Supplement to "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

This file serves as a supplement to the main paper. Section S1 provides the detailed derivation of the EM algorithm. Section S2 describes the validity of SSC initialization. Section S3 includes the proof of theoretical results. Section S4 illustrates extensive simulation studies to assess the validity of the theoretical results.

S1 Derivation of the EM algorithm

As is discussed in Section 3.2, we propose to perform an EM algorithm to obtain the MLE of this model. The detailed procedure of the EM algorithm is presented in algorithm 1. Based on the complete likelihood (3.3), we provide detailed derivation of the EM algorithm in this section.

First of all, some conditional probabilities are denoted below to help

take expectation,

$$
\begin{aligned}
p_{ij}^{s,s;[k]} &= \frac{\lambda_{ic_j}^{[k]}\lambda_{jc_i}^{[k]}}{\lambda_{ic_j}^{[k]}\lambda_{jc_i}^{[k]} + \sum_{u=1}^{s-1}\alpha_{c_ic_j}^{[k]}e^{-\beta_{c_ic_j}^{[k]}(t_{ij}^{(s)}-t_{ij}^{(u)})}}, \\
p_{ij}^{s,u;[k]} &= \frac{\alpha_{c_ic_j}^{[k]}e^{-\beta_{c_ic_j}^{[k]}(t_{ij}^{(s)}-t_{ij}^{(u)})}}{\lambda_{ic_j}^{[k]}\lambda_{jc_i}^{[k]} + \sum_{v=1}^{s-1}\alpha_{c_ic_j}^{[k]}e^{-\beta_{c_ic_j}^{[k]}(t_{ij}^{(s)}-t_{ij}^{(v)})}}.
\end{aligned}
\tag{S1.1}
$$

Here, the superscript $^{[k]}$ indicates the kth iteration. Then with $\mathrm{E}^{[k]}(d_{ii}) = \lambda_{ic_i}^{[k]^2}T$, $\mathrm{E}^{[k]}(d_{ij}) = \sum_{s=1}^{n_{ij}} p_{ij}^{s,s;[k]}$ and $\mathrm{E}^{[k]}(1_{\mathrm{trig}_{ij}^{(s)}=u<s}) = p_{ij}^{s,u;[k]}$, the expectation of the complete likelihood becomes

$$
\begin{aligned}
Q &= \mathrm{E}^{[k]}(\ell_c) \\
&= \sum_{i=1}^{N}\sum_{j<i}\left[-\lambda_{ic_j}\lambda_{jc_i}T + \sum_{s=1}^{n_{ij}} p_{ij}^{s,s;[k]}\log\lambda_{ic_j} + \sum_{s=1}^{n_{ij}} p_{ij}^{s,s;[k]}\log\lambda_{jc_i}\right] \\
&+ \sum_{i=1}^{N}\Big[-\frac{1}{2}\lambda_{ic_i}^2 T + \lambda_{ic_i}^{[k]^2}T\log\lambda_{ic_i}\Big] \\
&+ \sum_{i=1}^{N}\sum_{j<i}\left[\sum_{s=1}^{n_{ij}}\left(-\frac{\alpha_{c_ic_j}}{\beta_{c_ic_j}}\right) + \sum_{s=1}^{n_{ij}}\sum_{u=1}^{s-1}\Big\{\log\alpha_{c_ic_j} - \beta_{c_ic_j}(t_{ij}^{(s)} - t_{ij}^{(u)})\Big\}\, p_{ij}^{s,u;[k]}\right].
\end{aligned}
\tag{S1.2}
$$

where we let the summation term in the third bracket be 0 when $n_{ij} = 0$ for convenience of notation.

$\alpha_{c_ic_j}$ and $\beta_{c_ic_j}$ can be solved directly by differentiating (S1.2) and setting

it to zero. For λ_{ib}, since

$$\frac{\partial Q}{\partial \lambda_{ib}} = \begin{cases} \sum_{j<i, c_j=b} \left[-\lambda_{jc_i} T + \sum_{s=1}^{n_{ij}} \frac{p_{ij}^{s,s;[k]}}{\lambda_{ib}} \right] + \sum_{j>i, c_j=b} \left[-\lambda_{jc_i} T + \sum_{s=1}^{n_{ji}} \frac{p_{ji}^{s,s;[k]}}{\lambda_{ib}} \right], & \\ & (c_i \neq b) \\ \sum_{j<i, c_j=b} \left[-\lambda_{jc_i} T + \sum_{s=1}^{n_{ij}} \frac{p_{ij}^{s,s;[k]}}{\lambda_{ib}} \right] + \sum_{j>i, c_j=b} \left[-\lambda_{jc_i} T + \sum_{s=1}^{n_{ji}} \frac{p_{ji}^{s,s;[k]}}{\lambda_{ib}} \right], & \\ -\lambda_{ib} T + \lambda_{ic_i}^{[k]^2} T & (c_i = b) \end{cases}$$

Now denote

$$M_{ib}^{[k]} = \begin{cases} \sum_{j<i, c_j=b} \sum_{s=1}^{n_{ij}} p_{ij}^{s,s;[k]} + \sum_{j>i, c_j=b} \sum_{s=1}^{n_{ji}} p_{ji}^{s,s;[k]} & c_i \neq b \\ \sum_{j<i, c_j=b} \sum_{s=1}^{n_{ij}} p_{ij}^{s,s;[k]} + \sum_{j>i, c_j=b} \sum_{s=1}^{n_{ji}} p_{ji}^{s,s;[k]} + \lambda_{ic_i}^{[k]^2} T & c_i = b \end{cases},$$

$$O_{ab}^{[k]} = \sum_{c_i=a} M_{ib}^{[k]},$$

then $\partial Q / \partial \lambda_{ib} = 0$ gives

$$\Lambda_{c_i b} T = \frac{M_{ib}^{[k]}}{\lambda_{ib}}.$$

where Λ_{ab} is defined as $\Lambda_{ab} = \sum_{c_i=a} \lambda_{ib}$.

Take summation over i with $c_i = a$, we get $\Lambda_{ab}^2 T = O_{ab}^{[k]}$, which directly leads to

$$\lambda_{ib} = \frac{M_{ib}^{[k]}}{\sqrt{O_{c_i b}^{[k]} T}}.$$

Finally if we denote for simplicity ($\sum_{s=1}^{0} t_s \triangleq 0$)

$$E_{ij}^{[k]} = \sum_{s=1}^{n_{ij}} \sum_{u=1}^{s-1} p_{ij}^{s,u;[k]}, \qquad F_{ij}^{[k]} = \sum_{s=1}^{n_{ji}} \sum_{u=1}^{s-1} p_{ji}^{s,u;[k]} (t_{ji}^{(s)} - t_{ji}^{(u)}),$$

we can obtain the updates for the $(i+1)$th iteration: $(a \geq b)$

$$\begin{cases} \lambda_{ib}^{[k+1]} = \dfrac{M_{ib}^{[k]}}{\sqrt{O_{c_i b}^{[k]} T}}, \\ \alpha_{ab}^{[k+1]} = \dfrac{\sum\limits_{c_i=a} \sum\limits_{c_j=b, j<i} E_{ij}^{[k]} + \sum\limits_{c_i=b} \sum\limits_{c_j=a, j<i} E_{ij}^{[k]}}{\sum\limits_{c_i=a} \sum\limits_{c_j=b, j<i} n_{ij} + \sum\limits_{c_i=b} \sum\limits_{c_j=a, j<i} n_{ij}} \beta_{ab}^{[k+1]}, \\ \beta_{ab}^{[k+1]} = \dfrac{\sum\limits_{c_i=a} \sum\limits_{c_j=b, j<i} E_{ij}^{[k]} + \sum\limits_{c_i=b} \sum\limits_{c_j=a, j<i} E_{ij}^{[k]}}{\sum\limits_{c_i=a} \sum\limits_{c_j=b, j<i} F_{ij}^{[k]} + \sum\limits_{c_i=b} \sum\limits_{c_j=a, j<i} F_{ij}^{[k]}}. \end{cases} \tag{S1.3}$$

S2 Validity of SSC initialization

As the complete likelihood function given in (3.1) is non-convex, it is important to find a good initializer for the community membership vector $\boldsymbol{c}^0$. In Arastuie et al. (2020), the number of interactions n_{ij} on each node pair (i, j) is counted to form the accumulated matrix $\mathbf{W}$. Then, a simple spectral clustering is applied to this accumulated matrix. In our case, this approach is not suitable because of the degree variation within communities. However, we find that the limit of the accumulated network in our

Algorithm 1: An EM algorithm for the CHHIP model

Input: Dynamic network data $\mathbf{X}$, community number K

Output: community membership $\boldsymbol{c}$, parameter estimates $(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$

Initialize $\boldsymbol{\alpha}^{[0]}$, $\boldsymbol{\beta}^{[0]}$, $\boldsymbol{\lambda}^{[0]}$, $\boldsymbol{c}^{[0]}$ and set $k = 0$.
while *Not convergence* **do**
  for *nodes* $i > j$ **do**
    for *each timestamp* t_{ij}^s *in node pair* (i, j) **do**
      Calculate $p_{ij}^{s,s;[k]}$ (see (S1.1) in Section S1 of the Supplement)
      for *each timestamp* $t_{ij}^u < t_{ij}^s$ *in node pair* (i, j) **do**
        Calculate $p_{ij}^{s,u;[k]}$ (see (S1.1) in Section S1 of the Supplement)
      end
    end
  end
  for *node* i **do**
    for *block* b **do**
      Use (S1.3) to calculate $\widehat{\lambda}_{ib}^{[k+1]}$
    end
  end
  for *blocks* (a, b) **do**
    Use (S1.3) to calculate $\widehat{\alpha}_{ab}^{[k+1]}$ and $\widehat{\beta}_{ab}^{[k+1]}$
  end
  for *node* i **do**
    for *block* b **do**
      Set $\boldsymbol{d}^{(b)} \leftarrow \boldsymbol{c}^{[k]}$ with the expection of $\boldsymbol{d}_i^{(b)} = b$
      Calculate log-likelihood $\ell_c(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{d}^{(b)})$
    end
    $\boldsymbol{c} \leftarrow \text{argmax}_{\boldsymbol{d}^{(b)}} \ell_c(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{d}^{(b)})$
  end
  $k \leftarrow k + 1$
end

case is also a weighted version of the PABM model. This motivates us to take advantage of the structure and initialize our algorithm with subspace sparse clustering as outlined in algorithm 2.

Algorithm 2: A SSC community initialization algorithm

Input: Dynamic network data $\mathbf{X}$, community number K
Output: block membership vector $\widehat{\boldsymbol{c}}$

for *each node pair* (i, j) **do**
  Calculate n_{ij}, the number of interactions between i and j, to form $\mathbf{W}$
end
for *each node* i **do**
  Use Orthogonal Matching Pursuit (OMP) to solve $\boldsymbol{S}_i$ from

$$\boldsymbol{S}_i \leftarrow \underset{\boldsymbol{S}}{\operatorname{argmin}} \left\{ \|\boldsymbol{W}_i - \mathbf{W}\boldsymbol{S}\|_2^2 \quad \text{s.t.} \quad \|W_i\|_0 \le K,\ W_{ii} = 0 \right\},$$

  where $\boldsymbol{W}_i$ is the ith column of $\mathbf{W}$ (refer to Noroozi et al. (2021) for more details)
end
$\mathbf{S} \leftarrow \mathbf{S} + \mathbf{S}^\mathsf{T}$
$\widehat{\boldsymbol{c}} \leftarrow$ (normalized) spectral clustering for the weighted similarity matrix $\mathbf{S}$ (refer to Noroozi et al. (2021) for more details)

Lemma 2. *Define $n_{ij}(T)$ as the number of events occurring on node pair (i, j) up to T. Then with Assumption 1 in the section 4.1, we have*

$$\frac{1}{T} n_{ij}(T) \mid (c_i = a, c_j = b) \sim \mathcal{N}\left(\frac{\lambda_{ic_j}\lambda_{jc_i}}{1 - \frac{\alpha_{c_i c_j}}{\beta_{c_i c_j}}}, \frac{\lambda_{ic_j}\lambda_{jc_i}}{\left(1 - \frac{\alpha_{c_i c_j}}{\beta_{c_i c_j}}\right)^3 T} \right)$$

as $T \to \infty$.

The proof is given in Section S3.2 of the supplement. With Lemma 2, the expectation of the accumulated network used in algorithm 2 can be considered

as a realization of the modified PABM when T is large. Recall that $\mathrm{E}(\mathbf{A}_{ij}) = \lambda_{ic_j}\lambda_{jc_i}$ under the PABM. Given that $\widetilde{\lambda}_{ib} = (T/(1-\alpha_{c_ib}/\beta_{c_ib}))^{1/2}\lambda_{ib}$, we have $\mathrm{E}\left(n_{ij}(T)\right) \approx \widetilde{\lambda}_{ic_j}\widetilde{\lambda}_{jc_i}$. Moreover, the identifiability condition is naturally satisfied with $\widetilde{\Lambda}_{ab} = (T/(1-\alpha_{ab}/\beta_{ab}))^{1/2}\Lambda_{ab} = (T/(1-\alpha_{ba}/\beta_{ba}))^{1/2}\Lambda_{ba} = \widetilde{\Lambda}_{ba}$.

Noroozi et al. (2021) provides a sparse subspace clustering algorithm for community detection under the PABM model. The algorithm stems from the observation that this model leads to a low-rank adjacency matrix, with all of its columns lying in the union of K subspaces, each of the dimension K. Therefore, with every fixed K they formulate the problem as

$$\widehat{\boldsymbol{c}} \in \underset{\boldsymbol{c}\in\mathbb{C}_{N,K}}{\operatorname{argmin}} \left\{ \sum_{k,l=1}^{K} \left\| W^{(k,l)}(\boldsymbol{c}) - \Pi\left(W^{(k,l)}(\boldsymbol{c})\right) \right\|_F^2 \right\} \tag{S2.4}$$

where $\mathbb{C}_{N,K} = \{1, 2, \cdots, K^0\}^N$, $W^{(k,l)}(\boldsymbol{c})$ is a submatrix of the permuted $\mathbf{W}$ such that its rows correspond to the kth community and its columns correspond to the lth community in $\boldsymbol{c}$; $\Pi(\cdot)$ is a rank-one projection of the matrix. We extend their clustering error results to our weighted case.

Theorem 2. *Define $\widehat{\boldsymbol{c}}$ as the solution of Equation S2.4, $\mathbf{P}^0 \in \mathbb{R}^{N\times N}$ as $P^0_{ij} = \mathbb{E}[n_{ij}]/T$, and the membership class with the error rate being at least*

ρ_n:

$$\Upsilon\left(\boldsymbol{c}^0, \rho_N\right) = \left\{\boldsymbol{c} \in \mathbb{C}_{N,K} : (2N)^{-1} \min_{\sigma(\cdot)} \left\|\boldsymbol{c} - \sigma(\boldsymbol{c}^0)\right\|_1 \geq \rho_N\right\}$$

Assume $\max_{\substack{i,j=1,\cdots,N \\ s,r=1,\cdots,K}} \frac{\lambda_{ir}^0 \lambda_{sj}^0}{\left(1 - \alpha_{sr}^0/\beta_{sr}^0\right)^3} \leq M < \infty$. *If Assumptions 1 and 4 in Section 4 hold, and for some* $\alpha_N \in (0, 1/2)$ *and* $\rho_N \in (0, 1)$, *we have*

$$\left\|\mathbf{P}^0\right\|_F^2 - (1 + \alpha_N) \max_{\boldsymbol{c} \in \Upsilon(\boldsymbol{c}^0, \rho_N)} \sum_{k,l=1}^{K} \left\|\mathbf{P}^{0(k,l)}(\boldsymbol{c})\right\|_{op}^2 \geq \frac{H_1}{\alpha_N T}\left(NK + K^2 \ln N\right) + \frac{H_2 N}{\alpha_N}$$

where H_1 *and* H_2 *are fixed constants determined by* M, α_N *and* c_0.

Then with a known $K = K^0$, *the misclassification rate* $(2N)^{-1} \min_{\sigma(\cdot)} \|\widehat{\boldsymbol{c}} - \sigma(\boldsymbol{c}^0)\|_1$ *is at most* ρ_N *with probability at least* $1 - 2e^{-c_0 NT}$.

The proof is provided in Section S3.3 of the Supplement. In this theorem, we require that the maximal number of events per unit time does not increase with the number of nodes, and this condition is used for the concentration inequality. As explained by Noroozi et al. (2021), the assumption of the error rate bound can be regarded as a condition for the difference between intragroup and intergroup connective probabilities. In spite of the perfect theoretical results, the solution is hard to obtain because this optimization problem is NP-hard, and they use SSC as a relaxation. Following Noroozi et al. (2021), we present our result of correctness at population level using

SSC in Theorem 3. The corresponding proof is given in S3.4.

Theorem 3. *Assume the number of communities $K = K^0$ is known. Then with Assumptions 1 and 4, the SSC algorithm recovers communities correctly up to a permutation on the noiseless accumulated matrix $\mathbf{P}^0$.*

S3 Proof of Theoretical Results

S3.1 Proof of Theorem 1

Proof. Denote $\boldsymbol{\theta}^0 = (\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0, \boldsymbol{\lambda}^0)$ as the true value of parameters and $\boldsymbol{c}^0$ as the true labels.

First, the log-likelihood function in (3.2) can be rewritten as

$$\ell(\boldsymbol{\theta}, \boldsymbol{c}) = \sum_{i=1}^{N} \sum_{j<i} \log \left(- \int_0^T \lambda_{i,j}^*(\boldsymbol{\theta}, \boldsymbol{c}, t, w) \mathrm{d}t + \int_0^T \log \lambda_{i,j}^*(\boldsymbol{\theta}, \boldsymbol{c}, t, w) \mathrm{d}N_{ij}(t) \right)$$

Our proof approximates $\lambda_{i,j}^*(\boldsymbol{\theta}, \boldsymbol{c}, t, w)$ with $\lambda_{i,j}^{**}(\boldsymbol{\theta}, \boldsymbol{c}, t, w)$ in this likelihood function, and prove the result with regard to the approximated log-likelihood. This approximation was first used in Ogata et al. (1978), and is sufficiently close to the original log-likelihood when assumption 3 is satisfied.

From assumption 1, $X_{i,j}(\boldsymbol{\theta}, \boldsymbol{c})$ is stationary, ergodicity with second–

order moment, and by lemma 2 of Ogata et al. (1978) we have

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \min_{\boldsymbol{c}'\in\mathbb{C}}\inf_{\boldsymbol{\theta}'\in\Theta}\lambda_{i,j}^{**}(\boldsymbol{\theta}',\boldsymbol{c},t,\omega)dt=\mathrm{E}\left\{\min_{\boldsymbol{c}'\in\mathbb{C}}\inf_{\boldsymbol{\theta}'\in\Theta}\lambda_{i,j}^{**}(\boldsymbol{\theta}',0,\omega)\right\} \quad \text{(S3.5)}$$

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \log\frac{\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,t,\omega)}{\max_{\boldsymbol{c}'\in\mathbb{C}}\sup_{\boldsymbol{\theta}'\in\Theta}\lambda_{i,j}^{**}(\boldsymbol{\theta}',\boldsymbol{c}',t,\omega)}dN(t)=\mathrm{E}\left\{\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,0,\omega)\log\frac{\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,0,\omega)}{\max_{\boldsymbol{c}'\in\mathbb{C}}\sup_{\boldsymbol{\theta}'\in\Theta}\lambda_{i,j}^{**}(\boldsymbol{\theta}',\boldsymbol{c}',0,\omega)}\right\} \quad \text{(S3.6)}$$

It is easy to verify that $|\sum_i\sum_j\lambda_{i,j}^{**}(\boldsymbol{\theta},\boldsymbol{c},0,w)|$ and $|\sum_i\sum_j\log\lambda_{i,j}^{**}(\boldsymbol{\theta},\boldsymbol{c},0,w)|$ can be dominated by $N(N+1)/2M_0(w)$ and $N(N+1)/2\max\{|\log M_0(w)|,|\log m^2|\}$. Use the dominated convergence theorem:

$$\begin{aligned}&\lim_{U\to\{\boldsymbol{\theta}\}}\mathrm{E}\left[\min_{\boldsymbol{c}\in\mathbb{C}}\inf_{\boldsymbol{\theta}'\in U}\left\{\sum_i\sum_j\lambda_{i,j}^{**}(\boldsymbol{\theta}',\boldsymbol{c},0,\omega)\right\}\right]\\ =&\mathrm{E}\left[\lim_{U\to\{\boldsymbol{\theta}\}}\min_{\boldsymbol{c}\in\mathbb{C}}\inf_{\boldsymbol{\theta}'\in U}\left\{\sum_i\sum_j\lambda_{i,j}^{**}(\boldsymbol{\theta}',\boldsymbol{c},0,\omega)\right\}\right]=\min_{\boldsymbol{c}\in\mathbb{C}}\sum_i\sum_j\mathrm{E}\left[\lambda_{i,j}^{**}(\boldsymbol{\theta},\boldsymbol{c},0,\omega)\right]\end{aligned} \quad \text{(S3.7)}$$

$$\begin{aligned}&\lim_{U\to\{\boldsymbol{\theta}\}}\mathrm{E}\left[\min_{\boldsymbol{c}\in\mathbb{C}}\inf_{\theta'\in U}\left\{\sum_i\sum_j\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,0,\omega)\log\frac{\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,0,\omega)}{\lambda_{i,j}^{*}(\boldsymbol{\theta}',\boldsymbol{c},0,\omega)}\right\}\right]\\ =&\min_{\boldsymbol{c}\in\mathbb{C}}\sum_i\sum_j\mathrm{E}\left[\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,0,\omega)\log\frac{\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,0,\omega)}{\lambda_{i,j}^{**}(\boldsymbol{\theta},\boldsymbol{c},0,\omega)}\right]\end{aligned} \quad \text{(S3.8)}$$

If we define the likelihood ratio on $[0,1]$ as

$$R_{ij}\left(\boldsymbol{\theta}^0,\boldsymbol{c}^0;\boldsymbol{\theta},\boldsymbol{c}\right)=\int_0^1\left\{\lambda_{i,j}^{**}(\boldsymbol{\theta},\boldsymbol{c},t,\omega)-\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,t,\omega)\right\}dt+\int_0^1\log\frac{\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,t,\omega)}{\lambda_{i,j}^{**}(\boldsymbol{\theta},\boldsymbol{c},t,\omega)}dN(t),$$

By lemma 3 of Ogata et al. (1978), for any $\boldsymbol{c}$, i and j we have

$$\mathrm{E}\left[R_{ij}\left(\boldsymbol{\theta}^0,\boldsymbol{c}^0;\boldsymbol{\theta},\boldsymbol{c}\right)\right]=\mathrm{E}\left[\lambda_{i,j}^{**}(\boldsymbol{\theta},\boldsymbol{c},0,\omega)-\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,0,\omega)+\log\frac{\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,0,\omega)}{\lambda_{i,j}^{**}(\boldsymbol{\theta},\boldsymbol{c},0,\omega)}\right]\geq 0 \tag{S3.9}$$

Besides, it is easy to see that $\lambda_{i,j}^{**}(\boldsymbol{\theta},\boldsymbol{c},0,w)\overset{\text{a.s.}}{=}\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,0,w)$ holds for any nodes i and j if and only if $\boldsymbol{\theta}=\boldsymbol{\theta}^0$ and $\boldsymbol{c}=\sigma(\boldsymbol{c}^0)$ up to a permutation $\sigma(\cdot)$. For example, for the "only if" part, we first know that $\lambda_{i,j}^{**}(\boldsymbol{\theta},\boldsymbol{c},0,w)\overset{\text{a.s.}}{=}\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,0,w)$ if and only if $\lambda_{ic_j}\lambda_{jc_i}=\lambda^0_{ic_j^0}\lambda^0_{jc_i^0}$, $\alpha_{c_ic_j}=\alpha_{ij}=\alpha^0_{ij}=\alpha^0_{c_i^0c_j^0}$ and $\beta_{c_ic_j}=\beta_{ij}=\beta^0_{ij}=\beta^0_{c_i^0c_j^0}$. Therefore we have $\boldsymbol{\mu}(\boldsymbol{c})=\boldsymbol{\mu}^0(\boldsymbol{c}^0)$ if we denote $\boldsymbol{\mu}(\boldsymbol{c})$ as $\mu_{ij}(\boldsymbol{c})=\lambda_{ic_j}\lambda_{jc_i}$. Note that $\boldsymbol{\mu}(\boldsymbol{c})$ also takes the form of PABM, so the proof of lemma 1 in Estimation and clustering in popularity adjusted block model tells us that assumption 4 guarantees the recovery of $\boldsymbol{c}$ up to permutation. From that point we can obtain $\boldsymbol{\theta}=\boldsymbol{\theta}^0$ up to permutation.

Now suppose U_0 to be any open neighbourhood of $\boldsymbol{\theta}^0$. For notational convenience, we use $\sigma(\boldsymbol{\theta}^0)$ and $\sigma(U_0)$ to represent the image of $\boldsymbol{\theta}^0$ and U_0 under any fixed permutation mapping $\sigma(\cdot)$. Then (S3.9) tells us that

$\sum_i \sum_j \mathrm{E}\left[R_{ij}(\boldsymbol{\theta}_0, \boldsymbol{c}^0; \boldsymbol{\theta}, \boldsymbol{c})\right] = 0$ if and only if $\boldsymbol{\theta} = \sigma(\boldsymbol{\theta}^0)$ and $\boldsymbol{c} = \sigma(\boldsymbol{c}^0)$. In other words, there exists $\varepsilon > 0$, for any $\boldsymbol{\theta} \in \Theta \backslash \sigma(U_0)$, we have

$$\min_{\boldsymbol{c} \in \mathbb{C}} \sum_{i=1}^{N} \sum_{j<i} \mathrm{E}\left[R_{ij}\left(\boldsymbol{\theta}^0, \boldsymbol{c}^0; \sigma(\boldsymbol{\theta}), \sigma(\boldsymbol{c})\right)\right] \geq 3\varepsilon \tag{S3.10}$$

and for $\boldsymbol{\theta} \in \sigma(U_0)$ and $\boldsymbol{c} \neq \boldsymbol{\sigma}(c^0)$,

$$\sum_{i=1}^{N} \sum_{j<i} \mathrm{E}\left[R_{ij}\left(\boldsymbol{\theta}^0, \boldsymbol{c}^0; \sigma(\boldsymbol{\theta}), \sigma(\boldsymbol{c})\right)\right] \geq 3\varepsilon \tag{S3.11}$$

Without loss of generality we suppose $\sigma(\boldsymbol{c}) = \boldsymbol{c}$ first. From the compactness of parameter space, we can select a finite number of $\boldsymbol{\theta}_s$ to construct a covering of $\Theta \backslash U_0$: $\{U_s := U_{\boldsymbol{\theta}_s, \varepsilon}\}$, where $U_{\boldsymbol{\theta}_s, \varepsilon}$ is a sufficiently small neighbourhood of $\boldsymbol{\theta}_s$. Using (S3.5) – (S3.8) and (S3.10) ,

$$\begin{aligned}
&\frac{1}{T}\ell\left(\boldsymbol{\theta}^0, \boldsymbol{c}^0\right) - \max_{\boldsymbol{c} \in \mathbb{C}} \sup_{\boldsymbol{\theta} \in U_s} \frac{1}{T}\ell(\boldsymbol{\theta}, \boldsymbol{c}) \\
\geq & -\frac{1}{T}\sum_{i=1}^{N}\sum_{j<i}\int_0^T \lambda_{i,j}^{**}(\boldsymbol{\theta}^0, \boldsymbol{c}^0, t, \omega)dt + \frac{1}{T}\sum_{i=1}^{N}\sum_{j<i}\int_0^T \log \lambda_{i,j}^{**}(\boldsymbol{\theta}^0, \boldsymbol{c}^0, t, \omega)dN_{ij}(t) \\
& - \frac{1}{T}\min_{\boldsymbol{c} \in \mathbb{C}} \inf_{\boldsymbol{\theta} \in U_s} \left(\sum_{i=1}^{N}\sum_{j<i}\int_0^T \log \lambda_{i,j}^{**}(\boldsymbol{\theta}, \boldsymbol{c}, t, \omega)dN_{ij}(t) - \sum_{i=1}^{N}\sum_{j<i}\int_0^T \lambda_{i,j}^{**}(\boldsymbol{\theta}, \boldsymbol{c}, t, \omega)dt\right) \\
\geq & \frac{1}{T}\int_0^T \min_{\boldsymbol{c} \in \mathbb{C}} \inf_{\boldsymbol{\theta} \in U_s} \sum_{i=1}^{N}\sum_{j<i}\left\{\lambda_{i,j}^{**}(\boldsymbol{\theta}, \boldsymbol{c}, t, \omega) - \lambda_{i,j}^{**}(\boldsymbol{\theta}^0, \boldsymbol{c}^0, t, \omega)\right\} dt \\
& + \frac{1}{T}\int_0^T \min_{\boldsymbol{c} \in \mathbb{C}} \inf_{\boldsymbol{\theta} \in U_s} \sum_{i=1}^{n}\sum_{j<i} \log \frac{\lambda_{i,j}^{**}(\boldsymbol{\theta}^0, \boldsymbol{c}^0, t, \omega)}{\lambda_{ij}^{**}(\boldsymbol{\theta}, \boldsymbol{c}, t, \omega)} dN(t)
\end{aligned}$$

$$
\begin{aligned}
&\geq \mathrm{E}\left\{\min_{\boldsymbol{c}\in\mathbb{C}}\inf_{\boldsymbol{\theta}\in U_s}\sum_{i=1}^{N}\sum_{j<i}\left(\lambda_{i,j}^{**}(\boldsymbol{\theta},\boldsymbol{c},0)-\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,0)\right)\right\} \\
&\quad+\mathrm{E}\left\{\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,0)\min_{\boldsymbol{c}\in\mathbb{C}}\inf_{\boldsymbol{\theta}\in U_s}\sum_{i=1}^{N}\sum_{j<i}\log\frac{\lambda_{i,j}^{**}(\boldsymbol{\theta}^0,\boldsymbol{c}^0,0)}{\lambda_{i,j}^{**}(\boldsymbol{\theta},\boldsymbol{c},0)}\right\}-\epsilon \\
&\geq \min_{\boldsymbol{c}\in\mathbb{C}}\sum_{i=1}^{N}\sum_{j<i}\mathrm{E}[R_{ij}(\boldsymbol{\theta}^0,\boldsymbol{c}^0;\boldsymbol{\theta},\boldsymbol{c})]-2\epsilon\geq\epsilon
\end{aligned}
$$

Therefore for any U_0 containing $\boldsymbol{\theta}^0$, there exists $T_1 = T_1(\varepsilon, U_0) > T_0$ for any $T > T_1$,

$$
\sup_{\boldsymbol{\theta}\in U_0}\ell(\boldsymbol{\theta},\boldsymbol{c}^0)\geq\max_{\boldsymbol{c}\in\mathbb{C}}\sup_{\boldsymbol{\theta}\in\Theta\setminus U_0}\ell(\boldsymbol{\theta},\boldsymbol{c})+\varepsilon T. \tag{S3.12}
$$

Likewise for any $\boldsymbol{c}\neq\boldsymbol{c}^0$

$$
\sup_{\boldsymbol{\theta}\in U_0}\ell(\boldsymbol{\theta},\boldsymbol{c}^0)\geq\sup_{\boldsymbol{\theta}\in U_0}\ell(\boldsymbol{\theta},\boldsymbol{c})+\varepsilon T. \tag{S3.13}
$$

To generalize the results to the permuted case, we can substitute U_0 with the union of $\sigma(U_0)$ for all of permutation $\sigma(\cdot)$. Combining (S3.12) and (S3.13), we prove the theorem. □

S3.2 Proof of Lemma 2

Lemma 2. Lemma 2 is a direct application of Theorem 4 in Hawkes and Oakes (1974).

Specifically, for each i and j, parameters there are set as $\gamma(u) = \alpha_{c_i c_j} e^{-\beta_{c_i c_j} u}$ and $\nu = \lambda_{ic_j}\lambda_{jc_i}$.

Then with $\int_0^\infty u\gamma(u)du = \alpha/\beta + \alpha/\beta^2 < \infty$, it follows that

$$\frac{\nu T}{1-m} = \frac{\lambda_{ic_j}\lambda_{jc_i}T}{1-\frac{\alpha_{c_ic_j}}{\beta_{c_ic_j}}} \qquad \text{and} \qquad \frac{\nu T}{(1-m)^3} = \frac{\lambda_{ic_j}\lambda_{jc_i}T}{\left(1-\frac{\alpha_{c_ic_j}}{\beta_{c_ic_j}}\right)^3},$$

and this accomplishes the proof of this lemma. □

S3.3 Proof of Theorem 2

Proof. First, note that $\widehat{\boldsymbol{c}}$ is the solution of the optimization objective, we have

$$\sum_{k,l=1}^{K} \left\|\mathbf{W}^{(k,l)}(\hat{\boldsymbol{c}}) - \Pi_{(1)}\left(\mathbf{W}^{(k,l)}(\hat{\boldsymbol{c}})\right)\right\|_F^2 \leq \sum_{k,l=1}^{K} \left\|\mathbf{W}^{(k,l)}\left(\boldsymbol{c}^0\right) - \Pi_{(1)}\left(\mathbf{W}^{(k,l)}\left(\boldsymbol{c}^0\right)\right)\right\|_F^2 \tag{S3.14}$$

Using the property of the one-rank projection, for any fixed $\boldsymbol{c}$, we have

$$\sum_{k,l=1}^{K} \left\|\mathbf{W}^{(k,l)}(\boldsymbol{c}) - \Pi_{(1)}\left(\mathbf{W}^{(k,l)}(\boldsymbol{c})\right)\right\|_F^2 = \sum_{k,l=1}^{K} \left\{\left\|\mathbf{W}^{(k,l)}(\boldsymbol{c})\right\|_F^2 - \left\|\Pi_{(1)}\left(\mathbf{W}^{(k,l)}(\boldsymbol{c})\right)\right\|_F^2\right\} \tag{S3.15}$$

By (S3.14), (S3.15) and $\sum_{k,l=1}^{K}\left\|\mathbf{W}^{(k,l)}(\boldsymbol{c})\right\|_F^2 = \|\mathbf{W}\|_F^2$, we have

$$\sum_{k,l=1}^{K} \left\|\Pi_{(1)}\left(\mathbf{W}^{(k,l)}(\hat{\boldsymbol{c}})\right)\right\|_F^2 \geq \sum_{k,l=1}^{K} \left\|\Pi_{(1)}\left(\mathbf{W}^{(k,l)}\left(\boldsymbol{c}^0\right)\right)\right\|_F^2 \tag{S3.16}$$

Now we denote $\boldsymbol{\Xi}^{(k,l)}(\boldsymbol{c}) = \mathbf{W}^{(k,l)}(\boldsymbol{c}) - \mathbf{P}^{\mathbf{0}(k,l)}(\boldsymbol{c})$. Note that $\Pi_{(1)}\left(\mathbf{W}^{(k,l)}\left(\boldsymbol{c}^0\right)\right)$ is of rank one. Besides, define $\boldsymbol{V}^{\mathbf{0}(k,l)} \in \mathbb{R}^{N_k}$ as $V_r^{0(k,l)} = \lambda^0_{i_k^r,l}$ where i_k^r is the rth node in community k, and we find that $\mathbf{P}^{\mathbf{0}(k,l)}(\boldsymbol{c}) = \boldsymbol{V}^{0(k,l)}(\boldsymbol{c})\boldsymbol{V}^{0(l,k)^{\mathrm{T}}}(\boldsymbol{c})$ is also of rank one. Therefore

$$\begin{aligned}\left\|\Pi_{(1)}\left(\mathbf{W}^{(k,l)}\left(\boldsymbol{c}^0\right)\right)\right\|_F &= \left\|\Pi_{(1)}\left(\mathbf{W}^{(k,l)}\left(\boldsymbol{c}^0\right)\right)\right\|_{op} \\ &\geq \left\|\mathbf{P}^{\mathbf{0}(k,l)}\left(\boldsymbol{c}^0\right)\right\|_{op} - \left\|\Pi_{(1)}\left(\boldsymbol{\Xi}^{(k,l)}\left(\boldsymbol{c}^0\right)\right)\right\|_{op}. \\ &\geq \left\|\mathbf{P}^{\mathbf{0}(k,l)}\left(\boldsymbol{c}^0\right)\right\|_F - \left\|\boldsymbol{\Xi}^{(k,l)}\left(\boldsymbol{c}^0\right)\right\|_{op}\end{aligned}$$

With the inequality $a \geq b - c \Rightarrow a^2 \geq b^2/(\tau+1) - c^2/\tau$ for any positive a, b, c, τ, the right hand side has a lower bound:

$$\sum_{k,l=1}^{K}\left\|\Pi_{(1)}\left(\mathbf{W}^{(k,l)}\left(\boldsymbol{c}^0\right)\right)\right\|_F^2 \geq \frac{1}{\tau+1}\left\|\mathbf{P}^0\right\|_F^2 - \frac{1}{\tau}\sum_{k,l=1}^{K}\left\|\boldsymbol{\Xi}^{(k,l)}\left(\boldsymbol{c}^0\right)\right\|_{op}^2; \tag{S3.17}$$

And with the fact that

$$\left\|\Pi_{(1)}\left(\mathbf{W}^{(k,l)}(\hat{\boldsymbol{c}})\right)\right\|_F = \left\|\Pi_{(1)}\left(\mathbf{W}^{(k,l)}(\hat{\boldsymbol{c}})\right)\right\|_{op} \leq \left\|\mathbf{P}^{\mathbf{0}(k,l)}(\hat{\boldsymbol{c}})\right\|_{op} + \left\|\boldsymbol{\Xi}^{(k,l)}(\hat{\boldsymbol{c}})\right\|_{op},$$

the left hand side has an upper bound with regards to any positive τ_0:

$$\sum_{k,l=1}^{K}\left\|\Pi_{(1)}\left(\mathbf{W}^{(k,l)}(\hat{\boldsymbol{c}})\right)\right\|_F^2 \leq (1+\tau_0)\sum_{k,l=1}^{K}\left\|\mathbf{P}^{0(k,l)}\left(\hat{\boldsymbol{c}}\right)\right\|_{op}^2 + \left(1+\frac{1}{\tau_0}\right)\sum_{k,l=1}^{K}\left\|\boldsymbol{\Xi}^{(k,l)}\left(\hat{\boldsymbol{c}}\right)\right\|_{op}^2. \tag{S3.18}$$

Now we follow the proof of Lemma 2 and Theorem B.3 in Arastuie et al. (2020) where the (i,j)th entry of $\boldsymbol{\Xi}(\boldsymbol{c}^0)$ can be approximated by the normal distribution $N\left(\lambda^0_{ic_j}\lambda^0_{jc_i}/((1-\alpha^0_{c_ic_j}/\beta^0_{c_ic_j})T), \lambda^0_{ic_j}\lambda^0_{jc_i}/((1-\alpha^0_{c_ic_j}/\beta^0_{c_ic_j})^3T)\right)$ when T is large enough. Thus we take the value of $\sigma_1, \sigma_2, \sigma_*$ in the Theorem 3.1 of Bandeira and Van Handel (2016) with

$$\begin{aligned}
\sigma_1 &= \frac{1}{\sqrt{T}} \max_{c_i=k} \sqrt{\frac{\lambda^0_{il}\Lambda^0_{lk}}{(1-\alpha^0_{kl}/\beta^0_{kl})^3}}, \\
\sigma_2 &= \frac{1}{\sqrt{T}} \max_{c_j=l} \sqrt{\frac{\lambda^0_{jk}\Lambda^0_{kl}}{(1-\alpha^0_{kl}/\beta^0_{kl})^3}}, \\
\sigma_* &= \frac{1}{\sqrt{T}} \max_{c_i=k, c_j=l} \sqrt{\frac{\lambda^0_{il}\lambda^0_{jk}}{(1-\alpha^0_{kl}/\beta^0_{kl})^3}}.
\end{aligned}$$

With the assumption of $\max_{\substack{i,j=1,\cdots,N \\ s,r=1,\cdots,K}} \lambda^0_{ir}\lambda^0_{sj}/(1-\alpha^0_{sr}/\beta^0_{sr})^3 \le M$, we have

$$\sigma_1 \le \frac{1}{\sqrt{T}}\sqrt{\hat{N}_k M},\ \sigma_2 \le \frac{1}{\sqrt{T}}\sqrt{\hat{N}_l M},\ \sigma_* \le \frac{1}{\sqrt{T}}\sqrt{M}.$$

So use the Theorem 3.1 of Bandeira and Van Handel (2016), we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{k,l=1}^{K} \left\|\boldsymbol{\Xi}^{(k,l)}(\boldsymbol{c}^0)\right\|_{\text{op}}^2\right] &\lesssim 4c_1^2 \sum_{k,l=1}^{K}\left(\sigma_1^{(k,l)^2} + \sigma_2^{(k,l)^2}\right) + 2c_2^2 \sum_{k,l=1}^{K} \sigma_*^2 \log(N_k \wedge N_l) \\
&\le C_1 NK/T + C_2 K^2/T \ln N
\end{aligned} \tag{S3.19}$$

Likewise,

$$\mathbb{E}\left[\sum_{k,l=1}^{K}\left\|\boldsymbol{\Xi}^{(k,l)}(\widehat{\boldsymbol{c}})\right\|_{\mathrm{op}}^{2}\right]\lesssim C_1NK/T+C_2K^2/T\ln N \tag{S3.20}$$

Besides, let $g_{k,l}(\boldsymbol{\Xi}^{'})=\left\|\mathrm{diag}(\sqrt{\lambda^0_{i_1l}},\cdots,\sqrt{\lambda^0_{i_{N_k}l}})[\Xi'_{i,j}]_{\widehat{c}_i=k,\widehat{c}_j=l}\,\mathrm{diag}(\sqrt{\lambda^0_{kj_1}},\cdots,\sqrt{\lambda^0_{kj_{N_l}}})\right\|_{\mathrm{op}}$ for each block pair (k,l), we can prove it is sub-Gaussian when $\boldsymbol{\Xi}^{'}$ contains independent Gaussian variables. To see this, we denote $f_{k,l}(\mathrm{vec}(\boldsymbol{\Xi}^{'}))=g_{k,l}(\boldsymbol{\Xi}^{'})$ and prove $f_{k,l}(\cdot)$ is Lipschitz. This is because for any $\boldsymbol{\Xi}^{1\prime},\ \boldsymbol{\Xi}^{2\prime}\in\mathbb{R}^{N_k\times N_l}$,

$$\begin{aligned}
&|f^0_{k,l}(\mathrm{vec}(\boldsymbol{\Xi}^{1^{'}}))-f^0_{k,l}(\mathrm{vec}(\boldsymbol{\Xi}^{2^{'}}))|^2\\
\leq&\left\|\mathrm{diag}(\sqrt{\lambda^0_{i_1l}},\cdots,\sqrt{\lambda^0_{i_{N_k}l}})[\Xi_i^{1^{'}}-\Xi_i^{2^{'}}]_i\mathrm{diag}(\sqrt{\lambda^0_{kj_1}},\cdots,\sqrt{\lambda^0_{kj_{N_l}}})\right\|_{\mathrm{op}}^2\\
\leq&\left\|\mathrm{diag}(\sqrt{\lambda^0_{i_1l}},\cdots,\sqrt{\lambda^0_{i_{N_k}l}})\right\|_{\mathrm{op}}^2\left\|[\Xi_i^{1^{'}}-\Xi_i^{2^{'}}]_i\right\|_{\mathrm{op}}^2\left\|\mathrm{diag}(\sqrt{\lambda^0_{kj_1}},\cdots,\sqrt{\lambda^0_{kj_{N_l}}})\right\|_{\mathrm{op}}^2\\
\leq&\max_{c_i=k,c_j=l}\{\lambda^0_{il}\lambda^0_{kj}\}\left\|[\Xi_i^{1^{'}}-\Xi_i^{2^{'}}]_i\right\|_{\mathrm{F}}^2\\
=&\max_{c_i=k,c_j=l}\{\lambda^0_{il}\lambda^0_{kj}\}\|\mathrm{vec}(\boldsymbol{\Xi}^{1^{'}})-\mathrm{vec}(\boldsymbol{\Xi}^{2^{'}})\|^2
\end{aligned}\tag{S3.21}$$

Define $\boldsymbol{\Xi}^{(k,l)'}(\boldsymbol{c})\in\mathbb{R}^{N_k\times N_l}$ where $\boldsymbol{\Xi}^{(k,l)'}_{ij}(\boldsymbol{c})=\sqrt{(1-\alpha^0_{kl}/\beta^0_{kl})^3T/(\lambda^0_{il}\lambda^0_{jk})}\boldsymbol{\Xi}^{(k,l)}_{ij}(\boldsymbol{c})$. Then use the fact that entries in $\boldsymbol{\Xi}^{(k,l)'}_{ij}(\boldsymbol{c}^0)$ are independent standard Gaussian variables when T is large enough, we know $\|\boldsymbol{\Xi}^{(k,l)}(\boldsymbol{c}^0)\|_{op}=g_{k,l}(\boldsymbol{\Xi}^{(k,l)'}(\boldsymbol{c}^0))/\sqrt{(1-\alpha^0_{kl}/\beta^0_{kl})^3T}$ is a sub-Gaussian variable with a parameter of $\sqrt{\max_{c_i=k,c_j=l}\{\lambda^0_{il}\lambda^0_{kj}\}/(1-\alpha^0_{kl}/\beta^0_{kl})^3T}$.

As $\max_{\substack{i,j=1,\cdots,N \\ s,r=1,\cdots K}}\{\lambda_{ir}^0\lambda_{sj}^0/[(1-\alpha_{sr}^0/\beta_{sr}^0)^3]\} \leq M$, we obtain the following bound for any $v>0$ using the Hoeffding inequality:

$$\mathbb{E}\left[\exp\left(v\left(\left\|\boldsymbol{\Xi}^{(k,l)}(\boldsymbol{c}^0)\right\|_{\text{op}}-\mathbb{E}\left[\left\|\boldsymbol{\Xi}^{(k,l)}(\boldsymbol{c}^0)\right\|_{\text{op}}\right]\right)\right)\right]\leq e^{Mv^2/(2T)}$$

Likewise, for any $\widehat{\boldsymbol{c}}$ we also have

$$\mathbb{E}\left[\exp\left(v\left(\left\|\boldsymbol{\Xi}^{(k,l)}(\widehat{\boldsymbol{c}})\right\|_{\text{op}}-\mathbb{E}\left[\left\|\boldsymbol{\Xi}^{(k,l)}(\widehat{\boldsymbol{c}})\right\|_{\text{op}}\right]\right)\right)\right]\leq e^{Mv^2/(2T)}$$

Now we define two vectors with sub-Gaussian entries $\boldsymbol{\eta}^0,\widehat{\boldsymbol{\eta}}\in\mathbb{R}^{K(K+1)/2}$, where $\boldsymbol{\eta}^0_{(k-1)k/2+l}=\left\|\boldsymbol{\Xi}^{(k,l)}(\boldsymbol{c}^0)\right\|_{\text{op}}-\mathbb{E}\left[\left\|\boldsymbol{\Xi}^{(k,l)}(\boldsymbol{c}^0)\right\|_{\text{op}}\right]$, $\widehat{\boldsymbol{\eta}}_{(k-1)k/2+l}=\left\|\boldsymbol{\Xi}^{(k,l)}(\widehat{\boldsymbol{c}})\right\|_{\text{op}}-\mathbb{E}\left[\left\|\boldsymbol{\Xi}^{(k,l)}(\widehat{\boldsymbol{c}})\right\|_{\text{op}}\right]$. Take $\mathbf{A}=\mathbf{I}_{K(K+1)/2}$ in Theorem 2.1 of Hsu et al. (2012), which is

$$\mathbb{P}\left\{\|\mathbf{A}\tilde{\boldsymbol{\eta}}\|^2\geq\sigma^2\left(\text{Tr}\left(\mathbf{A}^T\mathbf{A}\right)+2\sqrt{\text{Tr}\left(\left(\mathbf{A}^T\mathbf{A}\right)^2\right)x}+2\left\|\mathbf{A}^T\mathbf{A}\right\|_{op}x\right)\right\}\leq\exp(-x)$$

Then for any $x>0$ we have

$$\mathbb{P}\left\{\|\boldsymbol{\eta}^0\|^2\geq 2MK(K+1)/T+6Mx/T)\right\}\leq\exp(-x) \tag{S3.22}$$

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{\eta}}\|^2\geq 2MK(K+1)/T+6Mx/T)\right\}\leq\exp(-x) \tag{S3.23}$$

Combining (S3.19) and (S3.22), we have

$$P\left(\sum_{k,l=1}^{K}\left\|\boldsymbol{\Xi}^{(k,l)}(\boldsymbol{c}^0)\right\|_{\mathrm{op}}^2 \leq C_1NK/T + C_2K^2/T\ln N + 2MK(K+1)/T + 6Mx/T\right) \gtrsim 1-\exp(-x) \tag{S3.24}$$

For any fixed $\widehat{\boldsymbol{c}}$, combining (S3.20) and (S3.23), we have

$$P\left(\sum_{k,l=1}^{K}\left\|\boldsymbol{\Xi}^{(k,l)}(\widehat{\boldsymbol{c}})\right\|_{\mathrm{op}}^2 \leq C_1NK/T + C_2K^2/T\ln N + 2MK(K+1)/T + 6Mx/T\right) \gtrsim 1-\exp(-x) \tag{S3.25}$$

Since $|\Upsilon(\boldsymbol{c}^0,\rho_N)| \leq \sum_{h\geq[\rho_N n]}\binom{N}{h}(K-1)^h \leq K^N$, take the union of $\widehat{\boldsymbol{c}} \in \Upsilon(\boldsymbol{c}^0,\rho_N)$ and let $x \leftarrow N\ln K + t$ in (S3.25), then

$$\mathbb{P}\left\{\max_{\widehat{\boldsymbol{c}}\in\Upsilon(\boldsymbol{c}^0,\rho_N)}\sum_{k,l=1}^{K}\left\|\boldsymbol{\Xi}^{(k,l)}(\widehat{\boldsymbol{c}})\right\|_{op}^2 \leq C_1NK/T + C_2K^2\ln(Ne)/T + 2MK(K+1)/T + C_3/T(N\ln K + t)\right\} \geq 1-\exp(-t) \tag{S3.26}$$

Let $\tau=\tau_0$, $t=c_0NT$, $1+\alpha_N=(1+\tau)^2$. Then $\tau^{-1}=\alpha_N^{-1}\left(1+\sqrt{1+\alpha_N}\right)$ and when $\alpha_N \leq 1$, we have

$$\tau^{-1}(1+\tau)^l \leq \alpha_N^{-1}2^{l/2}(\sqrt{2}+1) \qquad l=0,1,2.$$

Combining (S3.16),(S3.17), (S3.18) ,(S3.24) and (S3.26), with probability

at least $1-2\exp(-c_0NT)$ the following inequality holds:

$$\left\|\mathbf{P}^0\right\|_F^2-(1+\alpha_N)\max_{\widehat{\boldsymbol{c}}\in\Upsilon(\boldsymbol{c}^0,\rho_N)}\sum_{k,l=1}^{K}\left\|\mathbf{P}^{\mathbf{0}(k,l)}(\widehat{\boldsymbol{c}})\right\|_{op}^2\leq\frac{H_1}{\alpha_NT}\left(NK+K^2\ln N\right)+\frac{H_2}{\alpha_N}N, \tag{S3.27}$$

where $H_1\leq\sqrt{2}(1+\sqrt{2})^2\max\left[C_1+C_3,2C_2\right]$ and $H_2\leq C_3c_0$. This inequality is obtained through $N\ln K\leq NK$ and $\ln(Ne)\leq 2\ln N$.

To conclude, if the solution of the optimization problem $\widehat{\boldsymbol{c}}$ is in $\Upsilon\left(\boldsymbol{c}^0,\rho_N\right)$, then (S3.27) holds. However, this result contradicts with the condition. Therefore, $\widehat{\boldsymbol{c}}$ does not belong to $\Upsilon\left(\boldsymbol{c}^0,\rho_N\right)$ and this accomplishes the proof.

□

S3.4 Proof of Theorem 3

As $\mathbf{P}^0$, the adjacency matrix at the population level, can be expressed as a weighted PABM structure by Lemma 2, the proof of this theorem is exactly the same as Theorem 4 of Noroozi et al. (2021).

S4 Simulation experiments

In this section, we simulate networks under several models, and evaluate clustering effects and parameter estimation performance. α and β are parameters related with the triggering mechanism. For simplicity, we let α_1,β_1

be the value for α_{ab}, β_{ab} when $a = b$ and α_2, β_2 be the value when $a \neq b$. Two additional parameters z and w are introduced to control the background intensities of the temporal network. In general, a larger z leads to a globally denser network, while a smaller w enlarges the gap of interaction frequency between intragroup and intergroup dyads. We fix $N = 150$ and $K = 3$ throughout the simulation unless otherwise specified, and assign a node to each block with an equal probability. The detailed data generating procedure, parameter settings and a related analysis are provided below in Section S4.1–S4.3.

S4.1 The simulation procedure

Given the ground membership for each node, a two-step procedure is designed to generate dynamic networks under CHHIP. First, we randomly generate a symmetric, $N \times N$ baseline intensity matrix $\mathbf{B}$. Then we independently simulate a Hawkes process for each node pair using the corresponding entry in $\mathbf{B}$ and excitation parameters α, β. Hawkes processes can be simulated using the *hawkes* package (Zaatour (2014)) in R.

We generate $\mathbf{B}$ using the approach adopted in Noroozi et al. (2021) to obtain fairly diverse intensities. This approach is based on the observation that the matrix can be partitioned into K^2 rank-one blocks. Without

loss of generality, we let the first N_1 rows of $\mathbf{B}$ correspond to nodes from the first community, the next N_2 rows correspond to nodes from the second community, and so on. Denote by $\mathbf{B}^{(a,b)}$ the (a,b)th block pair of $\mathbf{B}$. Besides, $\boldsymbol{V}^{(a,b)} \in \mathbb{R}^{N_a}$ is defined by $V_r^{(a,b)} = \lambda_{i_a^r,b}$, where i_a^r is the rth node in block a. By definition, $\mathbf{B}_{r,s}^{(a,b)} = \lambda_{i_a^r b}\lambda_{i_b^s a} = V_r^{(a,b)}V_s^{(b,a)}$, resulting in $\mathbf{B}^{(a,b)} = \boldsymbol{V}^{(a,b)}\boldsymbol{V}^{(b,a)}$. Therefore, all we need to do is to construct the $N \times K$ matrix

$$\mathbf{V} = \begin{bmatrix} \boldsymbol{V}^{(1,1)} & \boldsymbol{V}^{(1,2)} & \dots & \boldsymbol{V}^{(1,K)} \\ \boldsymbol{V}^{(2,1)} & \boldsymbol{V}^{(2,2)} & \dots & \boldsymbol{V}^{(2,K)} \\ \vdots & \vdots & \dots & \vdots \\ \boldsymbol{V}^{(K,1)} & \boldsymbol{V}^{(K,2)} & \dots & \boldsymbol{V}^{(K,K)} \end{bmatrix}. \tag{S4.28}$$

We first generate diagonal blocks in $\boldsymbol{V}$ by uniformly sampling values from the interval $(y,z), 0 < y < z < 1$. Here we set a fixed lower bound, $y = 0.01$, so that the intragroup intensity for each node is guaranteed to be larger than the intergroup intensity. In general, a bigger z leads to denser networks with higher node heterogeneity. For non-diagonal $\boldsymbol{V}^{(a,b)} \in \mathbb{R}^{N_a}(a \neq b)$, we start from $\boldsymbol{V}^{(a,a)} \in \mathbb{R}^{N_a}$ and multiply the N_a entries with N_a elements sampled from $(0,1)$ to adapt to the normal situation where interaction heterogeneity with different communities exists for each node. All non-diagonal entries

are multiplied by an additional parameter $w(w > 0)$ which controls the strength of community structure.

S4.2 Parameter settings

Here we list ten settings used in our paper for reference:

I. $w = 0.8$, $z = 0.05$, $\alpha_1 = \alpha_2 = 0.005$, $\beta_1 = \beta_2 = 0.01$;

II. $w = 0.5$, $z = 0.05$, $\alpha_1 = \alpha_2 = 0.005$, $\beta_1 = \beta_2 = 0.01$;

III. $w = 1.5$, $z = 0.05$, $\alpha_1 = \alpha_2 = 0.005$, $\beta_1 = \beta_2 = 0.01$;

IV. $w = 0.8$, $z = 0.05$, $\alpha_1 = 0.009$, $\alpha_2 = 0.005$, $\beta_1 = 0.012$, $\beta_2 = 0.01$;

V. $w = 0.8$, $z = 0.05$, $\alpha_1 = 0.004$, $\alpha_2 = 0.005$, $\beta_1 = 0.012$, $\beta_2 = 0.01$;

VI. $w = 0.8$, $z = 0.03$, $\alpha_1 = \alpha_2 = 0.005$, $\beta_1 = \beta_2 = 0.01$;

VII. $w = 0.8$, $z = 0.08$, $\alpha_1 = \alpha_2 = 0.005$, $\beta_1 = \beta_2 = 0.01$;

In these settings, I,II,III are used to study the simulation performance with varied w, I,IV,V are used to study the simulation performance with α_1/β_1, I,VI,VII are used to study the simulation performance with varied z.

S4.3 An analysis of the effect of parameters

In our simulation study, two additional parameters z and w are introduced to control the network density and the community strength, respectively.

In this subsection we analyse the influence of the triggering ratio α/β and the community background strength parameter w.

Table S1: Some summary statistics for accumulated networks with 150 nodes when $T = 1000$ under different parameter settings. All values are rounded up to 2 decimal places.

Settings	$\mathbf{MD}_{\text{within}}$	$\mathbf{SD}_{\text{within}}$	$\mathbf{MD}_{\text{between}}$	$\mathbf{SD}_{\text{between}}$
I	41.63	32.10	30.23	20.01
II	41.63	32.10	10.71	9.19
III	41.63	32.10	105.99	61.86
IV	69.61	53.31	30.59	18.38
V	32.55	24.40	29.53	17.99

$\mathbf{MD}_{\text{within}}$ represents the mean degree within communities, $\mathbf{SD}_{\text{within}}$ represents the standard deviation of degrees within communities, $\mathbf{MD}_{\text{between}}$ represents the mean degree between communities, and $\mathbf{SD}_{\text{between}}$ represents the standard deviation of degrees between communities.

Table S1 summarizes the degree distribution of the accumulated network under settings I, II, III, IV, and V when $N = 150$ and $T = 1000$. A smaller $w < 1$ or a larger $w > 1$ enlarges the gap of background intensities between intragroup and intergroup block pairs and leads to a stronger community structure and a varying node heterogeneity (see settings I, II and III). A larger α/β triggers more subsequent events for dyads with high background intensities, causing stronger node heterogeneity. This fact can also be verified by the intensity parameter $\widetilde{\lambda}_{ib} = (T/(1-\alpha_{c_ib}/\beta_{c_ib}))^{1/2}\lambda_{ib}$ established in Lemma 2. In particular, when fixing α_2/β_2 and varying α_1/β_1, we will also change the community structure due to the baseline differences between intergroup and intragroup intensities (see settings I, IV, and V).

S4.4 Performance with a known K

Our first simulation study demonstrates the convergence of membership and intensity parameter with T when varying w, z and α/β. First, we simulate networks under CHHIP and perform the estimation procedure in Algorithm 2. The end time T of these networks takes values of $200, 500, 1000, 3000$ and 5000. In this simulation study, we assume that the community number $K = 3$ is known in advance so that we can calculate the Frobenius norms of the estimation errors with an appropriate permutation of community labels (i.e., the minimal MSE when community labels are permuted), $\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0\|_F$, $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0\|_F$, and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_F$. In addition, we use the adjusted rand index (ARI, Rand (1971), Hubert and Arabie (1985)), a widely used measure, to compare the clustering performance of our method under different settings.

Table S2: Averaged event number for different parameter settings I-VII.

setting\T	200	500	1000	3000	5000
I	2415.00	7191.88	15713.40	50500.00	86986.40
II	2059.36	6047.64	13584.96	43113.68	73060.44
III	3918.08	11614.04	25629.92	81686.20	139620.20
IV	2926.24	9871.80	23955.84	84747.60	148875.52
V	2220.76	6384.72	13376.20	41618.08	69564.00
VI	1034.12	3156.96	7062.88	22527.00	38193.20
VII	5438.16	16312.96	35386.92	115508.96	196640.64

In Figure S1 and Figure S2, cases I, II, III and I, IV, V correspond to settings with different baseline community strength parameter w and the intragroup triggering ratio α_1/β_1, respectively. The average numbers

of events of setting I are 2415.00, 7191.88, 15713.40, 50500.00 and 86986.40 corresponding to different values of T. More information about event numbers in other settings are provided in Table S2 above.

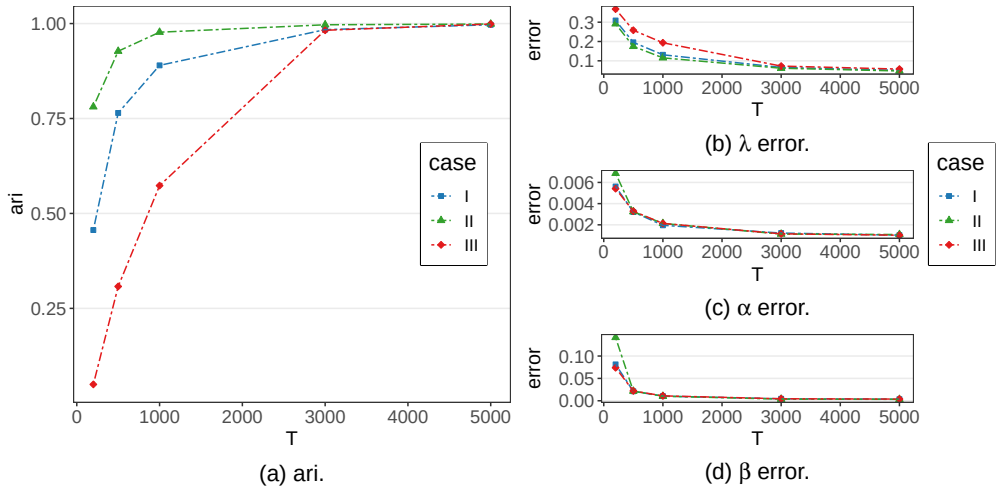


Figure S1: Average ARI (left) and Frobenius error (right) for estimation for 50 replicates when varying w.

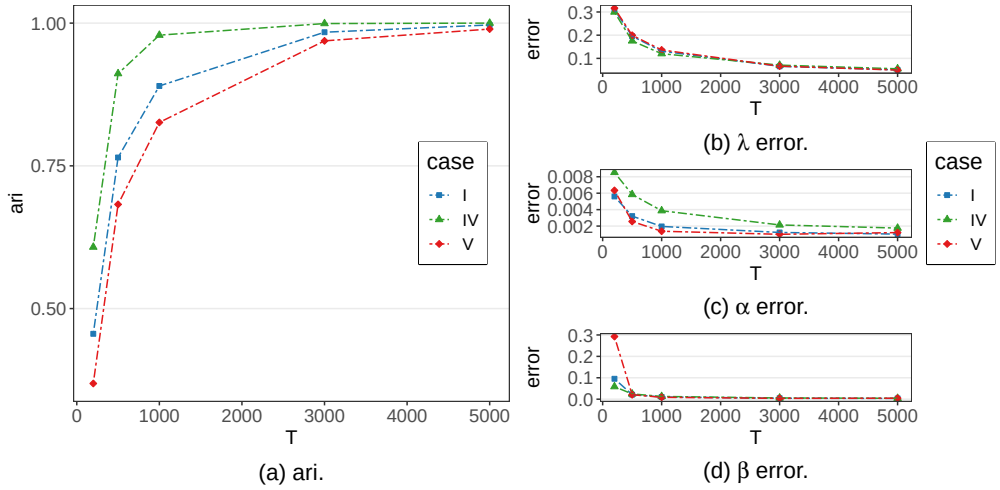


Figure S2: Average ARI (left) and Frobenius error (right) for estimation for 50 replicates when varying α_1/β_1.

In the left panels, we observe relatively low ARIs when the end time T is small because the network is too sparse to display its community structure. The ARI rises rapidly with T and converges to 1 in all cases. Note that even in case III where nodes in different communities may interact more than in the same community, ARI can still converge when T is large. Indeed, our method does not require the assumption that intragroup connection is stronger than intergroup ones, which is often needed for spectral or modularity clustering methods, and therefore can perform well as long as the event information is sufficiently large. Additionally, ARI for networks with stronger communities (smaller $w < 1$ and larger α_1/β_1, or larger $w > 1$) converge faster, which is consistent with our intuition.

The right panels demonstrate the parameter estimation errors. The estimation improves in all cases when T increases, although $\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0\|_F$ takes a relatively larger value due to its high dimensionality. We also find that the estimation performance of $\boldsymbol{\lambda}$ is sensitive to the clustering results, while those of triggering parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are less sensitive. This is reasonable because $\boldsymbol{\lambda}$ contains the baseline intensity specific to each node and thus directly relies on the correctness of clustering, while $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are specific to each group. In general, the parameters are better estimated as the clustering results become more accurate.

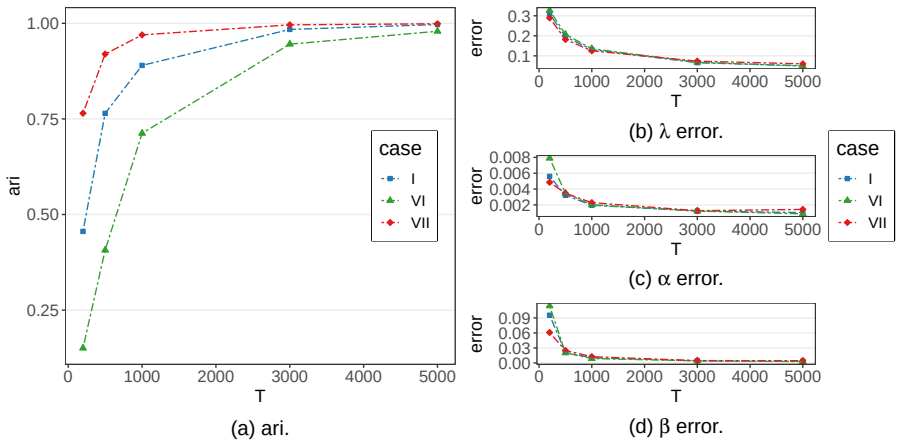


Figure S3: Average ARI (left) and Frobenius error (right) for estimation for 50 replicates when varying z.

To assess the convergence of parameters under different sparsity levels, we vary the baseline density parameter z and consider settings I, VI, VII. The clustering results are shown in the left panel of Figure S3, and we find that a bigger z leads to faster convergence of the ARI. This finding is intuitively acceptable because a denser network provides more information about communities. Correspondingly, a denser network generally makes it easier for intensity parameters to converge. However, the effects of sparsity levels on parameter convergence are not that obvious compared to node heterogeneity and community structure.

For each case in I-VII, we compare the initial clustering results via SSC and the final clustering results after the EM updates. The corresponding results are presented in Figure S4-Figure S6. In all cases the final clustering

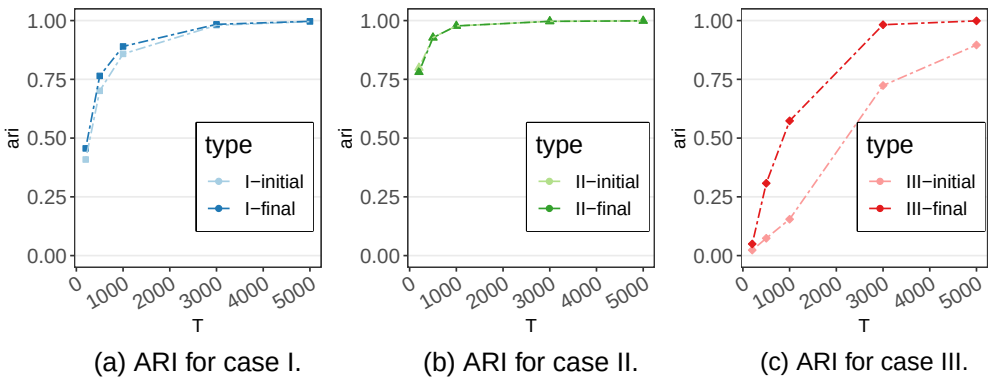
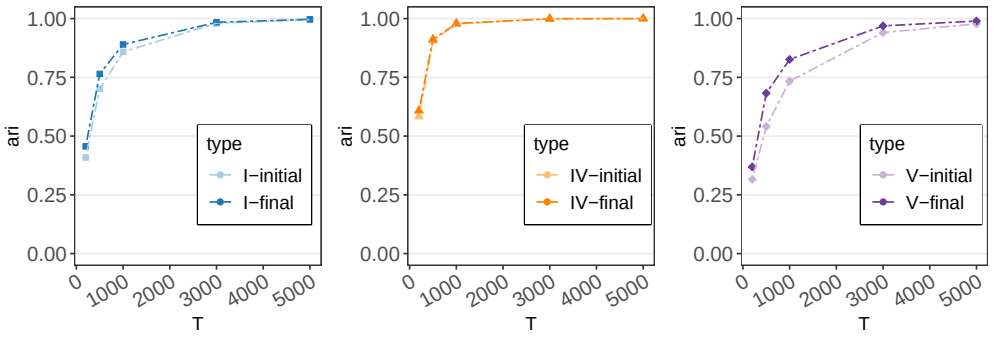


Figure S4: Average ARI for initial and final clustering results estimated for 50 replicates when varying w.

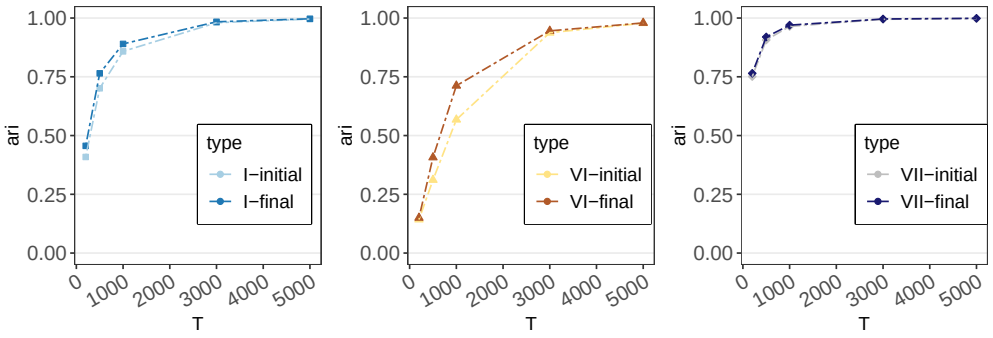
results are better than the initial one, especially in case III where intergroup background intensities are stronger than intragroup background intensities. Despite that the SSC clustering has been established to be consistent with T, it omits the timestamp of interactions and does not make full use of the data information. Therefore, when the end time T is finite, it will always be inferior to the final community membership. This sufficiently demonstrates the effectiveness of updating community labels in the EM algorithm.

Finally, we vary the node number n while fixing $K = 3$ and $T = 1000$ to assess the scalability of the proposed method. From Table S3, our method is scalable to a moderate-sized dynamic network with millions of interactions.



(a) ARI for case I. (b) ARI for case IV. (c) ARI for case V.

Figure S5: Average ARI for initial and final clustering results estimated for 50 replicates when varying α_1/β_1.



(a) ARI for case I. (b) ARI for case VI. (c) ARI for case VII.

Figure S6: Average ARI for initial and final clustering results estimated for 50 replicates when varying z.

Table S3: Averaged event number and corresponding computational time for 20 replicates when varying n. Computations are performed on a linux platform with a 24-core CPU.

Node number	Average event number	Computation time (s)
90	5647.4	35.4
150	15832.5	56.6
300	63706.7	321.3
600	255744.8	3701.3
1200	1023493.3	58822.6

S4.5 Model superiority under different generative settings

The second simulation compares our model with two existing models to verify its rationality. For example, the CHIP model accounts only for the community structure but not the node heterogeneity. To introduce a model with node heterogeneity but without community structure, we design a simple model named additive degree corrected Hawkes process (ADCHP). In this model, triggering parameters are identical on all node pairs and the background intensity is of an additive form $\gamma_i + \gamma_j$. This additive form is motivated by the famous β-model for static networks. By doing so, we have $\lambda^*_{ij}(t) = \gamma_i + \gamma_j + \sum_{t^{(s)}_{ij}<t} \alpha e^{-\beta(t-t^{(s)}_{ij})}$ for the node pair (i, j) under ADCHP.

In reality, the generative mechanism of networks with communities may not accord with our proposed CHHIP model. To make a fair comparison of the methods introduced above, we simulate data from the following three models:

Model 1: $\lambda^*_{ij}(t) = \mu_{c_i c_j} + \sum_{t^{(s)}_{ij}<t} \alpha_{c_i c_j} e^{-\beta_{c_i c_j}(t-t^{(s)}_{ij})}$,

Model 2: $\lambda^*_{ij}(t) = \theta_i W_{c_i c_j} \theta_j + \sum_{t^{(s)}_{ij}<t} \alpha_{c_i c_j} e^{-\beta_{c_i c_j}(t-t^{(s)}_{ij})}$,

Model 3: $\lambda^*_{ij}(t) = \lambda_{ic_j} \lambda_{jc_i} + \sum_{t^{(s)}_{ij}<t} \alpha_{c_i c_j} e^{-\beta_{c_i c_j}(t-t^{(s)}_{ij})}$.

All these models incorporate the community structure and fit a univariate Hawkes process for each dyad. Models 1 and 3 are the the original CHIP and the proposed CHHIP, respectively. Model 2 serves as a compromise between

the other two models by assuming a DCBM-type background intensity.

We assume that $\alpha = 0.005$ and $\beta = 0.01$ for all node pairs in these three models so the self-exciting parts are identical. Then, we specify distinct background intensities, which is the main difference between them. For Model 1, we let $\mu_{c_i c_j} = 0.001$ when $c_i = c_j$ and $\mu_{c_i c_j} = 0.0001$ when $c_i \neq c_j$. For Model 2, we fix $W_{c_i c_j} = 1$ when $c_i = c_j$ and $W_{c_i c_j} = 0.8$ when $c_i \neq c_j$. Then, for each node we generate a background-degree-related parameter $\tilde{\theta}_i$ following the discrete power law distribution with lower bound $m_{\min} = 1$ and scaling parameter $\alpha = 2$. For identifiability, $\sum_i \theta_i 1_{\{c_i=k\}} = 1$ for any $k \in \{1, \cdots, K\}$ is required in Model 2, so we further normalize $\tilde{\boldsymbol{\theta}}$ by $\theta_i = \widetilde{\theta}_i / \sum_j \widetilde{\theta}_j 1_{\{c_j=c_i\}}$ to obtain $\boldsymbol{\theta}$. For Model 3, we set $z = 0.05$ and $w = 0.8$. We still assume a known community number $K = 3$ in this simulation study. Because the parameter numbers of these models may differ from the fitting methods, we use Akaike information criterion (AIC) instead of parameter error for assessment. We divide the AIC by the total event number to calculate mean AIC, which keeps the result relatively constant with T and presents the difference between these methods more clearly.

Figure S7 exhibits estimation results using the three methods, CHIP, CHHIP, and ADCHP when networks are generated according to the model settings explained above. From Figure S7, ADCHP does not perform well in

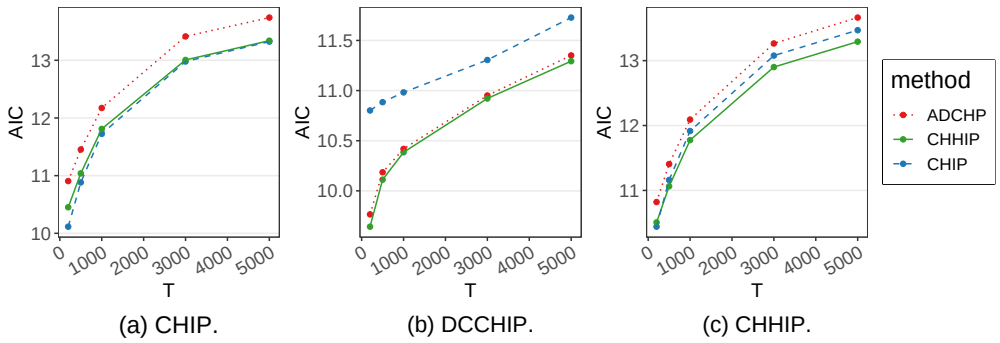


Figure S7: Mean AICs of different fitting methods under Models 1–3.

cases of CHIP and CHHIP for the lack of community structure. When the true model is Model 1, CHHIP is quite comparable to CHIP. In other words, the additional parameters in CHHIP would not increase the model complexity too much. However, the importance of these parameters emerges when the true model is Model 2 or Model 3, as nodal heterogeneity appears. The CHHIP always outperforms other models in these situations, especially when the duration is long enough.

S4.6 Selection of the community number K

Finally, we consider the case of an unknown community number, which is more common in reality. Let $\widehat{\boldsymbol{\theta}}^{\widehat{\boldsymbol{c}}}$ be the estimated parameters depending on $\widehat{\boldsymbol{c}}$, which is the membership vector given by the clustering method. In this experiment, we vary the true value of K from 3 to 5 and repeat the simulation-and-estimation process 50 times for each K. Specifically, we

simulate data from Model 3 with $\alpha = 0.005$, $\beta = 0.01$, $z = 0.05$ and $w = 0.8$. We perform the estimation procedure for every candidiate $\widehat{K}$ and select the $\widehat{K}$ that minimizes the HQ defined in Section 3.2 from $\{2, 3, 4, 5\}$. We report the times of correct selection (i.e., $\widehat{K} = K$) and provide the Frobenius error averaged over samples where the community numbers are correctly selected in Table S4.

It is easily seen that an increasing K leads to slower convergence rates for both clustering membership and parameter estimators because the model becomes more complex. As T increases, the HQ criterion can select the correct K in almost all cases, as long as the event information is sufficient.

Table S4: Times of $\hat{K} = K$ and average Frobenius error without ground community labels. Error values are rounded up to 4 decimal places.

K	T	$\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0\|_F$	$\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0\|_F$	$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_F$	No. of $(\widehat{K} = K)$
3	200	0.3169	0.0064	0.0925	2
	500	0.1932	0.0035	0.0224	47
	1000	0.1328	0.0019	0.0100	50
	3000	0.0670	0.0014	0.0049	50
	5000	0.0498	0.0011	0.0037	49
4	200	-	-	-	0
	500	-	-	-	0
	1000	0.1813	0.0029	0.0133	49
	3000	0.0912	0.0017	0.0053	50
	5000	0.0675	0.0016	0.0046	50
5	200	-	-	-	0
	500	-	-	-	0
	1000	0.2223	0.0043	0.0177	8
	3000	0.1194	0.0022	0.0061	50
	5000	0.0889	0.0021	0.0061	50

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.

Bandeira, A. S. and R. Van Handel (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability 44*(4), 2479 – 2506.

Hawkes, A. G. and D. Oakes (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability 11*(3), 493–503.

Hsu, D., S. Kakade, and T. Zhang (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability 17*(52), 1 – 6.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*(1), 193–218.

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.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*(336), 846–850.

Zaatour, R. (2014). *hawkes: Hawkes process simulation and calibration toolkit*. R package version 0.0-4.