k=1}^{\ell} \mathrm{E}\Big[\big\{ \mathrm{I}(I_k = i) - \lambda_{i,k-1}\, \tau_k \big\}^2 \Big] \\ & + \sum_{1 \leq k \neq r \leq \ell} \sum \mathrm{E}\Big[\big\{ \mathrm{I}(I_r = i) - \lambda_{i,r-1}\, \tau_r \big\} \big\{ \mathrm{I}(I_k = i) - \lambda_{i,k-1}\, \tau_k \big\} \Big] \\ = \;& \sum_{k=1}^{\ell} \mathrm{E}\big\{ \mathrm{I}(I_k = i) \big\} = \mathrm{E}(M_{i,\ell}). \end{aligned}$$

This completes the proof. ■

B.7.2 Part 2: t-truncated EMPP and its probabilistic properties

We next derive the probabilistic results of EMPP $(\boldsymbol{T}, \boldsymbol{I})$ when the point process $\boldsymbol{T} = \{T_0, T_1, \ldots\}$ reaches a pre-specified time point $t \in (0, \infty)$. Definition 6 introduces the notion of *t-truncated* EMPP.

Definition 6 (t-truncated EMPP) *Consider an* EMPP $(\boldsymbol{T}, \boldsymbol{I}) = (\{T_\ell\}_{\ell\geq 0}, \{I_\ell\}_{\ell\geq 0})$. *Let* $t \in (0,\infty)$ *be a given deterministic time point. Define the marked point process* $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]}) = (\{T_\ell^{[t]}\}_{\ell\geq 0}, \{I_\ell^{[t]}\}_{\ell\geq 0})$, *with*

$$T_\ell^{[t]} = T_\ell \wedge t, \qquad I_\ell^{[t]} = I_\ell \, \mathrm{I}(T_\ell \leq t). \tag{B.43}$$

We call $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ *the* t*-truncated* EMPP *from* $(\boldsymbol{T}, \boldsymbol{I})$.

Lemma B.6 states that any double sequence $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ defined in (B.43) is an EMPP.

Lemma B.6 *Let* $(\boldsymbol{T}, \boldsymbol{I}) = (\{T_\ell\}_{\ell\geq 0}, \{I_\ell\}_{\ell\geq 0})$ *be an* EMPP *defined in Definition* 5 *associated with* $\{\lambda_{i,\ell}\}_{i\in\mathcal{V}}$ *and* Δ_ℓ *in* (B.30)–(B.32). *For* $t \in (0,\infty)$, *let* $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ *be the corresponding* t*-truncated* EMPP *as in Definition* 6. *Then the probability distributions of* $\boldsymbol{T}^{[t]}$ *and* $\boldsymbol{I}^{[t]}$) *meet the conditions* (i) *and* (ii) *in Definition* 5 *associated with* $\{\lambda_{i,\ell}\}_{i\in\mathcal{V}}$ *and* $\Delta_{\ell,t}$ (*instead of* Δ_ℓ), *where*

$$\Delta_{\ell,t} \;=\; \Delta_\ell \wedge (t - T_\ell^{[t]}), \tag{B.44}$$

and thus $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ *is an* EMPP.

Proof: To prove Lemma B.6, it suffices to show that for each integer $\ell \geq 0$, $(T_{\ell+1}^{[t]}, I_{\ell+1}^{[t]})$ follows the conditional distribution in (B.30)–(B.32) with T_ℓ and Δ_ℓ replaced by $T_\ell^{[t]}$ and $\Delta_{\ell,t}$. We proceed by cases of $T_\ell^{[t]}$.

Case (i) $T_\ell^{[t]} < t - \Delta_\ell$.

From (B.43) and (B.44), we observe that $T_\ell^{[t]} = T_\ell$ and $\Delta_{\ell,t} = \Delta_\ell$. Also, from (B.30) and (B.31), we know that $T_{\ell+1} \leq T_\ell + \Delta_\ell \leq t$. Thus, $(T_{\ell+1}^{[t]},\ I_{\ell+1}^{[t]}) = (T_{\ell+1} \wedge t,\ I_{\ell+1}\mathrm{I}(T_{\ell+1} \leq t)) = (T_{\ell+1},\ I_{\ell+1})$, follows the conditional distribution in (B.30)–(B.32).

Case (ii) $t - \Delta_\ell \leq T_\ell^{[t]} < t$.

By (B.43) and (B.44), we have $T_\ell^{[t]} = T_\ell$ and $\Delta_{\ell,t} = t - T_\ell$. Using (B.30) and (B.31), we obtain

$$\mathrm{P}((T_{\ell+1} \wedge t) = t \mid \mathcal{F}_{T_\ell}) \;=\; \exp\{-\lambda_\ell^{\mathrm{sum}} \cdot (t - T_\ell)\},$$

and

$$f_{(T_{\ell+1}\wedge t)|\mathcal{F}_{T_\ell}}(x \mid \mathcal{F}_{T_\ell}) \;=\; \lambda_\ell^{\mathrm{sum}} \exp\{-\lambda_\ell^{\mathrm{sum}} \cdot (x - T_\ell)\}, \quad \text{for } x \in (T_\ell, t).$$

This indicates that $T_{\ell+1}^{[t]} = T_{\ell+1} \wedge t$ follows the distribution in (B.30) and (B.31) with T_ℓ and Δ_ℓ replaced by $T_\ell^{[t]} = T_\ell$ and $\Delta_{\ell,t} = t - T_\ell$. Also, by checking (B.32), we have that $I_{\ell+1}^{[t]} = I_{\ell+1}\mathrm{I}(T_{\ell+1} \leq t)$ conditional on $T_{\ell+1}^{[t]}$, follows the distribution in (B.32).

Case (iii) $T_\ell^{[t]} = t$.

By (B.44), we know $\Delta_{\ell,t} = 0$. Then $(T_{\ell+1}^{[t]}, I_{\ell+1}^{[t]}) = (t, 0)$ follows the distribution in (B.30)–(B.32) with T_ℓ and Δ_ℓ replaced by $T_\ell^{[t]} = t$ and $\Delta_{\ell,t} = 0$.

Summarizing the above three cases completes the proof. ■

Similar to (B.33) and (B.34), we use notations $\tau_\ell^{[t]}$ and $M_{i,\ell}^{[t]}$ for the duration and event counts respectively of the t-truncated EMPP $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$, i.e.,

$$\tau_\ell^{[t]} = T_\ell^{[t]} - T_{\ell-1}^{[t]}, \quad \ell \geq 1, \tag{B.45}$$

$$M_{i,0}^{[t]} = 0, \qquad M_{i,\ell}^{[t]} = \sum_{k=1}^{\ell} \mathrm{I}(I_k^{[t]} = i), \quad \ell \geq 1. \tag{B.46}$$

Define $M_{i,\infty}^{[t]} = \lim_{\ell\to\infty} M_{i,\ell}^{[t]}$. Let $L_t = \sum_{\ell=1}^{\infty} \mathrm{I}(T_\ell \leq t)$. From (B.43),

$$M_{i,\infty}^{[t]} = \sum_{k=1}^{\infty} \mathrm{I}(I_k^{[t]} = i) = \sum_{k=1}^{L_t} \mathrm{I}(I_k = i) = M_{i,L_t} \tag{B.47}$$

is the total event counts of node i in the time interval $[0, t]$. Lemma B.7 shows an upper bound for $\mathrm{E}(M_{i,\infty}^{[t]})$.

Lemma B.7 (Upper bound for $\mathrm{E}(M_{i,\infty}^{[t]})$**)** *Let* $(\boldsymbol{T}, \boldsymbol{I})$ *be an* EMPP*, and* $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ *be the corresponding* t*-truncated* EMPP *from* $(\boldsymbol{T}, \boldsymbol{I})$ *as in Definition* 6*. Assume that* $\sup_{i\in\mathcal{V},\ell\geq 0} \lambda_{i,\ell} \leq c$ *for a constant* $c \in (0, \infty)$*. Then*

$$\mathrm{E}(M_{i,\infty}^{[t]}) \;\leq\; c\,t. \tag{B.48}$$

Proof: Lemma B.6 verified that $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ is an EMPP. Applying Lemma B.4 to $(\boldsymbol{T}^{[t]}, \boldsymbol{I}^{[t]})$ gives that for each integer $\ell \geq 1$, $\mathrm{E}(M_{i,\ell}^{[t]}) = \mathrm{E}(\sum_{k=1}^{\ell} \lambda_{i,k-1}\tau_k^{[t]})$. Note that $M_{i,\ell}^{[t]}$ and $\sum_{k=1}^{\ell} \lambda_{i,k-1}\tau_k^{[t]}$ are monotonically increasing in ℓ. By the monotone convergence theorem, we obtain

$$\begin{aligned} & \mathrm{E}(M_{i,\infty}^{[t]}) = \lim_{\ell\to\infty} \mathrm{E}(M_{i,\ell}^{[t]}) \\ = \;& \lim_{\ell\to\infty} \mathrm{E}\Big(\sum_{k=1}^{\ell} \lambda_{i,k-1}\,\tau_k^{[t]}\Big) = \mathrm{E}\Big(\sum_{k=1}^{\infty} \lambda_{i,k-1}\,\tau_k^{[t]}\Big) \\ \leq \;& c\,\mathrm{E}\Big(\sum_{k=1}^{\infty} \tau_k^{[t]}\Big) \leq c\,t. \end{aligned} \tag{B.49}$$

This completes the proof. ■

Lemma B.8 (Upper bound for $\mathrm{var}\{\sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i) - \sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\}$**)** *Let* $(\boldsymbol{T},\boldsymbol{I})$ *be an* EMPP*, and* $(\boldsymbol{T}^{[t]},\boldsymbol{I}^{[t]})$ *be the corresponding* t*-truncated* EMPP *from* $(\boldsymbol{T},\boldsymbol{I})$ *as in Definition* 6*. Assume that* $\sup_{i\in\mathcal{V},\ell\geq 0}\lambda_{i,\ell}\leq c$ *for a constant* $c\in(0,\infty)$*. Let* $\{X_\ell\}_{\ell\geq 0}$ *be a sequence of random variables, such that* $X_\ell\geq 0$ *is measurable with respect to* $\mathcal{F}_{T_\ell}$ *for each integer* $\ell\geq 0$*, and* $\sup_{\ell\geq 0}X_\ell\leq c_2$ a.s. *for a constant* $c_2\in(0,\infty)$*. Then*

$$\mathrm{E}\Big\{\sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i) - \sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big\} = 0, \tag{B.50}$$

and

$$\begin{aligned}&\mathrm{var}\Big\{\sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i) - \sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big\}\\ &\qquad = \mathrm{E}\Big\{\sum_{k=1}^{\infty} X_{k-1}^2\,\mathrm{I}(I_k^{[t]}=i)\Big\} \leq c\,c_2^2\,t.\end{aligned} \tag{B.51}$$

Proof: An argument similar to (B.49) gives that

$$\begin{aligned}&\mathrm{E}\Big\{\sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i)\Big\} = \lim_{\ell\to\infty}\mathrm{E}\Big\{\sum_{k=1}^{\ell} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i)\Big\}\\ =\ &\lim_{\ell\to\infty}\mathrm{E}\Big(\sum_{k=1}^{\ell} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big) = \mathrm{E}\Big(\sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big),\end{aligned}$$

which proves (B.50). Using the similar proof as that of Lemma B.5, we have

$$\mathrm{E}\Big\{\Big(\sum_{k=1}^{\ell} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i) - \sum_{k=1}^{\ell} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big)^2\Big\} = \mathrm{E}\Big\{\sum_{k=1}^{\ell} X_{k-1}^2\,\mathrm{I}(I_k^{[t]}=i)\Big\}. \tag{B.52}$$

Note that $\lim_{\ell\to\infty}\sum_{k=1}^{\ell} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i) = \sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i)$, a.s., and

$$\lim_{\ell\to\infty}\sum_{k=1}^{\ell} X_{k-1}\,\lambda_{i,k-1}\tau_k^{[t]} = \sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]} \leq \sum_{k=1}^{\infty} c\,c_2\,\tau_k^{[t]} = c\,c_2\,t < \infty, \quad \text{a.s..}$$

It follows that

$$\begin{aligned}&\sum_{k=1}^{\ell} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i) - \sum_{k=1}^{\ell} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\\ \overset{\text{a.s.}}{\to}\ &\sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i) - \sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}, \quad \text{as } \ell\to\infty.\end{aligned} \tag{B.53}$$

Also, for each integer $\ell\geq 1$, we have

$$\Big\{\sum_{k=1}^{\ell} X_{k-1}\,\mathrm{I}(I_k^{[t]}=i) - \sum_{k=1}^{\ell} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}\Big\}^2$$

$$\leq \Big\{ \sum_{k=1}^{\ell} X_{k-1}\, \mathrm{I}(I_k^{[t]} = i) \Big\}^2 + \Big(\sum_{k=1}^{\ell} X_{k-1}\, \lambda_{i,k-1}\, \tau_k^{[t]} \Big)^2$$
$$\leq (c_2\, M_{i,\infty}^{[t]})^2 + (c\, c_2\, t)^2. \tag{B.54}$$

By (B.52), (B.53), (B.54) and the dominated convergence theorem, we have

$$\begin{aligned} & \mathrm{E}\Big\{ \Big(\sum_{k=1}^{\infty} X_{k-1}\, \mathrm{I}(I_k^{[t]} = i) - \sum_{k=1}^{\infty} X_{k-1}\, \lambda_{i,k-1}\, \tau_k^{[t]} \Big)^2 \Big\} \\ = \;& \mathrm{E}\Big\{ \sum_{k=1}^{\infty} X_{k-1}^2\, \mathrm{I}(I_k^{[t]} = i) \Big\} \\ \leq \;& c_2^2\, \mathrm{E}(M_{i,\infty}^{[t]}) \;\leq\; c\, c_2^2\, t, \end{aligned}$$

where the last inequality is from (B.48). Thus, (B.51) is proved. ■

B.7.3 Part 3: Proofs of Lemmas 5–7 and Theorem 2

Remark 6 verified that the *marked point process* $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ *for intensity discontinuities* is an EMPP. Let $(\breve{\boldsymbol{T}}^{[t]}, \boldsymbol{I}^{[t]})$ be the *t-truncated* EMPP from $(\breve{\boldsymbol{T}}, \boldsymbol{I})$ as in Definition 6, i.e., for each integer $\ell \geq 1$, $(\breve{T}_\ell^{[t]}, I_\ell^{[t]}) = (\breve{T}_\ell \wedge t\,,\, I_\ell\, \mathrm{I}(\breve{T}_\ell \leq t))$. Recall that $M_{i,\ell}^{[t]}$ defined in (B.46) is the event counts corresponding to $(\breve{\boldsymbol{T}}^{[t]}, \boldsymbol{I}^{[t]})$, and $M_{i,\infty}^{[t]} = \lim_{\ell\to\infty} M_{i,\ell}^{[t]}$. Recall L_t defined in (4.18). Using (4.8), (4.15) and (B.47), $N_i(t)$ has the equivalent expressions:

$$N_i(t) \;=\; M_{i,L_t} \;=\; \sum_{k=1}^{L_t} \mathrm{I}(I_k = i) \;=\; M_{i,\infty}^{[t]}. \tag{B.55}$$

Following (B.45), let $\tau_\ell^{[t]} = \breve{T}_\ell^{[t]} - \breve{T}_{\ell-1}^{[t]}$ be the duration between two consecutive time points $\breve{T}_{\ell-1}^{[t]}$ and $\breve{T}_\ell^{[t]}$. It is easy to check that this $\tau_\ell^{[t]}$ is identical to that defined in (4.19). Using (4.16) and (4.19), the integral $\int_0^t \lambda_i(u \mid \mathscr{F}_u)\, \mathrm{d}u$ has the following equivalent expressions:

$$\int_0^t \lambda_i(u \mid \mathscr{F}_u)\, \mathrm{d}u \;=\; \sum_{k=1}^{L_t+1} \lambda_{i,k-1}\, \tau_k^{[t]} \;=\; \sum_{k=1}^{\infty} \lambda_{i,k-1}\, \tau_k^{[t]}. \tag{B.56}$$

After clarifying the facts above, we next prove Lemmas 5–7 and Theorem 2.

B.7.4 Proof of Lemma 5

Lemma 5 is directly obtained by Lemmas B.3 and B.5. ■

B.7.5 Proof of Lemma 6

Lemma 6 is directly obtained by Lemma B.4. ■

B.7.6 Proof of Lemma 7

From (B.43), (B.46) and (B.56), we have $\sum_{k=1}^{L_t} X_{k-1}\,\mathrm{I}(I_k = i) = \sum_{k=1}^{\infty} X_{k-1}\,\mathrm{I}(I_k^{[t]} = i)$, and $\sum_{k=1}^{L_t+1} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]} = \sum_{k=1}^{\infty} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}$. Then all the results in Lemma 7 could be directly obtained by Lemma B.8. ■

B.7.7 Proof of Theorem 2

Recall that $N_i(t) = M_{i,L_t}$ in (B.55) and $\int_0^t \lambda_i(u \mid \mathscr{F}_u)\,\mathrm{d}u = \sum_{k=1}^{L_t+1} \lambda_{i,k-1}\,\tau_k^{[t]}$ in (B.56). For each integer $\ell \geq 0$, let $X_\ell = x(\breve{T}_\ell)$. We further have $\int_0^t x(u-)\,\mathrm{d}N_i(u) = \sum_{k=1}^{L_t} X_{k-1}\,\mathrm{I}(I_k = i)$ and $\int_0^t x(u)\,\lambda_i(u \mid \mathscr{F}_u)\,\mathrm{d}u = \sum_{k=1}^{L_t+1} X_{k-1}\,\lambda_{i,k-1}\,\tau_k^{[t]}$. Then (4.22) and (4.24) in Theorem 2 are directly implied by (4.20) and (4.21) respectively in Lemma 7. The finiteness of $\boldsymbol{N}(t)$ in (4.23) is directly implied by (4.22). ■

B.8 Proof of non-stationarity of $\boldsymbol{N}(t)$ in Section 4.2

Lemma B.9 ($\boldsymbol{N}(t)$ is not strict-sense stationary) *Assume conditions* A1, A2, A3, A4, A5 *and* A8 *in Appendix* B. *Then there exists some node $i_0 \in \mathcal{V}$, such that $N_{i_0}(t)$ is not strict-sense stationary. Hence, the multivariate counting process $\boldsymbol{N}(t)$ is not strict-sense stationary.*

Before proving Lemma B.9, we first present Lemma B.10.

Lemma B.10 (Probabilistic inequalities for $\lambda_i(t \mid \mathscr{F}_t)$ in our (3.1)) *Assume conditions* A1, A2, A3, A4, A5 *and* A8 *in Appendix* B. *Let $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ be the conditional intensity functions defined in* (3.1). *Then for any distinct $i, j \in \mathcal{V}$, there exist constants $c_0, c_1, c_2, c_3 \in (0, \infty)$, such that for any $t \in (0, \phi)$, the following assertions hold:*

$$\mathrm{P}(\lambda_i(t \mid \mathscr{F}_t) = \exp(\beta_{0;i})) \;\geq\; \exp(-c_0\,t), \tag{B.57}$$

$$\mathrm{P}\big(\lambda_i(t \mid \mathscr{F}_t) = \exp\{\beta_{0;i} + \beta_{j,i} \cdot g(1/\phi)\}\big) \;\geq\; c_1 \cdot \exp(-c_2\,t) \cdot \{1 - \exp(-c_3\,t)\}. \tag{B.58}$$

Proof: For any $t \in (0, \phi)$, (3.1), (3.5) and (3.6) indicate that $\{\boldsymbol{N}(t) = \boldsymbol{0}\} \subseteq \{\lambda_i(t \mid \mathscr{F}_t) = \exp(\beta_{0;i})\}$ and $\{N_j(t) = 1 \text{ and } N_k(t) = 0 \text{ for all } k \in \mathcal{V} \setminus j\} \subseteq \{\lambda_i(t \mid \mathscr{F}_t) = \exp(\beta_{0;i} + \beta_{j,i} \cdot g(1/\phi))\}$. To verify (B.57) and (B.58), it suffices to show that for any $t \in (0, \phi)$,

$$\begin{aligned} &\mathrm{P}(\boldsymbol{N}(t) = \boldsymbol{0}) \;=\; \exp(-c_0\,t), \\ &\mathrm{P}(N_j(t) = 1, \text{ and } N_k(t) = 0 \text{ for all } k \in \mathcal{V} \setminus j) \end{aligned} \tag{B.59}$$

$$\geq\ c_1 \cdot \exp(-c_2\, t) \cdot \{1 - \exp(-c_3\, t)\}. \tag{B.60}$$

Recall that in Theorem 1, we have the following facts: $\breve{T}_0 = 0$, $\mathcal{F}_{\breve{T}_0} = \mathscr{F}_0 = \{\Omega, \varnothing\}$, $\mathcal{T}_0 = \varnothing$ and $T_0^* = \infty$. By Theorem 1(ii) and the fact that $T_0^* = \infty$, we have that $\breve{T}_1$ is a continuous random variable with the p.d.f.:

$$f_{\breve{T}_1}(x) = \lambda^{\rm sum}(0) \exp\{-\lambda^{\rm sum}(0) \cdot x\}, \quad x \in (0, \infty). \tag{B.61}$$

On the other hand, it is easy to verify from (4.5) that the first discontinuity point $\breve{T}_1 = \min\{T_{j,\ell} : j \in \mathcal{V}, \ell \geq 1\}$ is exactly the first event time point, i.e.,

$$\boldsymbol{N}(t) = \mathbf{0} \quad \text{ if and only if } \quad 0 \leq t < \breve{T}_1. \tag{B.62}$$

Combining (B.61) and (B.62), we have $\mathrm{P}(\boldsymbol{N}(t) = \mathbf{0}) = \exp\{-\lambda^{\rm sum}(0) \cdot t\}$. This proves (B.59) with a constant $c_0 = \lambda^{\rm sum}(0) = \sum_{i=1}^V \exp(\beta_{0;i})$.

Similarly, parts (ii) and (iii) of Theorem 3 yield that

$$\begin{aligned}
&\mathrm{P}\big(N_j(t) = 1, \text{ and } N_k(t) = 0 \text{ for all } k \in \mathcal{V} \setminus j\big) \\
=\ & \int_0^t \lambda^{\rm sum}(0) \exp\{-\lambda^{\rm sum}(0) \cdot x\} \cdot \frac{\lambda_j(0)}{\lambda^{\rm sum}(0)} \\
&\qquad \times \Big[1 - \int_x^t \lambda^{\rm sum}(x \mid \mathscr{F}_x) \exp\{-\lambda^{\rm sum}(x \mid \mathscr{F}_x) \cdot (u - x)\}\, \mathrm{d}u\Big]\, \mathrm{d}x \\
=\ & \begin{cases} \lambda_1 \cdot \{\exp(-\lambda_3 \cdot t) - \exp(-\lambda_2 \cdot t)\}/(\lambda_2 - \lambda_3), & \text{if } \lambda_2 \neq \lambda_3, \\ \lambda_1 \cdot t \cdot \exp(-\lambda_2 \cdot t), & \text{if } \lambda_2 = \lambda_3, \end{cases}
\end{aligned} \tag{B.63}$$

where $\lambda_1 = \lambda_j(0) = \exp(\beta_{0;j})$, $\lambda_2 = \lambda^{\rm sum}(0) = \sum_{i=1}^V \exp(\beta_{0;i})$, and $\lambda^{\rm sum}(x \mid \mathscr{F}_x)$ reduces to $\lambda_3 = \sum_{i=1}^V \exp\{\beta_{0;i} + \beta_{j,i}\, g(1/\phi)\}$.

For $\lambda_2 \neq \lambda_3$, if $\lambda_2 - \lambda_3 = \delta > 0$, then (B.63) gives that

$$\mathrm{P}\big(N_j(t) = 1, \text{ and } N_k(t) = 0 \text{ for all } k \in \mathcal{V} \setminus j\big) \ = \ \lambda_1/\delta \cdot \exp(-\lambda_3 \cdot t)\{1 - \exp(-\delta\, t)\}.$$

Thus, (B.60) holds with $c_1 = \lambda_1/\delta$, $c_2 = \lambda_3$ and $c_3 = \delta$. Due to the symmetry between λ_2 and λ_3 in (B.63), the similar argument holds when $\lambda_2 - \lambda_3 < 0$.

For $\lambda_2 = \lambda_3$, (B.63) yields that

$$\begin{aligned}
&\mathrm{P}\big(N_j(t) = 1, \text{ and } N_k(t) = 0 \text{ for all } k \in \mathcal{V} \setminus j\big) \\
=\ & \lambda_1 \cdot t \cdot \exp(-\lambda_2 \cdot t) \geq \lambda_1 \cdot \exp(-\lambda_2 \cdot t) \cdot \{1 - \exp(-t)\}.
\end{aligned}$$

Hence, (B.60) holds with $c_1 = \lambda_1$, $c_2 = \lambda_2$ and $c_3 = 1$. This completes the proof. ■

Next we prove Lemma B.9 using a proof by contradiction. For any $(j_0, i_0) \in \mathcal{E}$ with $\beta_{j_0,i_0} \neq 0$, if $\{N_{i_0}(t)\}_{t\geq 0}$ is stationary, then property (P1) implies that the conditional intensity function $\lambda_{i_0}(t \mid \mathscr{F}_t)$ has the same distribution for all $t \in (0, \phi)$. Thus, for the two possible values $\exp(\beta_{0;i_0})$ and $\exp\{\beta_{0;i_0} + \beta_{j_0,i_0} \cdot g(1/\phi)\}$ of $\lambda_{i_0}(t \mid \mathscr{F}_t)$, there exist some constants $c_4, c_5 \in [0, 1]$, such that $\mathrm{P}(\lambda_{i_0}(t \mid \mathscr{F}_t) = \exp(\beta_{0;i_0})) \equiv c_4$ and $\mathrm{P}(\lambda_{i_0}(t \mid \mathscr{F}_t) = \exp\{\beta_{0;i_0} + \beta_{j_0,i_0} \cdot g(1/\phi)\}) \equiv c_5$ hold for any $t \in (0, \phi)$. Combining this with (B.57) and (B.58), for any $t \in (0, \phi)$, we have

$$\begin{aligned}
& \mathrm{P}\big(\lambda_{i_0}(t \mid \mathscr{F}_t) = \exp(\beta_{0;i_0}),\ \text{or } \lambda_{i_0}(t \mid \mathscr{F}_t) = \exp\{\beta_{0;i_0} + \beta_{j_0,i_0} \cdot g(1/\phi)\}\big) \\
\equiv\ & c_4 + c_5 \\
\geq\ & \sup_{t\in(0,\phi)} \big\{ \exp(-c_0\, t)\big\} + \sup_{t\in(0,\phi)} \big\{c_1 \cdot \exp(-c_2\, t) \cdot \{1 - \exp(-c_3\, t)\}\big\} \\
\geq\ & 1 + c_1 \cdot \exp(-c_2 \cdot \phi/2) \cdot \{1 - \exp(-c_3 \cdot \phi/2)\} > 1,
\end{aligned}$$

which obviously contradicts. This completes the proof. ■

B.9 Proof of Lemma 8

From (4.25), for each integer $\ell \geq 1$, we have $R_\ell = \min(\mathcal{U}_\ell)$, where $\mathcal{U}_\ell = \{t \geq R_{\ell-1} + \phi : \boldsymbol{N}((t-\phi, t]) = \mathbf{0}\}$. Thus, R_ℓ exists if and only if the following two conditions hold:

(i) $\mathcal{U}_\ell \neq \varnothing$. (Since $\mathcal{U}_\ell$ is bounded below, this indicates $\inf(\mathcal{U}_\ell)$ exists.)

(ii) $\inf(\mathcal{U}_\ell) \in \mathcal{U}_\ell$. (This indicates $\min(\mathcal{U}_\ell) = \inf(\mathcal{U}_\ell)$.)

We start by proving the existence of R_1. We first prove that condition (i) holds with probability one for $\mathcal{U}_1$. Note that $\mathcal{U}_1 \neq \varnothing$ if and only if there exists $t \geq \phi$ such that $\boldsymbol{N}((t-\phi, t]) = \mathbf{0}$. It suffices to show that

$$\mathrm{P}\Big(\bigcup_{t\geq\phi} \big\{\boldsymbol{N}((t-\phi, t]) = \mathbf{0}\big\}\Big) = 1. \tag{B.64}$$

By condition A5, there exists some constant $c \in (0, \infty)$ such that $\lambda^{\mathrm{sum}}(t \mid \mathscr{F}_t) \leq c$. This together with (B.10) gives that $\lambda^{\mathrm{sum}}(t; \mathscr{F}_s) \leq c$. Then by (B.9), for $t > s \geq 0$, we have

$$\begin{aligned}
\mathrm{P}\big(\boldsymbol{N}(t) = \boldsymbol{N}(s) \mid \mathscr{F}_s\big) &= \exp\Big\{ - \int_s^t \lambda^{\mathrm{sum}}(u; \mathscr{F}_s)\, \mathrm{d}u\Big\} \\
&\geq \exp\{-c \cdot (t - s)\}.
\end{aligned} \tag{B.65}$$

For each integer $k \geq 1$, plugging $s = (k-1)\,\phi$ and $t = k\,\phi$ into (B.65), we obtain

$$\mathrm{P}(A_k^* \mid \mathscr{F}_{(k-1)\,\phi}) \ \geq\ \exp(-c\,\phi), \tag{B.66}$$

where the event

$$A_k^* \;=\; \big\{\boldsymbol{N}\big(((k-1)\,\phi, k\,\phi]\big) = \boldsymbol{0}\big\} \;=\; \big\{\boldsymbol{N}\big((k-1)\,\phi\big) = \boldsymbol{N}(k\,\phi)\big\}\,.$$

Letting $k = 1$ in (B.66) yields that

$$\mathrm{P}(A_1^*) \;\geq\; \exp(-c\,\phi). \tag{B.67}$$

Also, for integers $k \geq 2$, by (B.66) and the fact that $\{\bigcap_{m=1}^{k-1} \overline{A_m^*}\} \in \mathscr{F}_{(k-1)\,\phi}$, we have $\mathrm{P}(A_k^* \mid \cap_{m=1}^{k-1} \overline{A_m^*}) \geq \exp(-c\,\phi)$. Combining this with (B.67), for any integer $\ell \geq 2$, we have

$$\begin{aligned} \mathrm{P}\Big(\bigcap_{k=1}^{\ell} \overline{A_k^*}\Big) \;&=\; \mathrm{P}(\overline{A_1^*}) \cdot \prod_{k=2}^{\ell} \mathrm{P}\Big(\overline{A_k^*} \,\Big|\, \bigcap_{m=1}^{k-1} \overline{A_m^*}\Big) \\ &\leq\; \big\{1 - \exp(-c\,\phi)\big\}^{\ell}, \end{aligned} \tag{B.68}$$

which gives that

$$\mathrm{P}\Big(\bigcup_{k=1}^{\ell} A_k^*\Big) \;=\; 1 - \mathrm{P}\Big(\bigcap_{k=1}^{\ell} \overline{A_k^*}\Big) \;\geq\; 1 - \{1 - \exp(-c\,\phi)\}^{\ell}.$$

Letting $\ell \to \infty$ in the above inequality yields that $\mathrm{P}(\bigcup_{k=1}^{\infty} A_k^*) = 1$. It follows that

$$\begin{aligned} \mathrm{P}\Big(\bigcup_{t \geq \phi} \big\{\boldsymbol{N}((t-\phi, t]) = \boldsymbol{0}\big\}\Big) \;&\geq\; \mathrm{P}\Big(\bigcup_{k \geq 1} \big\{\boldsymbol{N}(((k-1)\,\phi, k\,\phi]) = \boldsymbol{0}\big\}\Big) \\ &=\; \mathrm{P}\Big(\bigcup_{k=1}^{\infty} A_k^*\Big) = 1, \end{aligned} \tag{B.69}$$

which proves (B.64).

We then prove that condition (ii) holds for $\mathcal{U}_1$ using a proof by contradiction. Let $R_1^* = \inf(\mathcal{U}_1)$. If $R_1^* \notin \mathcal{U}_1$, then there exists a sequence of time points $\{u_k\}_{k \geq 1} \in \mathcal{U}_1$ such that $u_1 > u_2 > \cdots$ and $\lim_{k \to \infty} u_k = R_1^*$. Note that $u_k \in \mathcal{U}_1$ implies that $\boldsymbol{N}(u_k) - \boldsymbol{N}(u_k - \phi) = \boldsymbol{N}((u_k - \phi, u_k]) = \boldsymbol{0}$. Using this and the fact that $\boldsymbol{N}(t)$ is right-continuous in $t \geq 0$, we have

$$\boldsymbol{N}((R_1^* - \phi, R_1^*]) \;=\; \boldsymbol{N}(R_1^*) - \boldsymbol{N}(R_1^* - \phi) \;=\; \lim_{k \to \infty} \{\boldsymbol{N}(u_k) - \boldsymbol{N}(u_k - \phi)\} \;=\; \boldsymbol{0},$$

which contradicts with $R_1^* \notin \mathcal{U}_1$.

Next, for each integer $\ell \geq 2$, we prove that R_ℓ exists with probability one. Following the same proof of condition (ii) for the case of $\mathcal{U}_1$, we can verify that condition (ii) also

holds for $\mathcal{U}_\ell$ with integers $\ell \geq 2$. Now we prove condition (i) holds with probability one for $\mathcal{U}_\ell$ with $\ell \geq 2$. Similar to (B.64), it suffices to show that

$$\mathrm{P}\Big(\bigcup_{t \geq R_{\ell-1}+\phi} \big\{ \boldsymbol{N}((t-\phi, t]) = \mathbf{0} \big\} \Big) = 1. \tag{B.70}$$

For each integer $k \geq 1$ and real number $r > 0$, define the event

$$A^*_{k,r} \;=\; \Big\{ \boldsymbol{N}\big(((k-1)\phi + r, k\phi + r]\big) = \mathbf{0} \Big\} \;=\; \Big\{ \boldsymbol{N}((k-1)\phi + r) = \boldsymbol{N}(k\phi + r) \Big\}.$$

Using the same proof as that of (B.66)–(B.69), we obtain $\mathrm{P}(\bigcup_{k=1}^{\infty} A^*_{k,r}) = 1$, and

$$\begin{aligned} \mathrm{P}\Big(\bigcup_{t \geq r+\phi} \big\{ \boldsymbol{N}((t-\phi, t]) = \mathbf{0} \big\} \Big) \;&\geq\; \mathrm{P}\Big(\bigcup_{k \geq 1} \big\{ \boldsymbol{N}\big(((k-1)\phi + r, k\phi + r]\big) = \mathbf{0} \big\} \Big) \\ &=\; \mathrm{P}\Big(\bigcup_{k=1}^{\infty} A^*_{k,r} \Big) = 1. \end{aligned}$$

It follows that for each realization $R_{\ell-1} = r$, we have

$$\mathrm{P}\Big(\bigcup_{t \geq R_{\ell-1}+\phi} \big\{ \boldsymbol{N}((t-\phi, t]) = \mathbf{0} \big\} \,\Big|\, R_{\ell-1} = r \Big) = 1,$$

which proves (B.70). ■

B.10 Proof of Theorem 3

B.10.1 Proof of part (i)

Before proving part (i), we first show Lemma B.11.

Lemma B.11 *Assume conditions* A1, A2, A3, A4 *and* A5 *in Appendix* B. *Let* $R \in [\phi, \infty)$ *be a stopping time with respect to the filtered probability space* $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t \geq 0}, \mathrm{P})$, *satisfying* $\boldsymbol{N}((R-\phi, R]) = \mathbf{0}$. *Then for each* $t \geq 0$, $\boldsymbol{N}(t+R) - \boldsymbol{N}(R)$ *is independent of* $\mathcal{F}_R$, *where* $\mathcal{F}_R$ *is the stopping-time* σ*-algebra with respect to* R.

Proof: Recall the sequence of discontinuity points $\{\breve{T}_\ell\}_{\ell \geq 0}$ defined in (4.5). Let $\ell_R = \max\{\ell \geq 0 : \breve{T}_\ell \leq R\}$. Define the *time-shifted-by-*R marked point process $(\{\vec{T}_\ell\}_{\ell \geq 0}, \{\vec{I}_\ell\}_{\ell \geq 0})$ as follows:

$$\begin{aligned} &\vec{T}_0 = R, \qquad \vec{I}_0 = 0, \\ \text{and} \quad &\vec{T}_\ell = \breve{T}_{\ell_R + \ell}, \quad \vec{I}_\ell = I_{\ell_R + \ell}, \quad \text{for } \ell \geq 1. \end{aligned}$$

Clearly, the intensity functions $\{\lambda_i(t \mid \mathscr{F}_t)\}_{i\in\mathcal{V}}$ are piecewise-constant in the time interval $t \in [R, \infty)$, with their discontinuity points belonging to the set $\{\vec{T}_\ell\}_{\ell\geq 1}$. For $\ell \geq 0$, define $\mathcal{F}_{\vec{T}_\ell} = \{A \in \mathscr{F} : A \cap \{\vec{T}_\ell \leq t\} \in \mathscr{F}_t \text{ for every } t > 0\}$ as the stopping-time σ-algebra with respect to $\vec{T}_\ell$. Then we have the following fact.

(Theorem 1′) The statements in Theorem 1 still hold if $(\breve{T}_\ell, I_\ell, \mathcal{F}_{\breve{T}_\ell})$ is replaced by $(\vec{T}_\ell, \vec{I}_\ell, \mathcal{F}_{\vec{T}_\ell})$ for each integer $\ell \geq 0$.

For simplicity, we refer to this new version of Theorem 1 (modified for $(\vec{T}_\ell, \vec{I}_\ell, \mathcal{F}_{\vec{T}_\ell})$) as Theorem 1′. The proof of Theorem 1′ is similar to that of Theorem 1. We just need to replace (B.15) by $\big\{\{(\vec{T}_0, \vec{I}_0), \ldots, (\vec{T}_{\ell_R}, \vec{I}_{\ell_R})\} = \{(\vec{t}_0, \vec{i}_0), \ldots, (\vec{t}_\ell, \vec{i}_\ell)\}\big\} = \bullet \in \mathcal{F}_{\vec{T}_\ell}$, and the remaining proof of Theorem 1′ follows the similar arguments as those in (B.16)–(B.29).

Next, we prove that

$$(\vec{T}_{\ell+1} - \vec{T}_\ell, \vec{I}_{\ell+1}, \boldsymbol{\lambda}(\vec{T}_{\ell+1} \mid \mathcal{F}_{\vec{T}_{\ell+1}})) \text{ is independent of } \mathcal{F}_R, \text{ for each integer } \ell \geq 0, \quad \text{(B.71)}$$

where $\boldsymbol{\lambda}(t \mid \mathscr{F}_t) = (\lambda_1(t \mid \mathscr{F}_t), \ldots, \lambda_V(t \mid \mathscr{F}_t))^\top$ denotes the vector of intensity functions. We start by proving the case of $\ell = 0$. Using the facts that $\vec{T}_0 = R$ and $\boldsymbol{N}((R-\phi, R]) = \mathbf{0}$, we have $\vec{\mathcal{T}}_0 := \bigcup_{i\in\mathcal{V}} \big\{t \in (\vec{T}_0 - \phi, \vec{T}_0] : N_i(\{t\}) = 1\big\} = \varnothing$, implying that $\vec{T}_0^* = \infty$ (where $\vec{T}_\ell^*$ is the modified version of T_ℓ^* in (4.10)). Following result (ii) in Theorem 1′, we have that $(\vec{T}_1 - \vec{T}_0) \mid \mathcal{F}_{\vec{T}_0} \sim \text{Exp}(\lambda^{\text{sum}}(\vec{T}_0 \mid \mathcal{F}_{\vec{T}_0}))$. Combining (3.1), (3.5), (3.6), and the fact that $\boldsymbol{N}((R-\phi, R]) = \mathbf{0}$ gives that $\lambda^{\text{sum}}(\vec{T}_0 \mid \mathcal{F}_{\vec{T}_0}) = \sum_{i\in\mathcal{V}} \exp(\beta_{0;i})$ is a non-random constant. It follows that $\vec{T}_1 - \vec{T}_0 \sim \text{Exp}(\sum_{i\in\mathcal{V}} \exp(\beta_{0;i}))$ is independent of $\mathcal{F}_{\vec{T}_0}$. Using result (iii) in Theorem 1′, we have $\text{P}(\vec{I}_1 = i \mid \sigma(\mathcal{F}_{\vec{T}_0}, \vec{T}_1)) = \lambda_i(\vec{T}_0 \mid \mathcal{F}_{\vec{T}_0})/\lambda^{\text{sum}}(\vec{T}_0 \mid \mathcal{F}_{\vec{T}_0}) = \exp(\beta_{0;i})/\sum_{j\in\mathcal{V}} \exp(\beta_{0;j})$, for each $i \in \mathcal{V}$. This implies that $(\vec{T}_1 - \vec{T}_0, \vec{I}_1)$ is independent of $\mathcal{F}_{\vec{T}_0}$. Using $\boldsymbol{N}((R-\phi, R]) = \mathbf{0}$, it is easy to show that $\boldsymbol{\lambda}(\vec{T}_1 \mid \mathcal{F}_{\vec{T}_1})$ deterministically depends on $(\vec{T}_1 - \vec{T}_0, \vec{I}_1)$. Therefore, we prove that $(\vec{T}_1 - \vec{T}_0, \vec{I}_1, \boldsymbol{\lambda}(\vec{T}_1 \mid \mathcal{F}_{\vec{T}_1})$ is independent of $\mathcal{F}_{\vec{T}_0} = \mathcal{F}_R$.

Suppose that (B.71) holds for $0, 1, \ldots, \ell-1$, with $\ell \geq 1$. By induction, it suffices to prove the case of ℓ for (B.71). Using the fact that $\boldsymbol{N}((R-\phi, R]) = \mathbf{0}$, we have that $\vec{T}_\ell^* - R$ deterministically depends on $\{(\vec{T}_k - \vec{T}_{k-1}, \vec{I}_k)\}_{k=1,\ldots,\ell}$, which is independent of $\mathcal{F}_R$. Also, $\boldsymbol{\lambda}(\vec{T}_\ell \mid \mathcal{F}_{\vec{T}_\ell})$ is independent of $\mathcal{F}_R$. By results (ii) and (iii) in Theorem 1′, the distribution of $(\vec{T}_{\ell+1} - \vec{T}_\ell, \vec{I}_{\ell+1})$ conditional on $\mathcal{F}_{\vec{T}_\ell}$ deterministically depends on $(\boldsymbol{\lambda}(\vec{T}_\ell \mid \mathcal{F}_{\vec{T}_\ell}), \vec{T}_\ell^* - R)$, and thus is independent of $\mathcal{F}_R$. Finally, since $\boldsymbol{\lambda}(\vec{T}_{\ell+1} \mid \mathcal{F}_{\vec{T}_{\ell+1}})$ deterministically depends on $\{(\vec{T}_k - \vec{T}_{k-1}, \vec{I}_k)\}_{k=1,\ldots,\ell+1}$, we prove that $(\vec{T}_{\ell+1} - \vec{T}_\ell, \vec{I}_{\ell+1}, \boldsymbol{\lambda}(\vec{T}_{\ell+1} \mid \mathcal{F}_{\vec{T}_{\ell+1}}))$ is independent of $\mathcal{F}_R$. This proved (B.71).

Using the similar arguments as in (4.7) and (4.8), we know that the counting process $\{\boldsymbol{N}(t+R)-\boldsymbol{N}(R)\}_{t\geq 0}$ and marked point process $\{(\vec{T}_{\ell+1}-\vec{T}_\ell, \vec{I}_{\ell+1})\}_{\ell\geq 0}$ are equivalent with each other. Therefore, from (B.71), we have that $\boldsymbol{N}(t+R)-\boldsymbol{N}(R)$ is independent of $\mathcal{F}_R$ for each $t\geq 0$. This completes the proof. ■

Now we prove part (i) of Theorem 3. Recall the random processes $\lambda_i(t)$ in (2.6) and $r_{i,\phi}(t)$ in (3.6). Let $\boldsymbol{\lambda}(t \mid \mathscr{F}_t) = (\lambda_1(t \mid \mathscr{F}_t), \ldots, \lambda_V(t \mid \mathscr{F}_t))^\top$ and $\boldsymbol{r}_\phi(t) = (r_{1,\phi}(t), \ldots, r_{V,\phi}(t))^\top$ be the vectors of these random processes. For $t\geq 0$, define the following *time-shifted-by-R_ℓ* random processes:

$$\begin{aligned}
\vec{\boldsymbol{N}}(t) &= \boldsymbol{N}(t+R_\ell) = (\vec{N}_1(t), \ldots, \vec{N}_V(t))^\top, \\
\vec{\boldsymbol{\lambda}}(t \mid \vec{\mathscr{F}}_t) &= \boldsymbol{\lambda}(t+R_\ell \mid \mathcal{F}_{t+R_\ell}) = (\vec{\lambda}_1(t \mid \vec{\mathscr{F}}_t), \ldots, \vec{\lambda}_V(t \mid \vec{\mathscr{F}}_t))^\top, \\
\vec{\boldsymbol{r}}_\phi(t) &= \boldsymbol{r}_\phi(t+R_\ell) = (\vec{r}_{1,\phi}(t), \ldots, \vec{r}_{V,\phi}(t))^\top.
\end{aligned} \tag{B.72}$$

where $\vec{\mathscr{F}}_t = \mathcal{F}_{t+R_\ell} = \{A\in\mathscr{F} : A\cap\{t+R_\ell\leq u\}\in\mathscr{F}_u \text{ for every } u>t\}$. Also, for $s<t$, let $\vec{N}_i((s,t]) = \vec{N}_i(t\vee 0) - \vec{N}_i(s\vee 0)$. We have the three facts below:

Fact (a) : $\vec{\lambda}_i(t \mid \vec{\mathscr{F}}_t)$ is the conditional intensity function of $\vec{N}_i(t)$. This is because

$$\begin{aligned}
\vec{\lambda}_i(t \mid \vec{\mathscr{F}}_t) &= \lambda_i(t+R_\ell \mid \mathcal{F}_{t+R_\ell}) \\
&= \lim_{\Delta\downarrow 0} \Delta^{-1}\mathrm{P}(N_i(t+R_\ell+\Delta) = N_i(t+R_\ell)+1 \mid \mathcal{F}_{t+R_\ell}) \\
&= \lim_{\Delta\downarrow 0} \Delta^{-1}\mathrm{P}(\vec{N}_i(t+\Delta) = \vec{N}_i(t)+1 \mid \vec{\mathscr{F}}_t), \quad t\geq 0,
\end{aligned}$$

agreeing with the definition in (2.6) of an intensity function.

Fact (b) : $\vec{\lambda}_i(t \mid \vec{\mathscr{F}}_t)$, $\vec{r}_{i,\phi}(t)$, and $\vec{N}_i(t)$ follow the same model as in (3.1), (3.5) and (3.6). This could be seen from the following identities:

$$\begin{aligned}
\vec{\lambda}_i(t \mid \vec{\mathscr{F}}_t) &= \exp\Big\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \beta_{j,i}\, g\big(r_{j,\phi}(t+R_\ell)\big)\Big\} \\
&= \exp\Big\{\beta_{0;i} + \sum_{j\in\mathcal{V}} \beta_{j,i}\, g\big(\vec{r}_{j,\phi}(t)\big)\Big\}, \quad t\geq 0,
\end{aligned} \tag{B.73}$$

which is identical to (3.1) and (3.5), and

$$\begin{aligned}
\vec{r}_{i,\phi}(t) &= N_i\big((t+R_\ell-\phi\,,\, t+R_\ell]\big)/\phi \\
&= N_i\big(((t+R_\ell-\phi)\vee R_\ell\,,\, t+R_\ell]\big)/\phi \\
&= \vec{N}_i\big((t-\phi\,,\, t]\big)/\phi, \qquad t\geq 0,
\end{aligned} \tag{B.74}$$

which is identical to (3.6). Here, (B.74) follows from the fact that $N_i((R_\ell-\phi, R_\ell]) = 0$, which is implied by (4.25).

Fact (c) : The following mappings are deterministic, with M1 = M1′, M2 = M2′, M3 = M3′, M4 = M4′, and M5 = M5′.

$$\begin{array}{lll}
(\mathrm{M1}) & \text{from } \{\boldsymbol{N}(t+R_\ell)-\boldsymbol{N}(R_\ell)\}_{t\geq 0} & \text{to } \{\boldsymbol{N}(t+R_\ell)-\boldsymbol{N}((t+R_\ell-\phi)\vee R_\ell)\}_{t\geq 0}.\\
(\mathrm{M2}) & \text{from } \{\boldsymbol{N}(t+R_\ell)-\boldsymbol{N}(R_\ell)\}_{t\geq 0} & \text{to } \{\vec{\boldsymbol{r}}_\phi(t)\}_{t\geq 0}.\\
(\mathrm{M3}) & \text{from } \{\boldsymbol{N}(t+R_\ell)-\boldsymbol{N}(R_\ell)\}_{t\geq 0} & \text{to } \{\vec{\boldsymbol{\lambda}}(t\mid \vec{\mathscr{F}}_t)\}_{t\geq 0}.\\
(\mathrm{M4}) & \text{from } \{\boldsymbol{N}(t+R_\ell)-\boldsymbol{N}(R_\ell)\}_{t\geq 0} & \text{to } R_{\ell+1}-R_\ell.\\
(\mathrm{M5}) & \text{from } \{\boldsymbol{N}(t+R_\ell)-\boldsymbol{N}(R_\ell)\}_{t\geq 0} & \text{to } \boldsymbol{N}((R_\ell,R_{\ell+1}]).
\end{array}$$

and

$$\begin{array}{lll}
(\mathrm{M1}') & \text{from } \{\boldsymbol{N}(t)\}_{t\geq 0} & \text{to } \{\boldsymbol{N}(t)-\boldsymbol{N}(t-\phi)\}_{t\geq 0}.\\
(\mathrm{M2}') & \text{from } \{\boldsymbol{N}(t)\}_{t\geq 0} & \text{to } \{\boldsymbol{r}_\phi(t)\}_{t\geq 0}.\\
(\mathrm{M3}') & \text{from } \{\boldsymbol{N}(t)\}_{t\geq 0} & \text{to } \{\boldsymbol{\lambda}(t\mid \mathscr{F}_t)\}_{t\geq 0}.\\
(\mathrm{M4}') & \text{from } \{\boldsymbol{N}(t)\}_{t\geq 0} & \text{to } R_1.\\
(\mathrm{M5}') & \text{from } \{\boldsymbol{N}(t)\}_{t\geq 0} & \text{to } \boldsymbol{N}(R_1).
\end{array}$$

To show this, note that for any $t\geq 0$, the following two identities hold

$$\begin{aligned}
& \boldsymbol{N}(t+R_\ell)-\boldsymbol{N}((t+R_\ell-\phi)\vee R_\ell)\\
= \;& \{\boldsymbol{N}(t+R_\ell)-\boldsymbol{N}(R_\ell)\}-\{\boldsymbol{N}((t+R_\ell-\phi)\vee R_\ell)-\boldsymbol{N}(R_\ell)\}\\
= \;& \{\boldsymbol{N}(u+R_\ell)-\boldsymbol{N}(R_\ell)\}|_{u=t}-\{\boldsymbol{N}(u+R_\ell)-\boldsymbol{N}(R_\ell)\}|_{u=(t-\phi)\vee 0},
\end{aligned}$$

$$\begin{aligned}
& \boldsymbol{N}(t)-\boldsymbol{N}(t-\phi)\\
= \;& \boldsymbol{N}(u)|_{u=t}-\boldsymbol{N}(u)|_{u=(t-\phi)\vee 0},
\end{aligned}$$

which implies that the mapping M1 is deterministic, and M1 = M1′. This combined with (B.74) and (3.6) gives that the mappings M2 = M2′ are deterministic. By using (B.73), (3.1), (3.5) and the fact that the mappings M2 = M2′ are deterministic, we have that the mappings M3 = M3′ are deterministic. From (4.25), we observe that

$$\begin{aligned}
R_1 &= \min\{t\geq\phi : \boldsymbol{N}(t)-\boldsymbol{N}(t-\phi)=\boldsymbol{0}\},\\
R_{\ell+1}-R_\ell &= \min\{t\geq\phi : \boldsymbol{N}(t+R_\ell)-\boldsymbol{N}((t+R_\ell-\phi)\vee R_\ell)=\boldsymbol{0}\}.
\end{aligned}$$

This together with the fact that M1 = M1′ are deterministic yields that the mappings M4 = M4′ are deterministic. From

$$\boldsymbol{N}((R_\ell,R_{\ell+1}])=\boldsymbol{N}((R_{\ell+1}-R_\ell)+R_\ell)-\boldsymbol{N}(R_\ell),$$

we have that the mapping from $(\{\boldsymbol{N}(t+R_\ell)-\boldsymbol{N}(R_\ell)\}_{t\geq 0},\, R_{\ell+1}-R_\ell)$ to $\boldsymbol{N}((R_\ell,R_{\ell+1}])$ is deterministic, and identical to that from $(\{\boldsymbol{N}(t)\}_{t\geq 0},\, R_1)$ to $\boldsymbol{N}(R_1)$. This together with M4 = M4′ yields that M5 = M5′ are deterministic.

Fact (a) verifies that $\vec{\boldsymbol{\lambda}}(t\mid\vec{\mathscr{F}}_t)$ is the vector of intensity processes of $\vec{\boldsymbol{N}}(t)$. From Fact (c), we have that the deterministic mappings M3 = M3′. Combining these yields that $(\boldsymbol{N}(t+R_\ell)-\boldsymbol{N}(R_\ell),\vec{\boldsymbol{\lambda}}(t\mid\vec{\mathscr{F}}_t)) \overset{\mathrm{D}}{=} (\boldsymbol{N}(t),\boldsymbol{\lambda}(t\mid\mathscr{F}_t))$ for any $t\geq 0$, which proves

$\boldsymbol{N}(t+R_\ell)-\boldsymbol{N}(R_\ell) \stackrel{\mathrm{D}}{=} \boldsymbol{N}(t)$. On the other hand, by (B.73) and (B.74), it is seen that for any $t \geq 0$, $\vec{\boldsymbol{\lambda}}(t \mid \vec{\mathscr{F}}_t)$ is a deterministic function of $\{\boldsymbol{N}(s+R_\ell)-\boldsymbol{N}(R_\ell)\}_{0\leq s\leq t}$. By Lemma B.11, we have that $\boldsymbol{N}(t+R_\ell)-\boldsymbol{N}(R_\ell)$ is independent of $\mathcal{F}_{R_\ell}$ for any $t \geq 0$. This combined with the fact that the mapping M3 is deterministic proves that $\vec{\boldsymbol{\lambda}}(t \mid \vec{\mathscr{F}}_t)$ is also independent of $\mathcal{F}_{R_\ell}$. The proof of part (i) completes. ■

B.10.2 Proof of part (ii)

Using the facts that the mappings M5 = M5′ are deterministic and that $\boldsymbol{N}(t+R_\ell)-\boldsymbol{N}(R_\ell) \stackrel{\mathrm{D}}{=} \boldsymbol{N}(t)$ for any $t \geq 0$ from Theorem 3 (i), we obtain $\boldsymbol{N}((R_\ell, R_{\ell+1}]) \stackrel{\mathrm{D}}{=} \boldsymbol{N}(R_1)$. Thus, $\{\boldsymbol{N}((R_\ell, R_{\ell+1}])\}_{\ell\geq 0}$ is a sequence of identically distributed random vectors. On the other hand, since $\boldsymbol{N}(t+R_\ell)-\boldsymbol{N}(R_\ell)$ is independent of $\mathcal{F}_{R_\ell}$ and the mapping M5 is deterministic, we have that $\boldsymbol{N}((R_\ell, R_{\ell+1}])$ is independent of $\mathcal{F}_{R_\ell}$. Also, $\boldsymbol{N}((R_k, R_{k+1}])$ is $\mathcal{F}_{R_\ell}$-measurable for any $0 \leq k \leq \ell-1$. It follows that $\boldsymbol{N}((R_\ell, R_{\ell+1}])$ is independent of $\{\boldsymbol{N}((R_k, R_{k+1}])\}_{0\leq k\leq \ell-1}$. Hence, $\{\boldsymbol{N}((R_\ell, R_{\ell+1}])\}_{\ell\geq 0}$ is a sequence of independent random vectors. The proof is completed. ■

B.10.3 Proof of part (iii)

Before proving part (iii), we first show Lemmas B.12 and B.13.

Lemma B.12 *Assume conditions* A1*,* A2*,* A3*,* A4 *and* A5 *in Appendix* B*. Let* R_1 *be the first recurrence time point defined in* (4.25) *with* $\ell = 1$*. Then* $\mathrm{E}(R_1^2) < \infty$*.*

Proof: Let the random variable $L^* = \min\{k \geq 1 : \mathrm{I}(A_k^*) = 1\}$, where the event $A_k^* = \{\boldsymbol{N}(((k-1)\,\phi, k\,\phi]) = \boldsymbol{0}\}$ is defined in (B.66). By (B.68), for any integer $\ell > 2$, we have

$$\mathrm{P}(L^* \geq \ell) \;=\; \mathrm{P}\Big(\bigcap_{k=1}^{\ell-1} \overline{A_k^*}\Big) \;\leq\; \big\{1-\exp(-c\,\phi)\big\}^{\ell-1}.$$

Thus, the tail probability of L^* is bounded by that of the geometric distribution with the constant success probability $\exp(-c\,\phi)$. Since geometric distribution has finite first and second moments, we have $\mathrm{E}\{(L^*)^2\} < \infty$. Also, $\mathrm{I}(A_{L^*}^*) = 1$ implies that $\boldsymbol{N}(((L^*-1)\,\phi, L^*\,\phi]) = \boldsymbol{0}$, which combined with (4.25) gives that $R_1 \leq L^*\,\phi$. Hence, we obtain

$$\mathrm{E}(R_1^2) \;\leq\; \mathrm{E}\{(L^*\,\phi)^2\} \;=\; \phi^2\,\mathrm{E}\{(L^*)^2\} \;<\; \infty.$$

This completes the proof. ■

Lemma B.13 *Assume conditions* A1, A2, A3, A4 *and* A5 *in Appendix* B. *Let* $h(\cdot) : \mathbb{R}^{2V} \to \mathbb{R}$ *be a continuous function. For each integer* $\ell \geq 1$, *define the random variable*

$$\mathrm{S}_\ell \;=\; \int_{R_{\ell-1}}^{R_\ell} h(\boldsymbol{\lambda}(t \mid \mathscr{F}_t), \boldsymbol{r}_\phi(t))\,\mathrm{d}t. \tag{B.75}$$

Then $\{\mathrm{S}_\ell\}_{\ell\geq 1}$ *is a sequence of* i.i.d. *random variables.*

Proof: Following the notations in (B.72), for $\ell \geq 1$, we have

$$\mathrm{S}_{\ell+1} = \int_{R_\ell}^{R_{\ell+1}} h(\boldsymbol{\lambda}(t \mid \mathscr{F}_t), \boldsymbol{r}_\phi(t))\,\mathrm{d}t = \int_0^{R_{\ell+1}-R_\ell} h(\vec{\boldsymbol{\lambda}}(t \mid \vec{\mathscr{F}}_t), \vec{\boldsymbol{r}}_\phi(t))\,\mathrm{d}t. \tag{B.76}$$

By comparing (B.76) with $\mathrm{S}_1 = \int_0^{R_1} h(\boldsymbol{\lambda}(t \mid \mathscr{F}_t), \boldsymbol{r}_\phi(t))\,\mathrm{d}t$, and using the fact that the mappings M2 = M2′, M3 = M3′ and M4 = M4′ are deterministic, we have that the mappings

$$\begin{aligned} &\text{(M6)} \quad \text{from } \{\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell)\}_{t\geq 0} \text{ to } \mathrm{S}_{\ell+1}, \\ &\text{(M6}') \quad \text{from } \{\boldsymbol{N}(t)\}_{t\geq 0} \text{ to } \mathrm{S}_1, \end{aligned}$$

are both deterministic, with M6 = M6′. Combining this with the fact that $\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell) \overset{\mathrm{D}}{=} \boldsymbol{N}(t)$ for any $t \geq 0$ from Theorem 3 (i), we obtain $\mathrm{S}_{\ell+1} \overset{\mathrm{D}}{=} \mathrm{S}_1$. Thus, $\{\mathrm{S}_\ell\}_{\ell\geq 1}$ is a sequence of identically distributed random variables. On the other hand, using the facts that $\boldsymbol{N}(t+R_\ell) - \boldsymbol{N}(R_\ell)$ is independent of $\mathcal{F}_{R_\ell}$ and the mapping M6 is deterministic, we have that $\mathrm{S}_{\ell+1}$ is independent of $\mathcal{F}_{R_\ell}$. Also, S_k is $\mathcal{F}_{R_\ell}$-measurable for $1 \leq k \leq \ell$. Hence, $\{\mathrm{S}_\ell\}_{\ell\geq 1}$ is a sequence of independent variables. The proof is completed. ■

Now we prove part (iii) of Theorem 3. Note that $D_\ell = R_\ell - R_{\ell-1}$ is a special case of S_ℓ in (B.75) with $h(\cdot) \equiv 1$. Applying Lemma B.13, we conclude that $\{D_\ell\}_{\ell\geq 1}$ is a sequence of i.i.d. random variables. Furthermore, Lemma B.12 proved that $D_1 = R_1$ has the finite second moment. This completes the proof. ■

B.11 Proof of Theorem 4

Before proving Theorem 4, we first show Lemma B.14.

Lemma B.14 (Asymptotic convergence of $\Lambda_i(t) = \int_0^t \lambda_i(u \mid \mathscr{F}_u)\,\mathrm{d}u$) *Assume conditions* A1, A2, A3, A4 *and* A5 *in Appendix* B. *For each* $i \in \mathcal{V}$, *consider the random process* $\Lambda_i(t) = \int_0^t \lambda_i(u \mid \mathscr{F}_u)\,\mathrm{d}u$ *for* $t > 0$. *Then there exists a constant* $c_i \in (0, \infty)$, *such that*

$$\Lambda_i(t)/t \xrightarrow{\mathrm{P}} c_i, \quad \text{as } t \to \infty. \tag{B.77}$$

Proof: Denote the increment of $\Lambda_i(t)$ in the ℓth *recurrence cycle* $(R_{\ell-1}, R_\ell]$ by

$$\mathrm{S}_{i,\ell} \,=\, \Lambda_i(R_\ell) - \Lambda_i(R_{\ell-1}) \,=\, \int_{R_{\ell-1}}^{R_\ell} \lambda_i(t \mid \mathscr{F}_t)\,\mathrm{d}t.$$

Applying $h(\boldsymbol{\lambda}(t \mid \mathscr{F}_t), \boldsymbol{r}_\phi(t)) = \lambda_i(t \mid \mathscr{F}_t)$ to (B.75) in Lemma B.13 indicates that $\{\mathrm{S}_{i,\ell}\}_{\ell\geq 1}$ is a sequence of i.i.d. random variables.

By condition A5, there exists a constant $c \in (0, \infty)$, such that

$$\mathrm{S}_{i,\ell} \,=\, \int_{R_{\ell-1}}^{R_\ell} \lambda_i(t \mid \mathscr{F}_t)\,\mathrm{d}t \,\leq\, \int_{R_{\ell-1}}^{R_\ell} c\,\mathrm{d}t \,=\, c\,D_\ell, \tag{B.78}$$

where $D_\ell \,=\, R_\ell - R_{\ell-1}$. Combining Lemma B.12 and (B.78) implies that the second moments of D_ℓ and $\mathrm{S}_{i,\ell}$ are finite. Applying the strong law of large numbers, we obtain

$$\frac{1}{\ell}\sum_{k=1}^{\ell} \mathrm{S}_{i,k} \overset{\text{a.s.}}{\to} \mathrm{E}(\mathrm{S}_{i,1}), \quad \frac{1}{\ell}\sum_{k=1}^{\ell} D_k \overset{\text{a.s.}}{\to} \mathrm{E}(D_1), \quad \text{as } \ell \to \infty.$$

Thus, for arbitrarily small $\delta > 0$ and $\epsilon > 0$, we could find sufficiently large C_1, such that

$$\mathrm{P}\Big(\sup_{\ell > C_1} \max\Big\{ \Big| \frac{1}{\ell}\sum_{k=1}^{\ell} \mathrm{S}_{i,k} - \mathrm{E}(\mathrm{S}_{i,1}) \Big|, \Big| \frac{1}{\ell}\sum_{k=1}^{\ell} D_k - \mathrm{E}(D_1) \Big| \Big\} > \epsilon \Big) < \delta. \tag{B.79}$$

On the other hand, for any time point $t > 0$, let $L_t = \sup\{\ell \geq 0 : R_\ell \leq t\}$ be the number of recurrence time points up to t. We have

$$\sum_{k=1}^{L_t} \mathrm{S}_{i,k} \leq \Lambda_i(t) \leq \sum_{k=1}^{L_t+1} \mathrm{S}_{i,k}, \quad \sum_{k=1}^{L_t} D_k \leq t \leq \sum_{k=1}^{L_t+1} D_k,$$

which directly yields that

$$\frac{\sum_{k=1}^{L_t} \mathrm{S}_{i,k}}{\sum_{k=1}^{L_t+1} D_k} \leq \frac{\Lambda_i(t)}{t} \leq \frac{\sum_{k=1}^{L_t+1} \mathrm{S}_{i,k}}{\sum_{k=1}^{L_t} D_k}. \tag{B.80}$$

Since $\mathrm{E}(D_1^2) < \infty$, it is easy to show that $L_t \overset{\mathrm{P}}{\to} \infty$ as $t \to \infty$. Thus for arbitrarily small $\delta_2 > 0$ and large $C_2 > C_1$, there exists $t_0 > 0$, such that for $\forall t > t_0$, $\mathrm{P}(L_t > C_2) > 1 - \delta_2$. Combining (B.79) and (B.80), the following (B.81) holds with probability at least $1 - \delta - \delta_2$ for $t > t_0$:

$$\frac{C_2\,\{\mathrm{E}(\mathrm{S}_{i,1}) - \epsilon\}}{(C_2+1)\,\{\mathrm{E}(D_1) + \epsilon\}} \leq \frac{\Lambda_i(t)}{t} \leq \frac{(C_2+1)\,\{\mathrm{E}(\mathrm{S}_{i,1}) + \epsilon\}}{C_2\,\{\mathrm{E}(D_1) - \epsilon\}}. \tag{B.81}$$

Since ϵ, δ, and δ_2 are arbitrarily small and C_2 is arbitrarily large, (B.81) implies that

$$\Lambda_i(t)/t \overset{\mathrm{P}}{\to} \mathrm{E}(\mathrm{S}_{i,1})/\mathrm{E}(D_1), \quad \text{as } t \to \infty.$$

We complete the proof with $c_i = \mathrm{E}(\mathrm{S}_{i,1})/\mathrm{E}(D_1)$ in (B.77). ■

Now we prove Theorem 4. By Theorem 2, for each $i \in \mathcal{V}$ and $t \in (0,\infty)$, we have $\mathrm{var}\{(N_i(t) - \Lambda_i(t))/t\} \le c_1/t$, which together with Lemma 7 implies that

$$\{N_i(t) - \Lambda_i(t)\}/t \xrightarrow{\mathrm{P}} 0, \quad \text{as } t \to \infty. \tag{B.82}$$

By Lemma B.14 and (B.82), we obtain

$$\boldsymbol{N}(t)/t \xrightarrow{\mathrm{P}} \boldsymbol{c}_0, \quad \text{as } t \to \infty,$$

where $\boldsymbol{c}_0 = (\mathrm{E}(\mathrm{S}_{1,1}), \ldots, \mathrm{E}(\mathrm{S}_{V,1}))^\top/\mathrm{E}(D_1)$. This completes the proof. ■

B.12 Proof of Theorem 5

Before proving Theorem 5, we first show Lemmas B.15, B.16 and B.17.

Lemma B.15 *Assume conditions* A1, A2, A3, A4 *and* A5 *in Appendix* B. *Let* $f(\cdot) : \mathbb{R}^V \to [0,\infty)$ *be a non-negative continuous function bounded above by* $c_0 \in (0,\infty)$. *For* $t \ge 0$, *let* $Y(t) = f(\boldsymbol{r}_\phi(t))$, *where* $\boldsymbol{r}_\phi(t)$ *is defined above* (B.72). *Let* R_1 *be the first recurrence time point defined in* (4.25) *with* $\ell = 1$. *Assume that* $\mathrm{E}\{\int_0^{R_1} Y(t)\,\mathrm{d}t\} > 0$. *Then there exists a constant* $c_i \in (0,\infty)$, *such that*

$$\int_0^t \lambda_i(u \mid \mathscr{F}_u)\, Y(u)\,\mathrm{d}u/t \xrightarrow{\mathrm{P}} c_i, \quad \textit{as } t \to \infty.$$

Proof: For each integer $\ell \ge 1$, define

$$\mathrm{S}^*_{i,\ell} = \int_{R_{\ell-1}}^{R_\ell} \lambda_i(t \mid \mathscr{F}_t)\, Y(t)\,\mathrm{d}t.$$

Applying Lemma B.13 with $h(\boldsymbol{\lambda}(t \mid \mathscr{F}_t), \boldsymbol{r}_\phi(t \mid \mathscr{F}_t)) = \lambda_i(t \mid \mathscr{F}_t) f(\boldsymbol{r}_\phi(t)) = \lambda_i(t \mid \mathscr{F}_t)\, Y(t)$, we have that $\{\mathrm{S}^*_{i,\ell}\}_{\ell \ge 1}$ is a sequence of i.i.d. random variables. By condition A5, there exist constants $c_2, c_3 \in (0,\infty)$ such that $c_2 \le \lambda_i(t \mid \mathscr{F}_t) \le c_3$ for any $t \in [0,\infty)$. We obtain the following moment inequalities:

$$\begin{aligned}
\mathrm{E}(\mathrm{S}^*_{i,1}) &= \mathrm{E}\Big\{ \int_0^{R_1} \lambda_i(t \mid \mathscr{F}_t)\, Y(t)\,\mathrm{d}t \Big\} \ge c_2\, \mathrm{E}\Big\{ \int_0^{R_1} Y(t)\,\mathrm{d}t \Big\} > 0, \\
\mathrm{E}\big\{(\mathrm{S}^*_{i,1})^2\big\} &= \mathrm{E}\Big[\Big\{ \int_0^{R_1} \lambda_i(t \mid \mathscr{F}_t)\, Y(t)\,\mathrm{d}t \Big\}^2 \Big] \le c_3^2\, c_0^2\, \mathrm{E}(R_1^2) < \infty.
\end{aligned}$$

Applying the same proof as Lemma B.14 with $\mathrm{S}_{i,\ell} = \mathrm{S}^*_{i,\ell}$, one can show that

$$\frac{\int_0^t \lambda_i(u \mid \mathscr{F}_u)\, Y(u)\,\mathrm{d}u}{t} \xrightarrow{\mathrm{P}} \frac{\mathrm{E}\{\int_0^{R_1} \lambda_i(u \mid \mathscr{F}_u)\, Y(u)\,\mathrm{d}u\}}{\mathrm{E}(R_1)} = \frac{\mathrm{E}(\mathrm{S}^*_{i,1})}{\mathrm{E}(R_1)} > 0, \quad \text{as } t \to \infty.$$

This completes the proof. ■

Lemma B.16 *Assume conditions* A1, A2, A3, A4 *and* A5 *in Appendix* B. *For* $i \in \mathcal{V}$, *let* $x_i(t) = g(r_{i,\phi}(t))$ *be the covariate defined in* (3.5), *and* $x_0(t) \equiv 1$. *Then for any* $i, j \in \mathcal{V} \cup \{0\}$ (*not necessarily distinct*), *we have*

$$\mathrm{E}\Big\{ \int_0^{R_1} x_i(u)\,\mathrm{d}u \Big\} \;>\; 0, \tag{B.83}$$

$$\mathrm{E}\Big\{ \int_0^{R_1} x_i(u)\, x_j(u)\,\mathrm{d}u \Big\} \;>\; 0. \tag{B.84}$$

Proof: If $i = 0$, then (B.83) obviously holds; if either i or j is zero, then (B.84) reduces to (B.83). Thus, to prove Lemma B.16, it suffices to verify (B.83) and (B.84) for the case of $i, j \in \mathcal{V}$.

By the similar proof to that below (B.62), for any $t > 0$, we have

$$\mathrm{P}(N_i(t) \geq 1) \;=\; 1 - \mathrm{P}(N_i(t) = 0) \;=\; 1 - \exp\{-\lambda_i(0) \cdot t\} \;>\; 0. \tag{B.85}$$

The OM condition (2.10) implies that

$$\mathrm{P}(N_i(t)\, N_j(t) \geq 1) \;=\; \lambda_i(0)\, \lambda_j(0)\, t^2 + o(t^2),$$

as $t \to 0$, and thus there exists $t_0 \in (0, \phi)$, such that

$$\mathrm{P}(N_i(t_0)\, N_j(t_0) \geq 1) \;>\; 0. \tag{B.86}$$

For $t \in (0, \phi)$, we observe that $r_{i,\phi}(t) = N_i((t-\phi, t])/\phi = N_i(t)/\phi$. Hence, $r_{i,\phi}(t)$ is increasing in $t \in (0, \phi)$, which implies that $x_i(t) = g(r_{i,\phi}(t))$ is also increasing in $t \in (0, \phi)$. Together with (B.85), (B.86) and the fact that $R_1 \geq \phi > t_0$, we obtain

$$\begin{aligned}
\mathrm{E}\Big\{ \int_0^{R_1} x_i(u)\,\mathrm{d}u \Big\} \;&\geq\; \mathrm{E}\{x_i(t_0) \cdot (\phi - t_0)\} \\
&>\; 0, \\
\mathrm{E}\Big\{ \int_0^{R_1} x_i^2(u)\,\mathrm{d}u \Big\} \;&\geq\; \mathrm{E}\{x_i^2(t_0) \cdot (\phi - t_0)\} \\
&>\; 0, \\
\mathrm{E}\Big\{ \int_0^{R_1} x_i(u)\, x_j(u)\,\mathrm{d}u \Big\} \;&\geq\; \mathrm{E}\{x_i(t_0)\, x_j(t_0) \cdot (\phi - t_0)\} \\
&>\; 0.
\end{aligned}$$

These complete the proof. ■

Lemma B.17 *Assume conditions* A1, A2, A3, A4 *and* A5 *in Appendix* B. *Let* $\widetilde{\boldsymbol{x}}(t) = (1, x_1(t), x_2(t), \ldots, x_V(t))^\top$ *be the vector of the covariates defined in* (3.5). *Then for any* $\widetilde{\boldsymbol{u}} \in \mathbb{R}^{V+1}$ *with* $\|\widetilde{\boldsymbol{u}}\| > 0$,

$$\mathrm{E}\Big[\int_0^{R_1} \big\{ \widetilde{\boldsymbol{x}}(t)^\top \widetilde{\boldsymbol{u}} \big\}^2 \,\mathrm{d}t \Big] > 0.$$

Proof: Let $\widetilde{\boldsymbol{u}} = (u_0, u_1, \ldots, u_V)^\top$. We proceed by cases of u_0.

Case (i) $u_0 \neq 0$. Consider $t_0 \in (0, \phi)$. By (B.65), we have

$$\mathrm{P}(\boldsymbol{N}(t_0) = \boldsymbol{0}) \;\; \geq \;\; \exp\{-c \cdot (t_0 - 0)\} > 0. \tag{B.87}$$

Note that $\boldsymbol{N}(t_0) = \boldsymbol{0}$ implies that $\widetilde{\boldsymbol{x}}(t)^\top \widetilde{\boldsymbol{u}} = u_0$ for $t \in [0, t_0]$. Combining (B.87) and the fact that $R_1 \geq \phi > t_0$, we obtain

$$\begin{aligned} \mathrm{E}\Big[\int_0^{R_1} \big\{\widetilde{\boldsymbol{x}}(t)^\top \widetilde{\boldsymbol{u}}\big\}^2 \,\mathrm{d}t \Big] \;\; &\geq \;\; \mathrm{E}\Big[\int_0^{t_0} \big\{\widetilde{\boldsymbol{x}}(t)^\top \widetilde{\boldsymbol{u}}\big\}^2 \cdot \mathrm{I}(\boldsymbol{N}(t_0) = \boldsymbol{0}) \,\mathrm{d}t \Big] \\ &= \;\; t_0 \, u_0^2 \, \mathrm{P}(\boldsymbol{N}(t_0) = \boldsymbol{0}) > 0. \end{aligned}$$

Case (ii) $u_0 = 0$. Since $\|\widetilde{\boldsymbol{u}}\| > 0$ and $u_0 = 0$, there exists some $i \in \mathcal{V}$ such that $u_i \neq 0$. We have

$$\begin{aligned} & \mathrm{P}(N_j(t) = 0 \text{ for all } j \in \mathcal{V} \setminus i, \quad N_i(t) = 1) \\ \geq \;\; & \mathrm{P}(N_i(t) = 1) - \sum_{j \in \mathcal{V} \setminus i} \mathrm{P}(N_i(t) = 1, N_j(t) = 1) - \sum_{j \in \mathcal{V} \setminus i} \mathrm{P}(N_j(t) > 1) \\ = \;\; & \lambda_i(0)\, t + o(t) - \sum_{j \in \mathcal{V} \setminus i} \{\lambda_i(0)\, \lambda_j(0)\, t^2 + o(t^2)\} - o(t) \qquad \text{(B.88)} \\ = \;\; & \lambda_i(0)\, t + o(t), \end{aligned}$$

as $t \to 0$, where (B.88) is derived from (2.6), (2.7), and (2.10). Hence, there exists $t_0 \in (0, \phi)$, such that

$$\mathrm{P}(N_j(t_0) = 0 \text{ for all } j \in \mathcal{V} \setminus i, \; N_i(t_0) = 1) > 0. \tag{B.89}$$

Let $t_1 \in (t_0, \phi)$. From (B.65) and (B.89), we have

$$\begin{aligned} & \mathrm{P}(N_j(t_1) = 0 \text{ for all } j \in \mathcal{V} \setminus i, \; N_i(t_0) = N_i(t_1) = 1) \\ = \;\; & \mathrm{P}(\boldsymbol{N}(t_1) = \boldsymbol{N}(t_0) \mid N_j(t_0) = 0 \text{ for all } j \in \mathcal{V} \setminus i, \; N_i(t_0) = 1) \\ & \times \mathrm{P}(N_j(t_0) = 0 \text{ for all } j \in \mathcal{V} \setminus i, \; N_i(t_0) = 1) \\ \geq \;\; & \exp\{-c \cdot (t_1 - t_0)\} \cdot \mathrm{P}(N_j(t_0) = 0 \text{ for all } j \in \mathcal{V} \setminus i, \; N_i(t_0) = 1) \\ > \;\; & 0. \end{aligned} \tag{B.90}$$

Note that the event $\{N_j(t_1) = 0 \text{ for all } j \in \mathcal{V} \setminus i, \text{ and } N_i(t_0) = N_i(t_1) = 1\}$ implies that $\boldsymbol{x}(t)^\top \widetilde{\boldsymbol{u}} = x_i(t)\, u_i = g(1/\phi)\, u_i$ for $t \in (t_0, t_1)$. Together with (B.90) and the fact that $R_1 \geq \phi > t_1$, we obtain

$$\begin{aligned} & \mathrm{E}\Big[\int_0^{R_1} \{\widetilde{\boldsymbol{x}}(t)^\top \widetilde{\boldsymbol{u}}\}^2 \,\mathrm{d}t \Big] \\ \geq \;\; & (t_1 - t_0)\, g^2(1/\phi)\, u_i^2 \cdot \mathrm{P}\big(N_j(t_1) = 0 \text{ for all } j \in \mathcal{V} \setminus i, \; N_i(t_0) = N_i(t_1) = 1\big) \\ > \;\; & 0. \end{aligned}$$

Combining the results of cases (i) and (ii) completes the proof. ■

Now we prove Theorem 5. Recall (5.4),

$$\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) \;=\; \frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} \Big[\exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top\widetilde{\boldsymbol{\beta}}_i\big\}\,\mathrm{d}t - \{\widetilde{\boldsymbol{x}}_i(t-)^\top\widetilde{\boldsymbol{\beta}}_i\}\,\mathrm{d}N_i(t)\Big].$$

With some algebra, we obtain the gradient vector and Hessian matrix of $\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$:

$$\nabla\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) \;=\; \frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} \Big[\widetilde{\boldsymbol{x}}_i(t)\,\exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top\widetilde{\boldsymbol{\beta}}_i\big\}\,\mathrm{d}t - \widetilde{\boldsymbol{x}}_i(t-)\,\mathrm{d}N_i(t)\Big], \tag{B.91}$$

$$\nabla^2\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) \;=\; \frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} \widetilde{\boldsymbol{x}}_i(t)\,\widetilde{\boldsymbol{x}}_i(t)^\top\exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top\widetilde{\boldsymbol{\beta}}_i\big\}\,\mathrm{d}t. \tag{B.92}$$

Let $\widetilde{x}_{i,j}(t)$ denote the jth component of $\widetilde{\boldsymbol{x}}_i(t)$, i.e.,

$$\widetilde{x}_{i,j}(t) = \begin{cases} 1, & \text{if } j = 1, \\ x_{j-1}(t), & \text{if } 1 < j \le i, \\ x_j(t), & \text{if } i < j \le V. \end{cases} \tag{B.93}$$

For each $j \in \mathcal{V}$, applying (4.20) in Lemma 7 gives that

$$\mathrm{E}\Big(\frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} \Big[\widetilde{x}_{i,j}(t)\,\exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top\widetilde{\boldsymbol{\beta}}_i^*\big\}\,\mathrm{d}t - \widetilde{x}_{i,j}(t-)\,\mathrm{d}N_i(t)\Big]\Big) \;=\; 0. \tag{B.94}$$

Also, using (4.24) in Theorem 2, we have

$$\mathrm{var}\Big(\frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} \Big[\widetilde{x}_{i,j}(t)\,\exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top\widetilde{\boldsymbol{\beta}}_i^*\big\}\,\mathrm{d}t - \widetilde{x}_{i,j}(t-)\,\mathrm{d}N_i(t)\Big]\Big) \;\le\; c_1/\mathrm{T}, \tag{B.95}$$

where $c_1 \in (0,\infty)$ is some constant. By Chebyshev's inequality, (B.94) and (B.95), we have

$$\frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} \Big[\widetilde{\boldsymbol{x}}_i(t)\exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top\widetilde{\boldsymbol{\beta}}_i^*\big\}\,\mathrm{d}t - \widetilde{\boldsymbol{x}}_i(t-)\,\mathrm{d}N_i(t)\Big] = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}}). \tag{B.96}$$

This verifies (5.5) in part (i).

Next we prove part (ii). Write $\mathbf{H}_{i,\mathrm{T}} = \nabla^2\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)$. For any $j,k \in \mathcal{V}$, from (B.92) and (B.93), the (j,k)th entry of $\mathbf{H}_{i,\mathrm{T}}$ is

$$\mathbf{H}_{i,\mathrm{T}}(j,k) = \frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} \widetilde{x}_{i,j}(t)\,\widetilde{x}_{i,k}(t)\,\exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top\widetilde{\boldsymbol{\beta}}_i^*\big\}\,\mathrm{d}t.$$

Let $Y_1(t) = \widetilde{x}_{i,j}(t)\,\widetilde{x}_{i,k}(t)$. Lemma B.16 proved that $\mathrm{E}\{\int_0^{R_1} Y_1(t)\,\mathrm{d}t\} > 0$. This enables us to apply Lemma B.15 with $Y(t) = Y_1(t)$, which yields that

$$\mathbf{H}_{i,\mathrm{T}}(j,k) = \frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} \widetilde{x}_{i,j}(t)\,\widetilde{x}_{i,k}(t)\,\lambda_i^*(t \mid \mathscr{F}_t)\,\mathrm{d}t \xrightarrow{\mathrm{P}} c_{j,k}, \quad \text{as } \mathrm{T}\to\infty,$$

where $c_{j,k} \in (0,\infty)$ is some constant. Denote by $\mathbf{C}_i = (c_{j,k})_{V\times V}$ the matrix consisting of the entries $\{c_{j,k} : j \in \mathcal{V}, k \in \mathcal{V}\}$. It follows that all entries in $\mathbf{C}_i$ are positive, and

$$\nabla^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) = \mathbf{H}_{i,\mathrm{T}} \xrightarrow{\mathrm{P}} \mathbf{C}_i, \quad \text{as } \mathrm{T}\to\infty. \tag{B.97}$$

This proves the asymptotic convergence in (5.6).

We next show $\mathbf{C}_i \succ \mathbf{0}$ using a proof by contradiction. If $\mathbf{C}_i$ is not positive definite, then there exists a vector $\widetilde{\boldsymbol{u}}$ with $\|\widetilde{\boldsymbol{u}}\| > 0$, such that $\widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}} \le 0$. Then (B.97) implies that

$$\widetilde{\boldsymbol{u}}^\top \mathbf{H}_{i,\mathrm{T}} \widetilde{\boldsymbol{u}} \xrightarrow{\mathrm{P}} \widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}} \le 0, \quad \text{as } \mathrm{T}\to\infty. \tag{B.98}$$

Let $Y_2(t) = \{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{u}}\}^2$. Lemma B.17 verifies that $\mathrm{E}\{\int_0^{R_1} Y_2(t)\,\mathrm{d}t\} > 0$. From Lemma B.15, there exists a constant $c_i \in (0,\infty)$ such that

$$\widetilde{\boldsymbol{u}}^\top \mathbf{H}_{i,\mathrm{T}} \widetilde{\boldsymbol{u}} = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{u}}\}^2 \lambda_i^*(t \mid \mathscr{F}_t)\,\mathrm{d}t \xrightarrow{\mathrm{P}} c_i > 0, \quad \text{as } \mathrm{T}\to\infty, \tag{B.99}$$

which contradicts (B.98). The proof is completed. ■

B.13 Proof of Theorem 6

Before proving Theorem 6, we first show Lemma B.18.

Lemma B.18 (Consistency of M-estimator) *Assume conditions* A1, A2, A3, A4, A5, A6 *and* A7 *in Appendix* B. *As* $\mathrm{T}\to\infty$, *there exists a local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ *of the loss function* $\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ *in* (5.4) *such that* $\|\widehat{\widetilde{\boldsymbol{\beta}}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}})$.

Proof: Let $r_{\mathrm{T}} = \sqrt{1/\mathrm{T}}$ and $\widetilde{\boldsymbol{u}} \in \mathbb{R}^V$. Following the arguments of Theorem 1 in [45], it suffices to show that for any given $\epsilon > 0$, there is a sufficiently large constant $C_\epsilon \in (0,\infty)$ such that, for sufficiently large T the following holds:

$$\mathrm{P}\Big(\inf_{\|\widetilde{\boldsymbol{u}}\|=C_\epsilon} \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^* + r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}) > \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)\Big) \ge 1-\epsilon.$$

Let $\widetilde{\boldsymbol{\beta}}_i = r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}} + \widetilde{\boldsymbol{\beta}}_i^*$ and $\|\widetilde{\boldsymbol{u}}\| = C_\epsilon$. By Taylor's expansion of $\mathcal{L}_{i,\mathrm{T}}(\cdot)$ at $\widetilde{\boldsymbol{\beta}}_i^*$, we get

$$\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) - \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) \equiv I_{1,1} + I_{1,2} + I_{1,3}, \tag{B.100}$$

with

$$I_{1,1} \;=\; \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\widetilde{\boldsymbol{x}}_i(t)^\top r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}} \exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^*\big\}\,\mathrm{d}t - \widetilde{\boldsymbol{x}}_i(t-)^\top r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}\,\mathrm{d}N_i(t)\Big],$$

$$
\begin{aligned}
I_{1,2} &= \frac{1}{2\mathrm{T}} \int_0^{\mathrm{T}} \{\widetilde{\boldsymbol{x}}_i(t)^\top r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}}\}^2 \exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^*\big\}\, \mathrm{d}t, \\
I_{1,3} &= \frac{1}{6\mathrm{T}} \int_0^{\mathrm{T}} \{\widetilde{\boldsymbol{x}}_i(t)^\top r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}}\}^3 \exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^{**}\big\}\, \mathrm{d}t,
\end{aligned} \tag{B.101}
$$

where $\widetilde{\boldsymbol{\beta}}_i^{**}$ lies between $\widetilde{\boldsymbol{\beta}}_i^*$ and $\widetilde{\boldsymbol{\beta}}_i$.

For the term $I_{1,1}$, using (5.5) in Theorem 5 gives that

$$
\begin{aligned}
|I_{1,1}| &\le \Big\| \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} \Big[\widetilde{\boldsymbol{x}}_i(t) \exp\big\{\widetilde{\boldsymbol{x}}_i(t)^\top \widetilde{\boldsymbol{\beta}}_i^*\big\}\, \mathrm{d}t - \widetilde{\boldsymbol{x}}_i(t-)\, \mathrm{d}N_i(t) \Big] \Big\| \|r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}}\| \\
&= O_{\mathrm{P}}(\sqrt{1/\mathrm{T}})\, r_{\mathrm{T}}\, \|\widetilde{\boldsymbol{u}}\|.
\end{aligned} \tag{B.102}
$$

For the term $I_{1,2}$, (5.6) in Theorem 5 implies that

$$
I_{1,2} = r_{\mathrm{T}}^2\, \widetilde{\boldsymbol{u}}^\top \nabla^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) \widetilde{\boldsymbol{u}} = r_{\mathrm{T}}^2\, \widetilde{\boldsymbol{u}}^\top \{\mathbf{C}_i + o_{\mathrm{P}}(1)\} \widetilde{\boldsymbol{u}}. \tag{B.103}
$$

For the term $I_{1,3}$, condition A5 implies that each component of $\widetilde{\boldsymbol{x}}_i(t)$ is bounded above by some positive constant. Hence, we have

$$
|I_{1,3}| \le c\, r_{\mathrm{T}}^3 \|\widetilde{\boldsymbol{u}}\|^3, \tag{B.104}
$$

for some constant $c \in (0, \infty)$.

Combining (B.102), (B.103) and (B.104), we obtain

$$
\begin{aligned}
& \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) - \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) \\
&= I_{1,1} + I_{1,2} + I_{1,3} \\
&= O_{\mathrm{P}}(\sqrt{1/\mathrm{T}})\, r_{\mathrm{T}}\, \|\widetilde{\boldsymbol{u}}\| + r_{\mathrm{T}}^2\, \{\widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}} + o_{\mathrm{P}}(1)\, \|\widetilde{\boldsymbol{u}}\|^2\} + c\, r_{\mathrm{T}}^3\, \|\widetilde{\boldsymbol{u}}\|^3 \\
&= \frac{1}{\mathrm{T}} \{O_{\mathrm{P}}(1)\, \|\widetilde{\boldsymbol{u}}\| + \widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}} + o_{\mathrm{P}}(1)\, \|\widetilde{\boldsymbol{u}}\|^2 + o_{\mathrm{P}}(1)\, \|\widetilde{\boldsymbol{u}}\|^3\}.
\end{aligned} \tag{B.105}
$$

By (B.105), we can choose some large C_ϵ, such that all terms in brackets in (B.105) are dominated by the term $\widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}}$, which is positive by the fact that $\mathbf{C}_i \succ \mathbf{0}$ from Theorem 5. This completes the proof. ■

Now we prove Theorem 6. Let $r_{\mathrm{T}} = \sqrt{1/\mathrm{T}}$ and $\widetilde{\boldsymbol{u}} = (u_0, u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_V)^\top \in \mathbb{R}^V$. Denote by $\ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)$ the objective function in (5.8), i.e.,

$$
\ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) = \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i) + \sum_{j \in \mathcal{V} \setminus i} w_{j,i,\mathrm{T}} |\beta_{j,i}|. \tag{B.106}
$$

Similar to the proof of Lemma B.18, it suffices to show that for any given $\epsilon > 0$, there is a sufficiently large constant $C_\epsilon \in (0, \infty)$ such that, for large T,

$$
\mathrm{P}\Big(\inf_{\|\widetilde{\boldsymbol{u}}\| = C_\epsilon} \ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^* + r_{\mathrm{T}}\, \widetilde{\boldsymbol{u}}) > \ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) \Big) \ge 1 - \epsilon.
$$

From (B.106), we have

$$\begin{aligned} & \ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^* + r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}) - \ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*) \\ \geq\ & \{\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^* + r_{\mathrm{T}}\,\widetilde{\boldsymbol{u}}) - \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)\} + \sum_{j\in \mathrm{Pa}^*(i)} w_{j,i,\mathrm{T}} \cdot (|\beta_{j,i}^* + r_{\mathrm{T}} u_j| - |\beta_{j,i}^*|) \\ \equiv\ & I_1 + I_2. \end{aligned}$$

For the term I_1, (B.105) yields that

$$I_1 = \frac{1}{\mathrm{T}}\{O_{\mathrm{P}}(1)\|\widetilde{\boldsymbol{u}}\| + \widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}} + o_{\mathrm{P}}(1)\|\widetilde{\boldsymbol{u}}\|^2 + o_{\mathrm{P}}(1)\|\widetilde{\boldsymbol{u}}\|^3\}.$$

For the term I_2, by triangle inequality and condition (5.11), we have

$$|I_2| \,\leq\, \sum_{j\in \mathrm{Pa}^*(i)} w_{j,i,\mathrm{T}}\, r_{\mathrm{T}}\, |u_j| \,\leq\, r_{\mathrm{T}}\, \|\widetilde{\boldsymbol{u}}\|_1 \max_{j\in \mathrm{Pa}^*(i)} w_{j,i,\mathrm{T}} \,=\, O_{\mathrm{P}}(1/\mathrm{T})\, \|\widetilde{\boldsymbol{u}}\|_1,$$

which is dominated by $\widetilde{\boldsymbol{u}}^\top \mathbf{C}_i \widetilde{\boldsymbol{u}}/\mathrm{T}$ for a sufficiently large C_ϵ. Hence, we conclude that I_2 is dominated by I_1. The remaining proof is the same as that of Lemma B.18. ■

B.14 Proof of Theorem 7

For a $\sqrt{1/\mathrm{T}}$-consistent estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ of $\widetilde{\boldsymbol{\beta}}_i^*$, for any $\epsilon > 0$, there exists constant C_ϵ, such that for sufficiently large T,

$$\mathrm{P}(\|\widehat{\widetilde{\boldsymbol{\beta}}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \leq r_{\mathrm{T}} C_\epsilon) \,>\, 1-\epsilon. \tag{B.107}$$

Let $r_{\mathrm{T}} = \sqrt{1/\mathrm{T}}$. Recall condition A5 implies that $\widetilde{\boldsymbol{x}}_i(t)$ is bounded above. This together with (B.92) yields that there exists a constant $c \in (0,\infty)$, such that

$$\begin{aligned} \sup_{\widetilde{\boldsymbol{\beta}}_i : \|\widetilde{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \leq r_{\mathrm{T}} C_\epsilon} \Big| \frac{\partial^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i} \partial \beta_{0;i}} \Big| \leq c, \qquad \text{for any } j \in \mathcal{V}\setminus i, \\ \sup_{\widetilde{\boldsymbol{\beta}}_i : \|\widetilde{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \leq r_{\mathrm{T}} C_\epsilon} \Big| \frac{\partial^2 \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i} \partial \beta_{k,i}} \Big| \leq c, \qquad \text{for any } j,k \in \mathcal{V}\setminus i. \end{aligned}$$

Combining this with Taylor's expansion, (B.91), and (B.96), for $j \in \mathcal{V}\setminus i$, we have

$$\begin{aligned} & \sup_{\widetilde{\boldsymbol{\beta}}_i : \|\widetilde{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}_i^*\| \leq r_{\mathrm{T}} C_\epsilon} \Big| \frac{\partial \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial \beta_{j,i}} \Big| \\ \leq\ & \Big| \frac{\partial \mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)}{\partial \beta_{j,i}} \Big| \end{aligned}$$

$$
\begin{aligned}
&+ \sup_{\widetilde{\boldsymbol{\beta}}_i:\|\widetilde{\boldsymbol{\beta}}_i-\widetilde{\boldsymbol{\beta}}_i^*\|\le r_{\mathrm{T}}C_\epsilon}\left[\Big\{\Big|\frac{\partial^2\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial\beta_{j,i}\partial\beta_{0;i}}\Big| + \sum_{k\in\mathcal{V}\setminus i}\Big|\frac{\partial^2\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial\beta_{j,i}\partial\beta_{k,i}}\Big|\Big\}\cdot\|\widetilde{\boldsymbol{\beta}}_i-\widetilde{\boldsymbol{\beta}}_i^*\|\right] \\
\le\ & \Big|\frac{\partial\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i^*)}{\partial\beta_{j,i}}\Big| + V\cdot c\, r_{\mathrm{T}}\, C_\epsilon \\
=\ & O_{\mathrm{P}}(\sqrt{1/\mathrm{T}}) + O(r_{\mathrm{T}}) = O_{\mathrm{P}}(\sqrt{1/\mathrm{T}}).
\end{aligned} \tag{B.108}
$$

Consider $\widetilde{\boldsymbol{\beta}}_i$ in the ball $\{\widetilde{\boldsymbol{\beta}}_i : \|\widetilde{\boldsymbol{\beta}}_i-\widetilde{\boldsymbol{\beta}}_i^*\|\le r_{\mathrm{T}}\,C_\epsilon\}$. For $j\in\mathcal{V}\setminus\{\mathrm{Pa}^*(i)\cup i\}$, if $\beta_{j,i}>0$, then (5.13) and (B.108) yield that

$$
\begin{aligned}
\frac{\partial\ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial\beta_{j,i}} &= \frac{\partial\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial\beta_{j,i}} + w_{j,i,\mathrm{T}}\,\mathrm{sign}(\beta_{j,i}) \\
&\ge -\sup_{\widetilde{\boldsymbol{\beta}}_i:\|\widetilde{\boldsymbol{\beta}}_i-\widetilde{\boldsymbol{\beta}}_i^*\|\le r_{\mathrm{T}}C_\epsilon}\Big|\frac{\partial\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial\beta_{j,i}}\Big| + \min_{j\in\mathcal{V}\setminus\{\mathrm{Pa}^*(i)\cup i\}} w_{j,i,\mathrm{T}} \\
&> 0,
\end{aligned} \tag{B.109}
$$

with probability tending to 1 as $\mathrm{T}\to\infty$. Similarly, if $\beta_{j,i}<0$, we have

$$
\begin{aligned}
\frac{\partial\ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial\beta_{j,i}} &= \frac{\partial\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial\beta_{j,i}} + w_{j,i,\mathrm{T}}\,\mathrm{sign}(\beta_{j,i}) \\
&\le \sup_{\widetilde{\boldsymbol{\beta}}_i:\|\widetilde{\boldsymbol{\beta}}_i-\widetilde{\boldsymbol{\beta}}_i^*\|\le r_{\mathrm{T}}C_\epsilon}\Big|\frac{\partial\mathcal{L}_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)}{\partial\beta_{j,i}}\Big| - \min_{j\in\mathcal{V}\setminus\{\mathrm{Pa}^*(i)\cup i\}} w_{j,i,\mathrm{T}} \\
&< 0,
\end{aligned} \tag{B.110}
$$

with probability tending to 1 as $\mathrm{T}\to\infty$.

By (B.109) and (B.110), the following argument holds with probability tending to 1 as $\mathrm{T}\to\infty$: for all $\widetilde{\boldsymbol{\beta}}_i$ with $\|\widetilde{\boldsymbol{\beta}}_i-\widetilde{\boldsymbol{\beta}}_i^*\|\le r_{\mathrm{T}}C_\epsilon$ and all $j\in\mathcal{V}\setminus\{\mathrm{Pa}^*(i)\cup i\}$, $\partial\ell_{i,\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i)/\partial\beta_{j,i}$ has the same sign as $\beta_{j,i}$. Together with (B.107) and the first order condition of $\widehat{\widetilde{\boldsymbol{\beta}}}_i$, it follows that

$$
\mathrm{P}\big(\widehat{\beta}_{j,i}=0,\ \text{for all } j\in\mathcal{V}\setminus\{\mathrm{Pa}^*(i)\cup i\}\big)\ge 1-2\epsilon \tag{B.111}
$$

holds for sufficiently large T. Since ϵ is arbitrary, letting $\epsilon\to 0$ in (B.111) yields that

$$
\mathrm{P}\big(\widehat{\beta}_{j,i}=0,\ \text{for all } j\in\mathcal{V}\setminus\{\mathrm{Pa}^*(i)\cup i\}\big)\to 1,\quad \text{as } \mathrm{T}\to\infty. \tag{B.112}
$$

Note that the vector $\widehat{\boldsymbol{\beta}}_i^{(\mathrm{II})}$ collects all the components in $\widehat{\widetilde{\boldsymbol{\beta}}}_i$ whose indices belong to the set $\mathcal{V}\setminus\{\mathrm{Pa}^*(i)\cup i\}$. Hence, (B.112) implies that $\mathrm{P}(\widehat{\boldsymbol{\beta}}_i^{(\mathrm{II})}=\mathbf{0})\to 1$, as $\mathrm{T}\to\infty$. This completes the proof. ■

B.15 Proof of Corollary 1

Recall that the true edge set $\mathcal{E}^* \neq \varnothing$ is implied by condition A8. To prove Corollary 1, it suffices to show that for each pair of distinct nodes $(j,i) \in \mathcal{V} \times \mathcal{V}$,

$$\begin{cases} \mathrm{P}((j,i) \in \widehat{\mathcal{E}}_+) \to 1, & \text{if } (j,i) \in \mathcal{E}_+^*, \\ \mathrm{P}((j,i) \in \widehat{\mathcal{E}}_-) \to 1, & \text{if } (j,i) \in \mathcal{E}_-^*, \quad \text{as } \mathrm{T} \to \infty. \\ \mathrm{P}((j,i) \notin \widehat{\mathcal{E}}) \to 1, & \text{if } (j,i) \notin \mathcal{E}^*, \end{cases}$$

If $(j,i) \in \mathcal{E}_+^*$, then (3.3) implies that $\beta_{j,i}^* > 0$. By Theorem 6, we have $\widehat{\beta}_{j,i} \xrightarrow{\mathrm{P}} \beta_{j,i}^* > 0$. Thus, $\mathrm{P}((j,i) \in \widehat{\mathcal{E}}_+) = \mathrm{P}(\widehat{\beta}_{j,i} > 0) \to 1$, as $\mathrm{T} \to \infty$.

If $(j,i) \in \mathcal{E}_-^*$, then similarly as above, we have $\mathrm{P}((j,i) \in \widehat{\mathcal{E}}_-) = \mathrm{P}(\widehat{\beta}_{j,i} < 0) \to 1$, as $\mathrm{T} \to \infty$.

If $(j,i) \notin \mathcal{E}^*$, then (3.2) implies that $\beta_{j,i}^* = 0$. By (5.14) in Theorem 7, we obtain $\mathrm{P}((j,i) \notin \widehat{\mathcal{E}}) = \mathrm{P}(\widehat{\beta}_{j,i} = 0) \to 1$, as $\mathrm{T} \to \infty$. This completes the proof. ■