Statistica Sinica Preprint No: SS-2022-0353

Title	A Community Hawkes Model for Continuous-time Networks with Interaction Heterogeneity
Manuscript ID	SS-2022-0353
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0353
Complete List of Authors	Haosheng Shi and Wenlin Dai
Corresponding Authors	Wenlin Dai
E-mails	wenlin.dai@ruc.edu.cn

Notice: Accepted version subject to English editing.

A community Hawkes model for continuous-time networks with interaction heterogeneity

Haosheng Shi and Wenlin Dai*

Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China

Abstract: Continuous-time networks have attracted significant attention due to their widespread applications in various disciplines. A rich literature considers the community structure of the nodes, while few have accounted for the node heterogeneity of interaction propensities. To simultaneously account for both the self-exciting feature and the node heterogeneity, we propose a model based on the Hawkes process, which allows the interaction intensity to vary flexibly with incurred nodes and their affiliated communities. We derive the likelihood function using the immigration–birth representation of the Hawkes process and develop an innovative expectation–maximization algorithm with membership refinement to tackle the computational challenge. Further, we establish the consistency of parameter estimation under mild assumptions. The effectiveness of our model is validated by extensive simulation studies on synthetic data as well as two real-world applications.

Key words and phrases: Community structure; Dynamic network; EM algorithm; Hawkes process; Node heterogeneity.

*Corresponding author

1. Introduction

Networks, especially dynamic networks, have attracted enormous research interest recently. In dynamic networks, nodal linkages are not fixed but may appear or disappear over time. Dynamic networks can be further divided into two types: discrete-time and continuous-time networks. The difference between these two types lies in the fact that the time points of interactions are either continuous-valued or discrete-valued. A crucial task of network analysis is understanding its generative mechanism. Specifically, we are interested in the linking probability of node pairs for static networks and the interaction frequency for dynamic networks. Statistical models inferred from these networks can then be used to discover existing generation patterns, to predict future links or to generate synthetic but realistic networks.

There has been significant research on static network generative models. The simplest model may be the Erdős–Rényi model in which all nodes are considered homogeneous (Erdős and Rényi (1959) and Erdős and Rényi (1960)). A natural extension accounting for nodes' inclination to cluster is the stochastic block model (SBM) proposed in Holland et al. (1983). It partitions nodes into blocks and assigns each block pair a specific linking probability. However, these simple models do not consider the possible degree heterogeneity, which motivates more relevant works, such as the β

model (Holland and Leinhardt (1981)), the degree-corrected block model (DCBM, Karrer and Newman (2011)) and the popularity-adjusted block model (PABM, Sengupta and Chen (2018)).

An analogy can be drawn between dynamic and static networks. A prevalent generalization is to combine the static models with stochastic processes. For example, there have been works discussing how to combine the SBM with a Markov structure on the node labels of discrete-time networks, such as Yang et al. (2011) and Matias and Miele (2017). When it comes to continuous-time networks, a regular practice assumes that events occur according to a continuous-time point process, in which the interaction intensity parallels the static linking probability. For example, Zhang et al. (2017) supposes that the connection between every pair of nodes obeys a continuous-time Markov process, the constant transition rate of which is determined by latent structures such as ER models or SBM. Matias et al. (2018) models interactions on each node pair by inhomogeneous Poisson processes based on the corresponding node labels, and thus each process in a block pair shares a common intensity function.

A common property in continuous-time networks is burstiness (Goh and Barabási (2008)) that events tend to cluster in time. Recently, several point processes have been designed to characterize this property, such as

Wu et al. (2022) and Zhang et al. (2023). The Hawkes process is the most commonly used one among these models because of its simple form with the self-excitation property (Zipkin et al. (2016)). Many studies have combined this process with static generative models to construct continuous-time network models. Some models deal with events related to nodes, such as Fox et al. (2016) and Delattre et al. (2016) with self excitation and Fang et al. (2023) and Chen et al. (2017) with mutual excitation. This kind of models is also appropriate for discovering triggering effects and recovering the latent network structure from time-to-event data, such as Cai et al. (2022), Nickel and Le (2020) and Bacry et al. (2020). Other models pinpoint both incurred nodes of the interaction instead, and can be further subdivided into several types. Models of the first type makes various attempts to incorporate mutual excitation between node pairs, such as Miscouridou et al. (2018), Yang and Koeppl (2020) and Huang et al. (2022) for reciprocal excitation, Passino and Heard (2023) with excitation of adjacent edges and Blundell et al. (2012), Junuthula et al. (2019), and Soliman et al. (2022) with excitation at the block level. However, when we focus on the interaction prediction of a given node pair, all these methods have to borrow historical information from other node pairs, thus complicating the analysis. An alternative type considers individual edges as basic units and attaches each of them with a univariate Hawkes process. Therefore, relationships between different node pairs are embodied in shared intensity parameters instead of shared historical interactions, which greatly simplifies the procedure. Some of these methods, such as Wu (2019), consider nodal linking propensity parameters so that the node heterogeneity is highlighted. Others, such as Arastuie et al. (2020), emphasize the community structure and assume shared parameters within the same block pair.

This paper aims to design an elegant and flexible continuous-time model embodying both community structure and nodal interaction heterogeneity. Consider a social network such as Twitter or other forums to motivate our model setting. Users of these networks tend to form some kind of social circles, such as communities of football, baseball, or basketball fans. While most users may maintain a relatively low presence, users known as activists in every community interact with others frequently, causing node heterogeneity in communities. Additionally, some football fans may interact more frequently with basketball fans than with baseball fans, while others behave in an opposite way. Therefore, the heterogeneity should be flexible enough across nodes and communities, so that the relative frequency of interaction when linking to a specific fan community, relevant to both the node itself and the target community, varies even for fans in the same community as in PABM.

Our main contributions in this work are as follows. First, we propose the community heterogeneous Hawkes independent pairs model (CHHIP). This model is constructed under the framework of independent dyads, so it does not need to consider the complicated mutual excitation and provides a more tractable procedure. Compared with existing models in this range, CHHIP is more flexible in modeling nodal interaction heterogeneity across different communities, and thus more applicable in practical settings. Second, we derive the complete likelihood function based on the immigration–birth representation of the Hawkes process and develop an innovative modified expectation–maximization (EM) algorithm with membership refinement to tackle the computational challenge. Besides, we propose an initialization method for the community structure based on subspace sparse clustering, and provide a criterion for selecting the number of communities. Finally, we prove that all estimators, including the community labels, are consistent as the time goes to infinity. The superiority of the proposed model is demonstrated through synthetic and real data examples.

The rest of the paper is organized as follows. In Section 2, we review some closely related works as the basis of our model. In Section 3, we formally present the CHHIP model and derive its likelihood function.

We also re-express the likelihood using the immigration–birth representation and provide an EM algorithm to alleviate the difficulty in finding the maximum likelihood estimators(MLE). In Section 4, we establish the consistency of parameter estimation and discuss the property of the parameter initialization method. We apply our method in Section 5 to two real-world applications. We conclude the paper with a discussion of directions for future work. Technical proofs for our theoretical results and simulation experiments to evaluate our method are provided in the Supplement. We provide the computer codes for this paper in a public gitee repository: https://gitee.com/bluesun2019/chhip.

2. Background

2.1 Notations

A static network is usually represented as $\mathcal{G} = \mathcal{G}_N(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of N nodes and $\mathcal{E}$ is the set of edges. An adjacency matrix $\mathbf{A}$ represents the linking relationship between nodes, where $\mathbf{A}_{ij} = 1$ indicates the presence of a link between nodes i and j and $\mathbf{A}_{ij} = 0$ indicates the reverse. Sometimes a specific weight w_{ij} is attached to each edge (i, j) in these networks. In this case, we use $\mathbf{W}$ to denote the weighted adjacency matrix with $(\mathbf{W})_{ij} = w_{ij}$. In comparison, a temporal or dynamic network often takes the form of a list of triplets (i, j, t) denoting the interaction timestamp (or duration) t incurred by nodes i and j. We use N to denote the number of nodes and M to denote the number of triplets in the dynamic network. For an integer $z > 1$, define $[z] = \{1, 2, \cdots, z\}$. In a network with K communities, we denote the membership of the node i as $c_i \in [K]$.

2.2 Block models for static networks

Many statistical models have been designed for static networks. Among these models, block models have received enormous attention for their functionality in capturing the community structure. The most classical block model is the SBM firstly proposed by Holland et al. (1983). Each node lies in one of the K communities under a K-block SBM. Under this simple model, each node pair (i, j) links independently with the probability $p_{ij} = P_{c_i c_j}$, where P is a $K \times K$ blockwise connection probability matrix. Obviously, nodes in the same community are assumed to be equivalent in creating links. DCBM (Karrer and Newman (2011)) has been proposed to account for the individual effect on creating links, where $p_{ij} = \theta_i W_{c_i c_j} \theta_j$, $W_{c_i c_j}$ is a group-level connection parameter between communities c_i and c_j, and θ_i adjusts the specific degree for node i.

The DCBM, however, is not flexible enough. In this model, nodes in the same community can only link to other communities with a fixed ratio. This is impractical, especially when we consider the case in which nodes have different connective inclinations toward different communities. To see this more clearly, define the node popularity of i in community a as $\eta_{ia} = \sum_{c_k=a} \mathrm{E}[A_{ik}]$. Then, the definition of the DCBM leads to a shared ratio for i, j in the same community:

$$\frac{\eta_{ia}}{\eta_{ib}} = \frac{W_{c_i a}}{W_{c_i b}} = \frac{W_{c_j a}}{W_{c_j b}} = \frac{\eta_{ja}}{\eta_{jb}}.$$

A more generalized model, PABM (Sengupta and Chen (2018)), successfully fixes the issue. In a K-block PABM, K linking parameters $\{\lambda_{ik}\}_{k=1}^{K}$ are assigned to node i. For any $i > j$, the connection probability between i and j is $p_{ij} = \lambda_{ic_j}\lambda_{jc_i}$. To ensure identifiability, it is assumed that $\Lambda_{ab} = \Lambda_{ba}$ with $\Lambda_{ab} = \sum_{c_i=a} \lambda_{ib}$. In this model, the nodal distinction in linking to different communities is modeled in a more flexible way:

$$\frac{\eta_{ia}}{\eta_{ib}} = \frac{\lambda_{ia} \cdot \Lambda_{c_i a}}{\lambda_{ib} \cdot \Lambda_{c_i b}} \neq \frac{\lambda_{ja} \cdot \Lambda_{c_j a}}{\lambda_{jb} \cdot \Lambda_{c_j b}} = \frac{\eta_{ja}}{\eta_{jb}}.$$

It is worth noting that the SBM and the DCBM are both special cases of the PABM. However, the PABM introduces more parameters than the previous ones, thus providing greater flexibility in fitting the network.

2.3 Hawkes process

Many dynamic networks in our daily lives exhibit a *burstiness* pattern, namely the occurrence of one event increases the probability of subsequent events. The Hawkes process (Hawkes, 1971) is widely used to capture such a phenomenon due to its self-exciting nature. For a Hawkes process $X(t)$ defined on $[0, T]$, we denote a realization as $\{t_1, t_2 \cdots, t_m\}$ where $0 \leq t_1 \leq \cdots \leq t_m \leq T$ and define $N(t) = \sum_{s=1}^{m} 1_{\{t_s<t\}}$ as its corresponding counting process. Given $\mathcal{H}(t) = \{t_s \mid t_s < t, s = 1, \cdots, m\}$, the arrival history up to time t, the conditional intensity takes the form of

$$\lambda^*(t \mid \mathcal{H}(t)) = \mu + \int_0^t \gamma(t-v)\mathrm{d}N(v) = \mu + \sum_{t_s<t} \gamma(t - t_s),$$

where $\mu > 0$ can be regarded as the background intensity or exogenous arrival rate of events, and $\gamma(\tau)$ is a non-negative self-excitation function to capture the endogenous triggering mechanism of past events. One typical choice of $\gamma(\cdot)$ is the exponential kernel $\gamma(\tau) = \alpha e^{-\beta\tau}$. α and β denote the jump size and the decay rate of the intensity, respectively. In this case, the conditional intensity of interaction increases by α immediately when an event occurs, and this effect decays in an exponential rate β with time.

2.4 CHIP model

The CHIP model was proposed by Arastuie et al. (2020) to provide a tractable tool for temporal network analysis. Under this model, nodes are partitioned into several communities as in block models. Interactions on each dyad independently follow a Hawkes process, whose parameters are related to affiliated communities of incurred nodes. In detail, the CHIP model classifies a node i to the block $c_i \in [K]$. Given the parameters $\mu_{ab}, \alpha_{ab}, \beta_{ab}$ attached to each block pair (a, b), nodes i and j interact with the conditional intensity as follows:

$$\lambda_{ij}^*(t \mid \mathcal{H}(t)) = \mu_{c_i c_j} + \int_0^t \alpha_{c_i c_j} e^{-\beta_{c_i c_j}(t-v)} \mathrm{d}N_{ij}(v).$$

Although this model has a simple form and can be solved quickly using the method of moments, sharing a parameter at the block level indicates a deficiency in node heterogeneity. Realistic situations where some nodes are much more sociable than others motivate us to design a new model.

3. CHHIP model

3.1 Definition

We propose the CHHIP model, which considers both node heterogeneity and the community structure. We focus on undirected dynamic networks without self-loops to simplify the model. Similar to the CHIP model, contacts between each node pair are assumed to follow the Hawkes process independently. By applying our model to undirected networks, we do not make a distinction between opposite interactions on each dyad, which is natural in some real data cases (see the example in Section 5.2). In this sense, we assume that the self-exciting and reciprocal effects on subsequent events exist simultaneously and are equal for simplicity. Given the block partition $\{c_i\}_{i=1}^N$, the intensity function of this process is defined as follows:

$$\lambda_{ij}^*(t \mid \mathcal{H}(t)) = \lambda_{ic_j}\lambda_{jc_i} + \int_0^t \alpha_{c_ic_j} e^{-\beta_{c_ic_j}(t-v)} \mathrm{d}N_{ij}(v), \quad 1 \leq j < i \leq N. \tag{3.1}$$

For convenience, we let $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ and represent the corresponding parameter space as Θ. All of the parameters in $\boldsymbol{\theta}$ should be non-negative. Also, we let $\alpha_{ab} = \alpha_{ba}$, $\beta_{ab} = \beta_{ba}$, and $\Lambda_{ab} = \Lambda_{ba}$ $(a, b \in \{1, 2, \cdots, K\})$ for model identifiability.

Unlike the CHIP model, CHHIP accounts for the popularity heterogeneity of the background intensity with a PABM-like term. This modification improves the estimation of routine communication frequency between two specific individuals by endowing each node with one of the K node-specific linking parameters. Nodes in the same community are no longer assumed to be equivalent, and the background intensity incorporates the specific linking pattern of a given node to the target community, providing a more flexible way to restore the possible node differences. However, this modification is not trivial, because the lack of stochastic equivalence within a community disables the method of moments used in Arastuie et al. (2020). We develop an EM algorithm to fix this problem. Whenever i interacts with j, a leap of $\alpha_{c_ic_j}$ with an exponential decay is introduced in the intensity to indicate self-excitation. This term incorporates historical information and reflects the burstiness of events. Besides, we assume nodes in the same community have similar self-exciting mechanisms, which is consistent with many other works such as Junuthula et al. (2019) and Arastuie et al. (2020).

3.2 Estimation procedure

Given the interaction timestamps $\{t_{ij}^{(s)}\}_{s=1}^{n_{ij}}$ on each node pair (i, j) $(i > j)$ of the observed network $\mathbf{X}$, we present the log-likelihood function as follows.

Lemma 1. *Assume that the model parameters $\boldsymbol{\theta}$ and the observation $\mathbf{X}$ are defined as above. Then, the log-likelihood function of the model is*

$$\begin{aligned}\ell = \sum_{i=1}^{N} \sum_{j<i} \Bigg\{ \sum_{s=1}^{n_{ij}} \log\left(\lambda_{ic_j}\lambda_{jc_i} + \alpha_{c_ic_j} A_{ij}(s)\right) - \lambda_{ic_j}\lambda_{jc_i} T \\ + \frac{\alpha_{c_ic_j}}{\beta_{c_ic_j}} \left(\sum_{s=1}^{n_{ij}} e^{-\beta_{c_ic_j}(T - t_{ij}^{(s)})} - n_{ij} \right) \Bigg\}\end{aligned} \tag{3.2}$$

where N is the number of nodes in the network, T is the end time, n_{ij} is the number of interactions between i and j up to T and $A_{ij}(s) = \sum_{u=1}^{s} e^{-\beta_{c_ic_j}(t_{ij}^{(s)} - t_{ij}^{(u)})}$.

However, the closed form of the MLE is unavailable because of the term $\log\left(\lambda_{ic_j}\lambda_{jc_i} + \alpha_{c_ic_j} A_{c_ic_j}(s)\right)$. Therefore, we propose an EM algorithm based on the immigration–birth representation to tackle the challenge; see Figure 1 for an intuitive illustration of this representation. Consider a stream of immigrants following the homogeneous Poisson process with intensity λ. The immigrants will continue to generate descendants, and the descendants can reproduce just like their triggering antecedents. All events but immigrants are named natives. More specifically, every point emerging at t_s, whether immigrant or native, independently gives birth to a new point with the conditional intensity function $\alpha e^{-\beta(t-t_s)}$ at time t. By combining all these timestamps, we obtain a realization of the Hawkes process with the intensity function $\lambda^*(t \mid \mathcal{H}(t)) = \lambda + \sum_{t_s<t} \alpha e^{-\beta(t-t_s)}$.

With the immigration–birth representation explained above, we now introduce some latent variables. For each (i, j), the immigrants should appear with intensity $\lambda_{ic_j}\lambda_{jc_i}$. Denote by d_{ij} the total number of immigrants. In many real networks, the number of self-interactions n_{ii} cannot be directly observed so we have not defined them yet. For convenience of subsequent calculations, we pretend there existed unobserved point processes following our model on self-loops and treat d_{ii}, the number of immigrants on the self-loop of i, as a latent variable here. Then conditional on λ_{ic_i}, d_{ii} follows a Poisson distribution with a mean of $\lambda_{ic_i}^2 T$. As described above, every event, whether immigrants or natives, can reproduce descendants, and all natives have their own triggering events. We define $e_{ij}^{(s)}$ as the number of descendants and $\text{trig}_{ij}^{(s)}$ as the triggering event for the sth interaction between nodes i and j $(i > j)$. If the sth interaction is an immigrant to the process, we define its triggering event as itself, namely $\text{trig}_{ij}^{(s)} = s$.

Now the complete likelihood of all interaction timestamps $\{t_{ij}^{(s)}\}_{s=1}^{n_{ij}}$ and these latent variables $\{d_{ij}, \text{trig}_{ij}^{(s)}, e_{ij}^{(s)}\}_{s=1}^{n_{ij}}$ on the node pair (i, j) $(i > j)$ can be divided into three parts:

1. The likelihood of the number of background events (immigrants):

$$L_1(\lambda_{ic_j}, \lambda_{jc_i}) = e^{-\lambda_{ic_j}\lambda_{jc_i}T} \frac{(\lambda_{ic_j}\lambda_{jc_i}T)^{d_{ij}}}{d_{ij}!},$$

2. Given the number of background events, the likelihood of the number of descendants for each event:

$$L_2(\alpha_{c_ic_j},\beta_{c_ic_j}) = \prod_{s=1}^{n_{ij}} \left\{ e^{-G(\alpha_{c_ic_j},\beta_{c_ic_j})} \frac{\left(G(\alpha_{c_ic_j},\beta_{c_ic_j})\right)^{e_{ij}^{(s)}}}{e_{ij}^{(s)}!} \right\},$$

   where the triggering ratio $G(\alpha_{c_ic_j},\beta_{c_ic_j}) = \int_0^\infty \alpha_{c_ic_j} e^{-\beta_{c_ic_j}t}\mathrm{d}t = \alpha_{c_ic_j}/\beta_{c_ic_j}$ is a commonly used approximation of the term $\int_0^{T-t_{ij}^{(s)}} \alpha_{c_ic_j} \exp\{-\beta_{c_ic_j}t\}\mathrm{d}t$ when T is large. This triggering ratio informs the average number of first-generation offspring and is often used as a measure of burstiness. More details are discussed in Lewis and Mohler (2011).

3. Given the number of background events d_{ij} and the number of descendants for each event $\{e_{ij}^{(s)}\}_{s=1}^{n_{ij}}$, the likelihood of the timestamps: As events with $\mathrm{trig}_{ij}^{(s)} = s$ contribute $1/T$ and those with $\mathrm{trig}_{ij}^{(s)} = h_s < s$ contribute $\alpha_{c_ic_j} e^{-\beta_{c_ic_j}\left(t_{ij}^{(s)}-t_{ij}^{(h_s)}\right)} / \int_{t_{ij}^{(h_s)}}^{T} \alpha_{c_ic_j} e^{-\beta_{c_ic_j}\left(z-t_{ij}^{(h_s)}\right)}\mathrm{d}z$, we can still substitute the denominator for $G(\alpha_{c_ic_j},\beta_{c_ic_j})$ and obtain

$$L_3(\alpha_{c_ic_j},\beta_{c_ic_j}) = \prod_{s=1}^{n_{ij}} \left[\left\{ \frac{\alpha_{c_ic_j} e^{-\beta_{c_ic_j}\left(t_{ij}^{(s)}-t_{ij}^{(h_s)}\right)}}{G(\alpha_{c_ic_j},\beta_{c_ic_j})} 1_{\{\mathrm{trig}_{ij}^{(s)}=h_s<s\}} \right\} \cdot \left(\frac{1}{T} 1_{\{\mathrm{trig}_{ij}^{(s)}=s\}} \right) \right].$$

With $\sum_{u=1}^{s} e_{ij}^{(u)} = n_{ij}$, we offset terms and formulate the complete log-likelihood as

$$\begin{aligned}\ell_c(\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{c}) = \sum_{i=1}^{N}\sum_{j<i}\Bigg[&-\lambda_{ic_j}\lambda_{jc_i}T + d_{ij}\log\lambda_{ic_j} + d_{ij}\log\lambda_{jc_i} + \Bigg\{\sum_{s=1}^{n_{ij}}\left(-\frac{\alpha_{c_ic_j}}{\beta_{c_ic_j}}\right) \\ &+ \sum_{s=1}^{n_{ij}}\left\{\log\alpha_{c_ic_j} - \beta_{c_ic_j}(t_{ij}^{(s)} - t_{ij}^{(\mathrm{trig}_{ij}^{(s)})})\right\}1_{\{\mathrm{trig}_{ij}^{(s)}<s\}}\Bigg\}1_{\{n_{ij}\neq 0\}}\Bigg] \\ &+ \frac{1}{2}\sum_{i=1}^{N}\Big[-\lambda_{ic_i}^2T + 2d_{ii}\log\lambda_{ic_i}\Big].\end{aligned} \tag{3.3}$$

With the complete likelihood (3.3), we develop a modified EM algorithm (summarized as Algorithm **??** in the Supplement) to obtain the MLE. Details of parameter updates are elaborated in S1 of the Supplement. A prominent problem here is how to determine the community membership vector $\boldsymbol{c}$. Arastuie et al. (2020) detects the community by applying a simple spectral clustering to an accumulated matrix $\mathbf{W}$ containing the number of interactions n_{ij} on each node pair, which is, however, not suitable for our model due to the degree variation within communities. Here, we regard the membership labels as parameters to be estimated instead of random variables, and update them by coordinate in each EM loop. In this way, the likelihood function remains increasing and the convergence is not disrupted. This technique, similar to which is used in Fang et al. (2023) and Soliman et al. (2022), can be seen as a community membership refinement step.

To select the number of communities, we use Hannan and Quinn information criterion (HQ, Hannan and Quinn (1979)), which is defined as $\mathrm{HQ}(K) = -2\ell + 2\left(NK + N(N+1)\right)\log\log M$ in our case, where ℓ is defined in Lemma 1. Chen et al. (2018) studies the performance of this criterion under Hawkes process models. Generally this criterion imposes a penalty lying between AIC and BIC, and has the slowest growing rate with the network size among strongly consistent information criterion taking the form of $\mathrm{IC}(K) = -2\ell + c(K, M)$. An information criterion is consistent if there is a $j \in [N]$ such that $\lim_{M\to\infty} P\left\{\min_{k\neq j}\left(\mathrm{IC}\left(k\right)\right) - \mathrm{IC}\left(j\right)\right) > 0\right\} = 1$, and $c(K, M)$ is a penalty term imposed on the community number K and the event number M. HQ performs well in our simulation studies and real data analysis, so we find it a suitable choice for model selection.

As the complete likelihood is non-convex, reasonable initial values for parameters should be selected. For node membership, we follow Arastuie et al. (2020) to obtain an accumulated matrix and perform a subspace sparse clustering (Noroozi et al. (2021)). Proof has been established that this estimator obtained by SSC is consistent under the weighted PABM model in the Section S2 of the Supplement, so we believe it is a suitable and valid initializer. The process of community detection is presented in Algorithm **??** (in the Supplement). For other intensity parameters, we pretend all

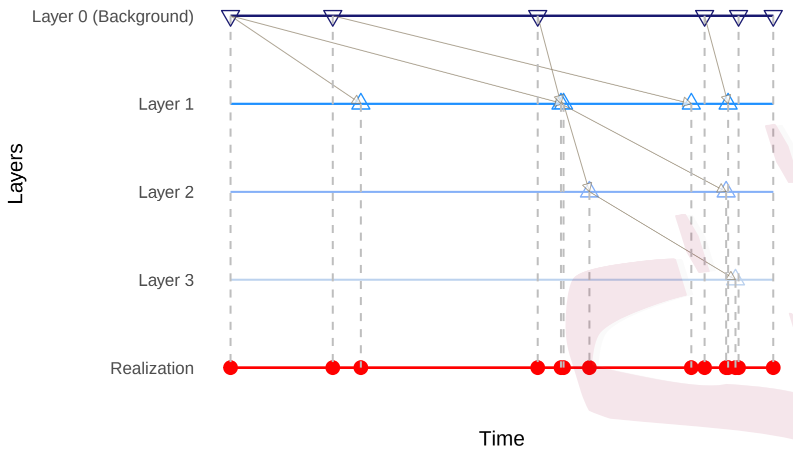


Figure 1: An illustration of the immigration–birth process. Events in layers indexed larger than 0 are triggered by their antecedents in the previous layer, and all of them constitute the realization process.

nodes were equivalent and take the moment estimators as initial values, which turns out to be effective in simulation studies. The procedure of initialization is the same as that in Arastuie et al. (2020): We calculate the mean and variance of interaction counts in each block pair, and use them to obtain the moment estimators. The only difference is that our initial block labels are obtained from SSC instead of spectral clustering.

4. Theoretical analysis

We now consider the asymptotic property of the MLE when the observation duration is long enough. The reasonableness of initializing communities by sparse subspace clustering is included in S2 of the supplement because of the space limit. In this section, we show that the community membership and the parameter estimation are both consistent with time. By combining these arguments, we establish the validity of this estimation procedure.

4.1 Consistency of parameter estimation

Suppose the dynamic network $\mathbf{X}$ is generated from CHHIP$(\boldsymbol{c}^0, \boldsymbol{\theta}^0)$, where $\boldsymbol{c}^0 \in \mathbb{C} = \{1, 2, \cdots, K^0\}^N$ and $\boldsymbol{\theta}^0 = (\boldsymbol{\lambda}^0, \boldsymbol{\alpha}^0, \boldsymbol{\beta}^0)$ with its parametric space Θ. Define the complete intensity for each entry as $\lambda^{**}_{i,j}(\boldsymbol{c}^0, \boldsymbol{\theta}^0, t, w) = \lambda^0_{ic_j}\lambda^0_{jc_i} + \alpha^0_{c_ic_j}\int_{-\infty}^{t} e^{-\beta^0_{c_ic_j}(t-v)} dN_{ij}(v)$, where w belongs to the sample space Ω including possible outcomes of the Hawkes process that is defined on the entire real line and $N_{ij}(t)$ is the corresponding counting process, and use this intensity to approximate the exact intensity (Ogata et al. (1978)). To derive the consistency of the MLE, we require assumptions below.

Assumption 1. $\lambda^0_{ia} > 0, \quad 0 < \alpha^0_{ab}/\beta^0_{ab} < 1$ for any $i \in \{1, \cdots, N\};\ a, b \in \{1, \cdots, K^0\}$.

Assumption 2. The parameter space Θ for $\boldsymbol{\theta}$ is a compact space, i.e., there exists $0 < m \leq M$ so that $m \leq \lambda_{ia} \leq M, m \leq \alpha_{ab} \leq M, m \leq \beta_{ab} \leq M$ for any $i \in \{1, \cdots, N\}$ and $a, b \in \{1, \cdots, K\}$.

Assumption 3. For any $\boldsymbol{\theta} \in \Theta$, there exists a neighborhood $U(\boldsymbol{\theta})$ such that for all $\boldsymbol{\theta}' \in U(\boldsymbol{\theta})$ and $\boldsymbol{c} \in \{1, 2, \cdots, K\}^N$:

$$\left|\lambda^{**}_{i,j}(\boldsymbol{\theta}', \boldsymbol{c}, 0, \omega)\right| \leq M_0(\boldsymbol{\theta}, \boldsymbol{c}, 0, \omega),$$

where M_0 is an L_1 integrable random variable.

Assumption 1 stems from a sufficient condition for the existence of a stationary and ergodic Hawkes process on each entry, established in Brémaud and Massoulié (1996). Our theoretical proof is built on these two properties of the point process as in Ogata et al. (1978). Assumptions 2 and 3 can be regarded as restrictions on the scale of parameters. As we focus on the case of a finite number of nodes, Assumption 2 can be easily satisfied whenever Assumption 1 holds. Many works such as Ogata et al. (1978) have used assumptions similar to Assumption 3 as an initial condition for stationary Hawkes process. It indicates that events from the infinite past will not affect the intensity at present, which is practically true. Specifically, for any finite S, we can always find a sufficiently small $U(\boldsymbol{\theta}, \delta(\boldsymbol{\theta}))$ so that $|\lambda^{**}_{i,j}(\boldsymbol{\theta}', \boldsymbol{c}, [-S, 0), w)|$, defined as $\left|\lambda_{ic_j}\lambda_{jc_i} + \alpha_{c_ic_j}\int_{-S}^{0} e^{-\beta_{c_ic_j}(t-v)} dN_{ij}(v)\right|$, can be bounded by $\max_{\boldsymbol{c}\in\mathbb{C}} \lambda^{**}_{i,j}(\boldsymbol{\theta}, \boldsymbol{c}, [-S, 0), w) + \epsilon$, and the former is L_1 integrable as $\mathbb{E}(\lambda^{**}_{i,j}(\boldsymbol{\theta}, \boldsymbol{c}, [-S, 0))) \leq M^2 + M/m\mathbb{E}(N_{ij}([0, 1)))$ (Puri and Tuan (1986)). Therefore, if there are no effects of the infinite past $(-\infty, -S)$ on the current intensity, Assumption 3 naturally holds.

To obtain the consistency of the community membership, an additional assumption is required to ensure identifiability of the community labels. For a block pair (a, b), We define $\boldsymbol{V}^{(a,b)} \in \mathbb{R}^{N_a}$ as $V_r^{(a,b)} = \lambda_{i_a^r,b}$, where i_a^r is the rth node in block a. This vector can be regarded as the intensity heterogeneity parameter of nodes in block a towards block b.

Assumption 4. For any $k = 1, \cdots, K^0$, vectors $\boldsymbol{V}^{(k,1)}, \cdots, \boldsymbol{V}^{(k,K^0)} \in \mathbb{R}^{n_k}$ are linearly independent.

This condition is also used by Noroozi et al. (2021), which guarantees that communities are identifiable with the noiseless matrix $\boldsymbol{\mu}^0 = [\lambda^0_{ic^0_j}\lambda^0_{jc^0_i}]_{i,j}$. Intuitively, it ensures that the column (row) space of $\boldsymbol{\mu}^0$ can be uniquely represented as the union of K^0 linearly independent subspaces, each of which has rank K^0. Therefore, the node partition where nodes lying in the same subspace belong to the same cluster is unique based on the PABM structure.

Theorem 1. *Under Assumptions 1–4, given the true number of communities $K = K^0$, the MLE $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{c}}$ is consistent up to a permutation $\sigma(\cdot)$ on $\{1, \cdots, K^0\}$:*

$$\widehat{\boldsymbol{\theta}} \xrightarrow{P} \sigma(\boldsymbol{\theta}^0) \ \textit{and}\ \widehat{\boldsymbol{c}} \xrightarrow{P} \sigma(\boldsymbol{c}^0), \qquad \textit{as} \quad T \to \infty,$$

where $\sigma(\boldsymbol{\theta}^0)$ *represents the label permuted version of* $\boldsymbol{\theta}^0$.

The proof is provided in S3.1 of the Supplement. Theorem 1 guarantees that given the community number, we can always obtain consistent estimators with respect to time duration T.

5. Real data analysis

5.1 Analysis of EU e-mail network

We analyze the EU e-mail network data from the Stanford Large Social Network Dataset (http://snap.stanford.edu/data/email-Eu-core-temporal.html). The network covers e-mail communication within a large European research institution. Nodes in this network represent the core institution members. Two nodes interact with each other once an e-mail is sent between the corresponding members. The affiliated department information for each member is also available, and can be regarded as the ground community structure. Members within the same department interact more frequently. In addition, each department has highly and poorly sociable members, leading to node heterogeneity. Although there might be more than one community in a department, we will show that simply viewing them as a whole community also works well for inference of CHHIP, thanks to the flexibility of endowing each node with K different baseline intensities.

In this study, we choose the most densely connected part consisting of 4 departments and omit the last 30% of the time duration when there are few interactions. We therefore obtain a sampled temporal network with 116 nodes and 55170 events over 563 days. As our method is designed for undirected networks, we make all interactions undirected. Figure 2 visualizes the network from two perspectives: plot (a) illustrates the community structure and the node heterogeneity of this network, while plots (b) and (c) exhibit the network burstiness pattern. Specifically, we use the signed total variation distance Δ proposed by Goh and Barabási (2008) from a baseline distribution, the Poisson distribution, to measure the burstiness level for each node pair. Δ is bound in $[-1, 1]$, and its magnitude indicates the level of burstiness: $\Delta = 1$ corresponds to the most bursty signals, $\Delta = 0$ implies a pattern similar to the Poisson, and $\Delta = -1$ represents a completely regular signal. From (b), we observe that although interactions are bursty on most dyads (see the first three dyads in (c)), there is still a sizeable proportion on which interactions are anti-bursty (see the last dyad in (c)).

First, we compare the communities detected by our proposed method with results from spectral clustering and ppSBM (Matias et al. (2018)), another popular community detection method for continuous-time networks.

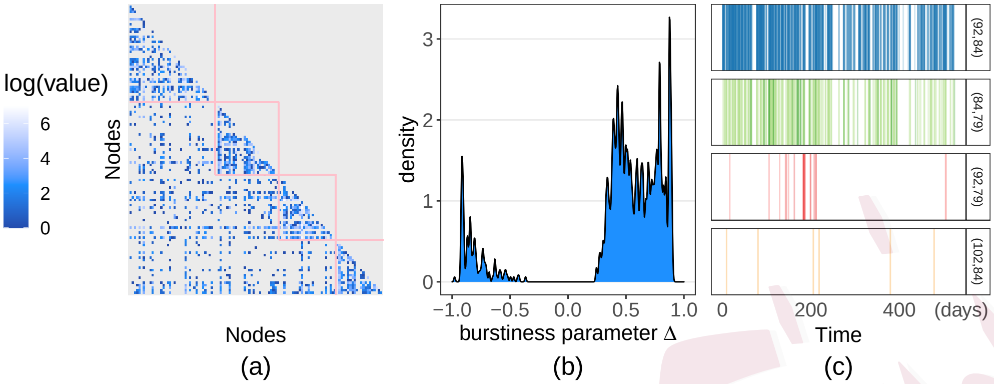
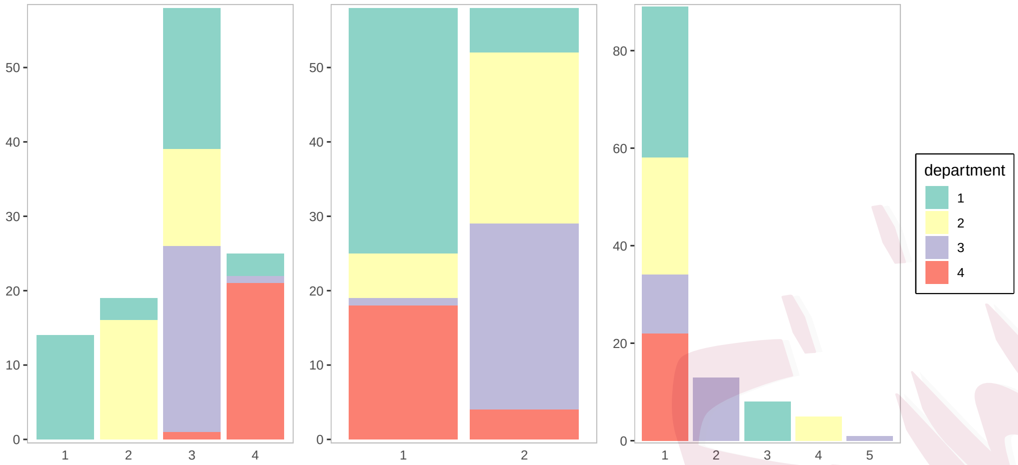


Figure 2: Visualization of the network. (a) shows the accumulated network, (b) shows the distribution of burstiness parameters on edges with at least two interactions and (c) shows the timestamps of events on four selected dyads. Dark and light colors in (c) correspond to different directions on this edge.

For spectral clustering, we simply accumulate the dynamic network as a weighted static network and perform this method on it. The number of communities is chosen by maximizing the famous average silhoutte width (Rousseeuw (1987)). The ppSBM method combines the stochastic block model and semi-parametric Poisson intensity functions. This work assumes that node pairs in the same community pair share the same intensity function, and proposes a semi-parametric variational EM method to estimate these functions. The number of communities is selected with the integrated classification likelihood criterion.

We select the community number from 2 to 5 using the three methods and present the results in Figure 3. Different colors in the plot correspond to members from different departments of the institution. The spectral clustering method focuses on finding statistically equivalent nodes in interaction counts and overlooks the dynamic information, so it falsely selects $\hat{K} = 2$ to mix all departments up, which is apparently not reasonable. The ppSBM method considers the dynamic information instead, and attempts to cluster statistically equivalent nodes in linking capability into the same cluster. However, this idea imposes an overly strong assumption for nodes in the same community, indicating that the method tends to cluster nodes into many small communities. When we restrict the community number within 5, it leads to a super unbalanced community membership. In comparison, our method is the only one to choose the correct $K = 4$. Accounting for the possible linking heterogeneity helps prevent communities containing only a few extreme nodes and produce a relatively balanced community structure.

We further evaluate our proposed CHHIP model as well as the CHIP and ADCHP models on this dataset by making predictions based on the entire history. The ADCHP model, short for additive degree corrected Hawkes process, is designed to characterize networks with node heterogeneity but without community structure. It can be deemed as a dynamic version of the static beta model. In this model, triggering parameters are identical on all node pairs and the background intensity is of an additive form $\gamma_i + \gamma_j$.



(a) CHHIP. (b) spectral clustering. (c) ppSBM.

Figure 3: Community detected on the EU e-mail network by CHHIP, spectral clustering and ppSBM.

For CHIP and CHHIP, we use both the department information (methods beginning with “department”) and the aforementioned data-driven results (methods beginning with “spectral” or “SSC”) as their community structures. The number of communities is still selected from 2 to 5 for the data-driven clustering methods. We first set a split point of time T_{split}, and use all timestamps before T_{split} to train the models. Then, we use these models to predict whether there exist interactions in the time interval $[T_{\text{split}}, T_{\text{split}} + \pi_{\text{duration}})$ for each edge. We choose $\pi_{\text{duration}} = 50$ days as in Yang and Koeppl (2020). Let the length of the training period vary from 40% to 90% of the whole interval. The probability that there is at least one interaction on (i, j) in our target time interval can be calculated as

$$1 - \exp\left\{ - \int_{T_{\text{split}}}^{T_{\text{split}}+\pi_{\text{duration}}} \widehat{\lambda}^*_{ij}(t \mid \mathcal{H}(T_{\text{split}}))dt \right\}$$

where $\widehat{\lambda}^*_{ij}(t \mid \mathcal{H}(t))$ is the intensity function estimated for node pair (i, j).

With this probability, we can obtain the rolling prediction precision in Figure 4, including the average area under the curve (AUC) of both the precision–recall (PR) curve and the receiver operating characteristic (ROC) curve. For convenience, we denote them as AUC-PR and AUC-ROC here. Both measurements are considered because although a classifier with a high AUC-ROC may perform well in predicting positive (or zero) links, it is insensitive to the classification results on negative (or non-zero) links. In comparison, AUC-PR considers non-zero links and is more accurate in unbalanced classification settings, which is more relevant to our case.

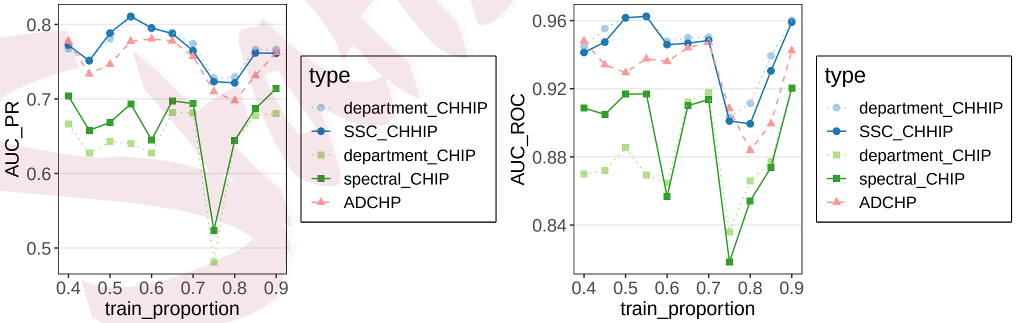
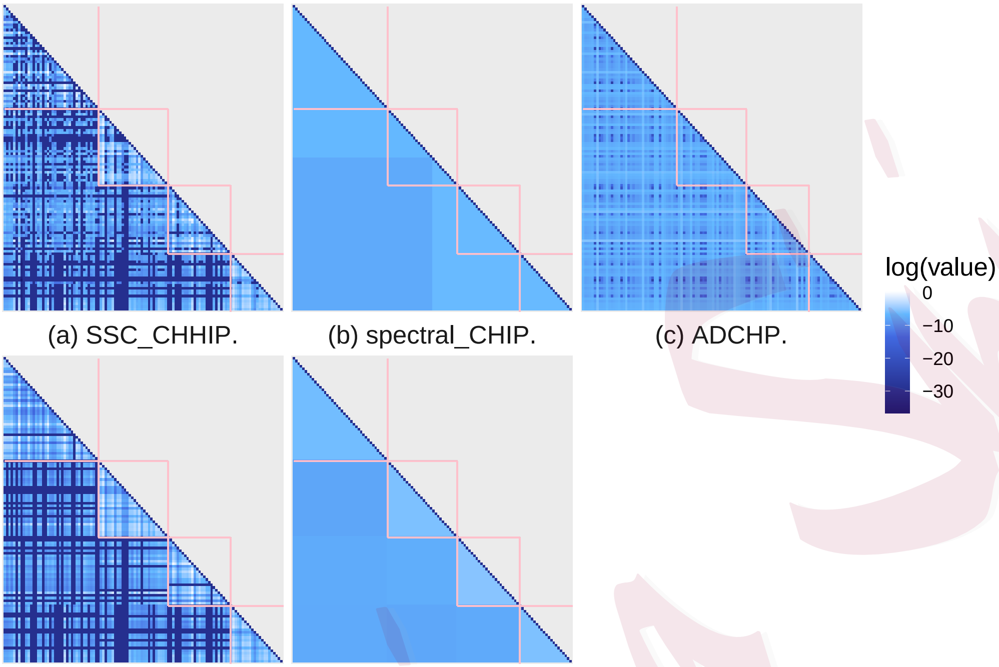


Figure 4: Prediction performance of different methods.

As shown in Figure 4, CHHIP outperforms the other methods on both AUC-ROC and AUC-PR scores, indicating that our model predicts both the presence and absence of the events well. An interesting phenomenon here is that the CHIP model with communities obtained by spectral clustering performs much better than that with departments. In fact, the spectral method tends to cluster nodes that are statistically equivalent to the same block, resulting in a clustering result catering for the use of CHIP model. However, the prediction result of this model is still inferior to ADCHP or CHHIP, due to the ignorance of node heterogeneity obviously present in this dataset. The prediction performance of CHHIP, in contrast to CHIP, remains stable under different community partitions owing to its node-specific parameters with different communities. Actually there might be different community assignments with reasonable interpretations. Our model can be flexibly adaptable by adjusting the value for parameters related to "misclassified" nodes and clusters and thus stabilize the prediction performance.

To compare the interpretability of the three models, we demonstrate the estimated background intensities in Figure 5. Our estimation, under either community structure, is more similar to the accumulated network demonstrated in Figure 2. However, ADCHP and CHIP tend to estimate a small baseline and, thus, require a larger triggering ratio as shown in Table



(a) SSC_CHHIP. (b) spectral_CHIP. (c) ADCHP.

(d) department_CHHIP. (e) department_CHIP.

Figure 5: Background intensity estimated for each entry using ADCHP, CHHIP and CHIP. Nodes have been permuted for ease of comparison according to their department membership in all cases.

Table 1: Triggering ratio α/β estimated for each block pair under CHIP and CHHIP models with given departments as blocks. For ADCHP, the estimated α/β is a constant 0.974 on all dyads.

CHIP CHHIP	Dept 1	Dept 2	Dept 3	Dept 4
Dept 1	0.945 0.683	0.916	0.932	0.829
Dept 2	0.734	0.921 0.603	0.873	0.752
Dept 3	0.872	0.623	0.951 0.752	0.814
Dept 4	0.623	0.647	0.514	0.907 0.689

1. When an event occurs, the large triggering parameter dominates these small background intensities; therefore, ADCHP and CHIP can predict the occurrence of subsequent events well through historical events. However, they cannot properly characterize the studied dynamic network, and consequently are not that suitable to serve as generative models.

We perform another practical predictive analysis to explain the facts above. In practice, it is sometimes hard to obtain historical data immediately. Historical events used for prediction may have occurred one day, one week or even one year previously. Hence, we investigate how CHHIP and ADCHP perform when there are missing events. Specifically, we fix the prediction interval at $[506, 556)$ and shift the end time of training data, T_{split}, from 169 to 506. Figure 6 reveals that our model can perform significantly better than other models when the missing interval is long because we estimate the background intensity more accurately.

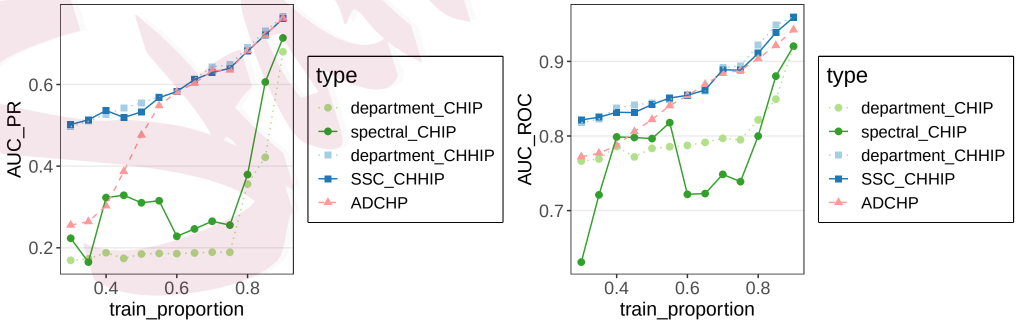


Figure 6: Prediction performance of different methods with missing history.

5.2 Analysis of primary school interaction network

We provide another example to show the performance of our method when the heterogeneity between nodes is not that significant. The dataset was firstly collected by Stehlé et al. (2011) and further analyzed in Matias et al. (2018). It consists of face-to-face contacts of 232 students from 10 classes and 10 teachers in a French primary school for 2 weekdays. Specifically, all participants were given a Radio-Frequency identification badge, which exchange radio packets when they are close to each other. The resolution of this experiment is 20s, which means that only radio packets lasting longer than 20s are considered as interactions in the network.

We focus on the second weekday of the data, combine consecutive con-

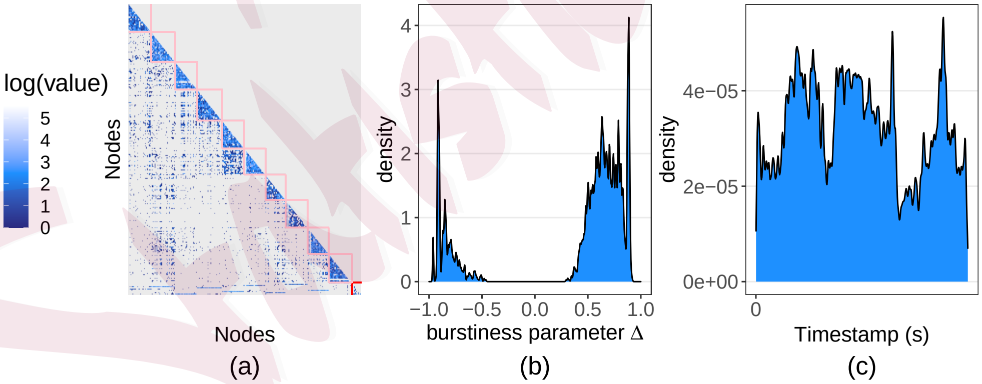
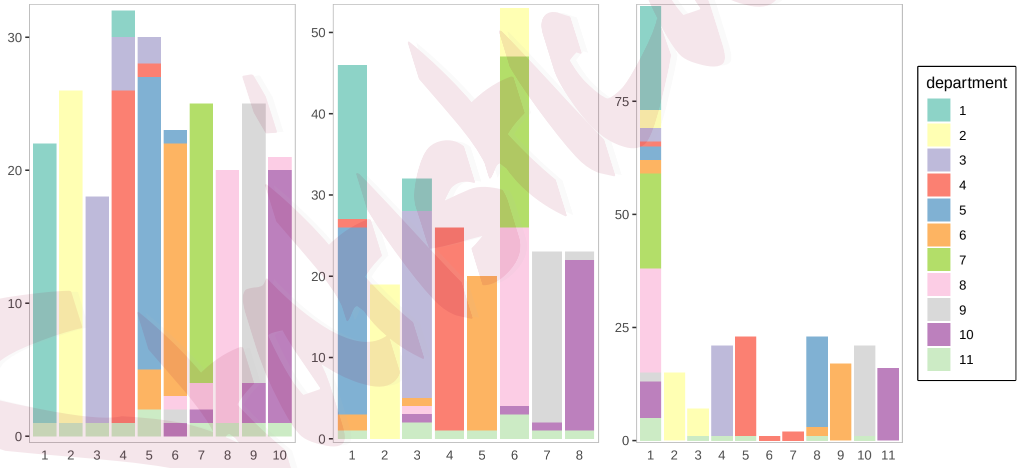


Figure 7: Visualization of the network. (a) shows the accumulated network, (b) shows the distribution of burstiness parameters on edges with at least two interactions and (c) shows the distribution of timestamps.

tacts which are probably identical communication, and obtain an undirected dynamic network with 242 nodes and 48388 events. Figure 7 summarizes the network with the accumulated matrix in plot (a), the distribution of burstiness parameters in plot (b) and and the distribution of timestamps in plot (c). Nodes in the red box and the pink boxes correspond to the teachers and the students in 10 classes, respectively. Note that students in the same class interact much more than students from different classes, leading to a strong community structure. As a result, the node heterogeneity for this network is not as strong as the previous example. The network is strongly bursty as illustrated in plot (b) and suitable for modelling with Hawkes process with similar analysis as discussed in Section 5.1.

First, the clustering results of the three methods discussed in Section 5.1 are presented in Figure 8. As the school has 10 classes, the number of communities are selected from 8 to 12. From Figure 8, the ppSBM method clusters nodes into 11 communities, where some of them are extremely large (for example, cluster 1) while others are small (for example, clusters 6 and 7). This phenomenon is still related to the property of ppSBM to split a community into many small blocks according to their intensity. Matias et al. (2018) shows that a larger community size such as 17 performs better; however, the computational cost of expanding the scope of parameter tuning for this method is rather high. The spectral clustering method also provides reasonable results except that it still mixes two classes up in the cluster 1 and cluster 8. In contrast, our method chooses $K = 10$ correctly, and each community detected mainly corresponds to one natural class of the school. In addition, our method assigns one teacher for each class except the two classes in Grade 3 (classes 5 and 6, the fifth and sixth blocks from top left in Figure 7(a)). Misclassified students (or teachers) are mainly from the other class of the grade that the natural cluster belongs to, and can be regarded as an "outlier" of this class when interacting with others. For example, the 68th node, belonging to the 1st class but misclassified into the 4th class,



(a) CHHIP. (b) spectral clustering. (c) ppSBM.

Figure 8: Clustering performance of different methods. Different colors represent students from different classes (departments 1-10) and teachers (department 11).

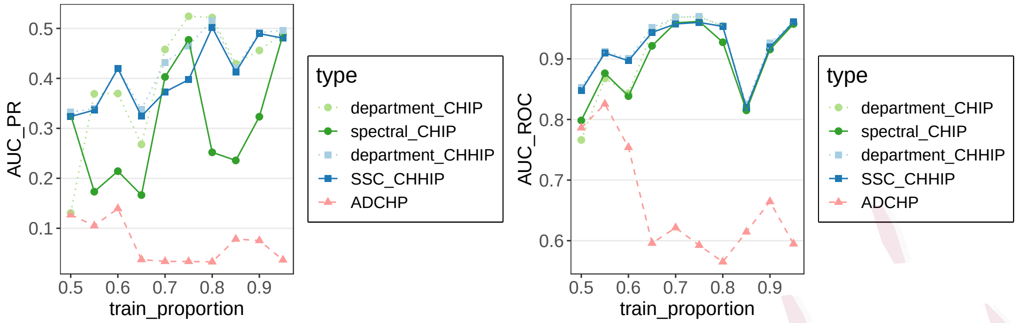


Figure 9: Prediction performance of different methods.

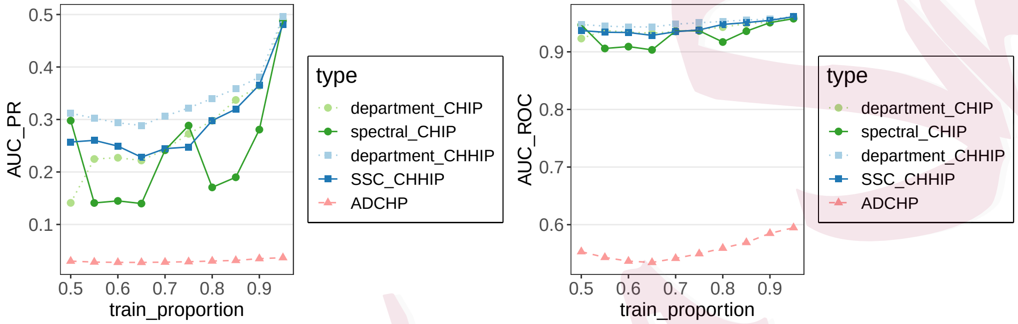


Figure 10: Prediction performance of different methods with missing history.

interacts with the latter class for 123 times which is far more than 23, the average level of the 1st class.

Link prediction of this network is also important for subsequent analysis of, for example, disease transmission in the school. We still perform the preceding two prediction procedures and plot results in Figure 9 and Figure 10. We vary the end time of the training period T_{split} from 50% to 95%. We select the optimal community number from 8 to 12 for SSC_CHHIP and spectral_CHIP methods, and predict the occurrence of interactions in

$[T_{\text{split}}, T_{\text{split}} + \pi_{\text{duration}})$ for Figure 9 and in $[0.95T, 0.95T + \pi_{\text{duration}})$ for Figure 10 where $\pi_{\text{duration}} = 1800$ seconds. We find that on average our method is slightly better than CHIP in both settings, while the ADCHP method has a very bad performance. In this example the community structure dominates node heterogeneity, so the method that only considers node heterogeneity is no longer suitable. The result also indicates that out method maintains its superiority even when the heterogeneity is not that significant.

6. Discussion

We proposed the CHHIP model as a dynamic network generative model designed for systems with community structure and nodal interaction heterogeneity. Compared with the existing literature, our model allows more flexible pair-wise intensities, which better characterizes the node heterogeneity. We derived the complete likelihood based on the immigration–birth representation and designed a modified EM algorithm to tackle the challenge in finding the MLEs. Further, we established the consistency of parameter estimation under mild assumptions. The effectiveness of our model was validated by extensive simulation studies on synthetic data and two real-world datasets.

While the independence assumption in the CHHIP model works well in many cases, it may be too simple to model networks with strong correlations among edges. Works to construct block models of Hawkes processes with reciprocal excitation (Yang and Koeppl, 2020; Huang et al., 2022) provide inspiration for extending the CHHIP model. Additionally, while we focused on undirected dynamic networks to simplify the model and capture reciprocity from opposite directions to some extent, this assumption regards opposite interactions as equivalent and enforces identical triggering effects for both kinds, which is not always the case. One possible solution is to consider a mutually-exciting bivariate Hawkes process on each dyad. Finally, the current EM algorithm for estimating the maximum likelihood estimator is not scalable enough for networks of large sizes. Recent efforts to reduce the burden of inference, such as online learning methods (Fang et al. (2023)), provide a potential solution. It remains to develop more efficient algorithms to reduce computational burdens.

Acknowledgement

The authors wish to extend their heartfelt appreciation to the editor, an associate editor, and two reviewers for their insightful and constructive feedback. Their valuable suggestions significantly enhanced the quality of the manuscript.

References

Arastuie, M., S. Paul, and K. Xu (2020). CHIP: a Hawkes process model for continuous-time networks with scalable and consistent estimation. *Advances in Neural Information Processing Systems 33*, 16983–16996.

Bacry, E., M. Bompaire, S. Gaïffas, and J.-F. Muzy (2020). Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research 21*(50), 1–32.

Blundell, C., J. Beck, and K. A. Heller (2012). Modelling reciprocating relationships with Hawkes processes. *Advances in Neural Information Processing Systems 25*, 2609–2617.

Brémaud, P. and L. Massoulié (1996). Stability of nonlinear Hawkes processes. *The Annals of Probability 24*(3), 1563–1588.

Cai, B., J. Zhang, and Y. Guan (2022). Latent network structure learning from high-dimensional multivariate point processes. *Journal of the American Statistical Association 0*(0), 1–14.

Chen, J., A. Hawkes, E. Scalas, and M. Trinh (2018). Performance of information criteria for selection of Hawkes process models of financial data. *Quantitative Finance 18*(2), 225–235.

Chen, S., A. Shojaie, E. Shea-Brown, and D. Witten (2017). The multivariate Hawkes process in high dimensions: Beyond mutual excitation. *arXiv preprint arXiv:1707.04928*.

Delattre, S., N. Fournier, and M. Hoffmann (2016). Hawkes processes on large networks. *The Annals of Applied Probability 26*(1), 216–261.

Erdős, P. and A. Rényi (1959). On random graphs I. *Publicationes Mathematicae 6*, 290–297.

Erdős, P. and A. Rényi (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci 5*(1), 17–60.

Fang, G., O. G. Ward, and T. Zheng (2023). Online estimation and community detection of network point processes for event streams. *Statistics and Computing 34*(1), 35.

Fang, G., G. Xu, H. Xu, X. Zhu, and Y. Guan (2023). Group network Hawkes process. *Journal of the American Statistical Association 0*(0), 1–17.

Fox, E. W., M. B. Short, F. P. Schoenberg, K. D. Coronges, and A. L. Bertozzi (2016). Modeling e-mail networks and inferring leadership using self-exciting point processes. *Journal of the American Statistical Association 111*(514), 564–584.

Goh, K.-I. and A.-L. Barabási (2008). Burstiness and memory in complex systems. *Europhysics Letters 81*(4), 48002.

Hannan, E. J. and B. G. Quinn (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B 41*(2), 190–195.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes.

Biometrika 58(1), 83–90.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. *Social Networks 5*(2), 109–137.

Holland, P. W. and S. Leinhardt (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association 76*(373), 33–50.

Huang, Z., H. Soliman, S. Paul, and K. S. Xu (2022). A mutually exciting latent space Hawkes process model for continuous-time networks. In *Uncertainty in Artificial Intelligence*, pp. 863–873. PMLR.

Junuthula, R., M. Haghdan, K. S. Xu, and V. Devabhaktuni (2019). The block point process model for continuous-time event-based dynamic networks. In *The World Wide Web Conference*, pp. 829–839.

Karrer, B. and M. E. Newman (2011). Stochastic blockmodels and community structure in networks. *Physical Review E 83*(1), 016107.

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

Matias, C. and V. Miele (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society. Series B 79*(4), 1119–1141.

Matias, C., T. Rebafka, and F. Villers (2018). A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika 105*(3), 665–680.

Miscouridou, X., F. Caron, and Y. W. Teh (2018). Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data. *Advances in Neural Information Processing Systems 31*, 2349–2358.

Nickel, M. and M. Le (2020). Learning multivariate Hawkes processes at scale. *arXiv preprint arXiv:2002.12501*.

Noroozi, M., R. Rimal, and M. Pensky (2021). Estimation and clustering in popularity adjusted block model. *Journal of the Royal Statistical Society: Series B 83*(2), 293–317.

Ogata, Y. et al. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics 30*(1), 243–261.

Passino, F. S. and N. A. Heard (2023). Mutually exciting point process graphs for modeling dynamic networks. *Journal of Computational and Graphical Statistics 32*(1), 116–130.

Puri, M. L. and P. D. Tuan (1986). Maximum likelihood estimation for stationary point processes. *Proceedings of the National Academy of Sciences 83*(3), 541–545.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, 53–65.

Sengupta, S. and Y. Chen (2018). A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society: Series B 80*(2), 365–386.

Soliman, H., L. Zhao, Z. Huang, S. Paul, and K. S. Xu (2022). The multivariate community Hawkes model for dependent relational events in continuous-time networks. In *International Conference on Machine Learning*, pp. 20329–20346. PMLR.

Stehlé, J., N. Voirin, A. Barrat, C. Cattuto, L. Isella, J. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems (2011). High-resolution measurements of face-to-face contact patterns in a primary school. *PLOS ONE 6*(8), e23176.

Wu, J. (2019). *Point process models for heterogeneous event time data*. Ph. D. thesis, Columbia University.

Wu, J., O. G. Ward, J. Curley, and T. Zheng (2022). Markov-modulated Hawkes processes for modeling sporadic and bursty event occurrences in social interactions. *The Annals of Applied Statistics 16*(2), 1171–1190.

Yang, S. and H. Koeppl (2020). The Hawkes edge partition model for continuous-time event-based temporal Networks. In *Conference on Uncertainty in Artificial Intelligence*, pp. 460–469. PMLR.

Yang, T., Y. Chi, S. Zhu, Y. Gong, and R. Jin (2011). Detecting communities and their evolutions in dynamic social networks—a Bayesian approach. *Machine learning 82*(2), 157–189.

Zhang, J., B. Cai, X. Zhu, H. Wang, G. Xu, and Y. Guan (2023). Learning human activity patterns using clustered point processes with active and inactive states. *Journal of Business & Economic Statistics 41*(2), 388–398.

Zhang, X., C. Moore, and M. E. Newman (2017). Random graph models for dynamic networks. *The European Physical Journal B 90*(10), 1–14.

Zipkin, J. R., F. P. Schoenberg, K. Coronges, and A. L. Bertozzi (2016). Point-process models of social network interactions: Parameter estimation and missing data recovery. *European Journal of Applied Mathematics 27*(3), 502–529.

First and second authors affiliation: Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China

E-mail: shihaosheng@ruc.edu.cn; wenlin.dai@ruc.edu.cn