CHIP: A Hawkes Process Model for Continuous-time Networks with Scalable and Consistent Estimation

Makan Arastuie
EECS Department
University of Toledo
makan.arastuie
@rockets.utoledo.edu

Subhadeep Paul
Department of Statistics
The Ohio State University
paul.963@osu.edu

Kevin S. Xu
EECS Department
University of Toledo
kevin.xu@utoledo.edu

Abstract

In many application settings involving networks, such as messages between users of an on-line social network or transactions between traders in financial markets, the observed data consist of timestamped relational events, which form a continuous-time network. We propose the *Community Hawkes Independent Pairs (CHIP)* generative model for such networks. We show that applying spectral clustering to an aggregated adjacency matrix constructed from the CHIP model provides *consistent community detection* for a growing number of nodes and time duration. We also develop consistent and computationally efficient estimators for the model parameters. We demonstrate that our proposed CHIP model and estimation procedure scales to large networks with tens of thousands of nodes and provides superior fits than existing continuous-time network models on several real networks.

1 Introduction

A variety of complex systems in the computer, information, biological, and social sciences can be represented as a network, which consists of a set of objects (nodes) and relationships (edges) between the objects. In many application settings, we observe edges in the form of distinct events occurring between nodes over time. For example, in on-line social networks, users interact with each other through events that occur at specific time instances such as liking, mentioning, or sharing another user's content. Such interactions form *timestamped relational events*, where each event is a triplet (i, j, t) denoting events from node i (sender) to node j (receiver) at timestamp t. The observation of these triplets defines a dynamic network that continuously evolves over time.

Timestamped relational event data are usually modeled by combining a point process model for the event times with a network model for the sender and receiver [1–9]. We refer to such models as *continuous-time network models* because they provide probabilities of observing events between two nodes during arbitrarily short time intervals. For model-based exploratory analysis and prediction of future events with relational event data, continuous-time network models are often superior to their discrete-time counterparts [10–14], which first aggregate events over time windows to form discrete-time network "snapshots" and thus lose granularity in modeling temporal dynamics.

We propose the *Community Hawkes Independent Pairs (CHIP)* model, which is inspired by the recently proposed Block Hawkes Model (BHM) [9] for timestamped relational event data. Both CHIP and BHM are based on the Stochastic Block Model (SBM) for static networks [15]. In the BHM, events between different pairs of nodes belonging to the same pair of communities are dependent, which makes it difficult to analyze. In contrast, for CHIP the pairs of nodes in the same community generate events according to *independent* univariate Hawkes processes with shared parameters, so that the number of parameters remains the same as in the BHM. The independence between node pairs enables tractable analysis of the CHIP model and more scalable estimation than the BHM.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Our main contributions are as follows. (1) We demonstrate that spectral clustering provides consistent community detection in the CHIP model for a growing number of nodes and time duration. (2) We propose consistent and computationally efficient estimators for the model parameters also for a growing number of nodes and time duration. (3) We show that the CHIP model provides better fits to several real datasets and scales to much larger networks than existing models, including a Facebook network with over $40{,}000$ nodes and over $800{,}000$ events. Other point process network models have demonstrated good empirical results, but to the best of our knowledge, this work provides the *first theoretical guarantee of estimation accuracy*. Our asymptotic analysis also has tremendous practical value given the scalability of our model to large networks with tens of thousands of nodes.

2 Background

2.1 Hawkes Processes

The Hawkes process [16] is a counting process designed to model continuous-time arrivals of events that naturally cluster together in time, where the arrival of an event increases the chance of the next event arrival immediately after. They have been used to model earthquakes [17], financial markets [18, 19], and user interactions on social media [3, 20].

A univariate Hawkes process is a *self-exciting* point process where its conditional intensity function given a sequence of event arrival times $\{t_1, t_2, t_3, ..., t_l\}$ for l events up to time duration T takes the general form $\lambda(t) = \mu + \sum_{t_i<t}^{t_l} \gamma(t - t_i)$, where μ is the background intensity and $\gamma(\cdot)$ is the kernel or the excitation function. A frequent choice of kernel is an exponential kernel, parameterized by $\alpha, \beta > 0$ as $\gamma(t - t_i) = \alpha e^{-\beta(t-t_i)}$, where the arrival of an event instantaneously increases the conditional intensity by the jump size α, after which the intensity decays exponentially back towards μ at rate β. Restricting $\alpha < \beta$ yields a stationary process. We use an exponential kernel for the CHIP model, since it has been shown to provide a good fit for relational events in social media [9, 21–23].

2.2 The Stochastic Block Model

Statistical models for networks typically consider a static network rather than a network of relational events. Many static network models are discussed in the survey by Goldenberg et al. [24]. A static network with n nodes can be represented by an $n \times n$ adjacency matrix A where $A_{ij} = 1$ if there is an edge between nodes i and j and $A_{ij} = 0$ otherwise. We consider networks with no self-edges, so $A_{ii} = 0$ for all i. For a directed network, we let $A_{ij} = 1$ if there is an edge from node i to node j.

One model that has received significant attention is the *stochastic block model* (SBM), formalized by Holland et al. [15]. In the SBM, every node i is assigned to one and only one community or *block* $c_i \in \{1, \ldots, k\}$, where k denotes the total number of blocks. For a directed SBM, given the block membership vector $\mathbf{c} = [c_i]_{i=1}^n$, all off-diagonal entries of the adjacency matrix A_{ij} are independent Bernoulli random variables with parameter p_{c_i,c_j}, where p is a $k \times k$ matrix of probabilities. Thus the probability of forming an edge between nodes i and j depends only on the block memberships c_i and c_j. There have been significant advancements in the analysis of estimators for the SBM. Several variants of spectral clustering [25], including regularized versions [26, 27], have been shown to be consistent estimators of the community assignments in the SBM and various extensions in several asymptotic settings [28–39]. Spectral clustering scales to large networks with tens of thousands of nodes and is generally not sensitive to initialization, so it is also a practically useful estimator.

2.3 Related Work

One approach for modeling continuous-time networks is to treat the edge strength of each node pair as a continuous-time function that increases when an event occurs between the node pair and then decays afterwards [40–42]. Another approach is to combine a point process model for the event times, typically some type of Hawkes process, with a network model. The conditional intensity functions of the point processes then serve as the time-varying edge strengths. Point process network models are used in two main settings. The first involves estimating the structure of a latent or unobserved network from observed events at the nodes [43–48]. These models are often used to estimate *static* networks of diffusion from information cascades.

In the second setting, which we consider in this paper, we directly observe events *between pairs of nodes* so that events take on the form (i, j, t) denoting an event from node i to node j at timestamp t. Our objective is to model the dynamics of such event sequences. In many applications, including messages on on-line social networks, most pairs of nodes either never interact and thus have no events between them. Thus, most prior work in this setting utilizes low-dimensional latent variable representations of the networks to parameterize the point processes.

The latent variable representations are often inspired by generative models for static networks such as continuous latent space models [49] and stochastic block models [15], resulting in the development of point process network models with continuous latent space representations [6] and latent block or community representations [1–5, 7–9], respectively. Point process network models with latent community representations are most closely related to the model we consider in this paper. Exact inference in such models is intractable due to the discrete nature of the community assignments. Approximate inference techniques including Markov Chain Monte Carlo (MCMC) [2, 3, 7] or variational inference [5, 8, 9] have been used in prior work. While such techniques have demonstrated good empirical results, to the best of our knowledge, they come with no theoretical guarantees.

3 The Community Hawkes Independent Pairs (CHIP) Model

We consider a generative model for timestamped relational event networks that we call the *Community Hawkes Independent Pairs (CHIP)* model. The CHIP model has parameters $(\boldsymbol{\pi}, \mu, \alpha, \beta)$. Each node is assigned to a community or block $a \in \{1, \ldots, k\}$ with probability π_a, where each entry of $\boldsymbol{\pi}$ is non-negative and all entries sum to 1. We represent the block assignments of all nodes either by a length n vector $\mathbf{c} = [c_i]_{i=1}^n$ or an $n \times k$ binary matrix C where $c_i = q$ is equivalent to $C_{iq} = 1$, $C_{iq'} = 0$ for all $q' \neq q$. Each of the parameters μ, α, β is a $k \times k$ matrix. While we assume that the number of blocks and the block assignments of the nodes do not change with time, the CHIP model captures time-varying behavior due to the incorporation of self-exciting point processes. Event times between node pairs (i, j) within a block pair (a, b) follow independent exponential Hawkes processes with shared parameters: baseline rate μ_{ab}, jump size α_{ab}, and decay rate β_{ab}. The generative process for our proposed CHIP model is as follows:

$$
\begin{aligned}
c_i &\sim \text{Categorical}(\boldsymbol{\pi}) && \text{for all nodes } i \\
\mathbf{t}_{ij} &\sim \text{Hawkes process}(\mu_{c_i c_j}, \alpha_{c_i c_j}, \beta_{c_i c_j}) && \text{for all } i \neq j \\
Y &= \text{Row concatenate } [(i\mathbf{1}, j\mathbf{1}, \mathbf{t}_{ij})] && \text{over all } i \neq j
\end{aligned}
$$

$\mathbf{1}$ denotes the all-ones vector of appropriate length. Let T denote the end time of the Hawkes process, which would correspond to the duration of the data trace. The column vector of event times $\mathbf{t}_{ij}$ has length $N_{ij}(T)$, which denotes the number of events from node i to node j up to time T. Let Y denote the event matrix with dimension $l \times 3$, where $l = \sum_{i,j} N_{ij}(T)$ denotes the total number of observed events over all node pairs. It is constructed by row concatenating triplets $(i, j, t_{ij}(q))$ over all events $q \in \{1, \ldots, N_{ij}(T)\}$ for all node pairs $i, j \in \{1, \ldots, n\}, i \neq j$.

3.1 Relation to Other Models

Our proposed CHIP model has a generative structure inspired by the SBM for static networks. Other point process network models in the literature have also utilized similar block structures, but they have been incorporated in two different approaches. One approach involves placing point process models at the level of block pairs [2, 4, 5, 9]. For a network with k blocks, k^2 different point processes are used to generate events between the k^2 block pairs. To generate events between pairs of nodes, rather than pairs of blocks, the point processes are thinned by randomly selecting nodes from the respective blocks so that all nodes in a block are stochastically equivalent, in the spirit of the SBM. Such models have demonstrated good empirical results, but the dependency between node pairs complicates analysis of the models.

The other approach involves modeling pairs of nodes with independent point processes that share parameters among nodes in the same block [1, 3, 8]. By having node pairs in the same block share parameters, the number of parameters is the same as for the models with block pair-level point processes. However, by using independent point processes for all node pairs, there is no dependency between node pairs, which simplifies analysis of the model. We use this approach in the proposed CHIP model and exploit this independence to perform the theoretical analysis in Section 4.

Algorithm 1 Estimation procedure for Community Hawkes Independent Pairs (CHIP) model

Input: Relational event matrix Y, number of blocks k

Result: Estimated block assignments $\hat{C}$ and CHIP model parameters $(\hat{\boldsymbol{\pi}}, \hat{\mu}, \hat{\alpha}, \hat{\beta})$

for all node pairs $i \neq j$ **do**
  N_{ij} = number of events from i to j in Y
$\hat{C} \leftarrow$ Spectral clustering(N, k)
for all block pairs (a, b) **do**
  Compute estimates $(\hat{m}_{ab}, \hat{\mu}_{ab})$ using (2)
  $\hat{\beta}_{ab} \leftarrow$ maximize log-likelihood by line search
  $\hat{\alpha}_{ab} \leftarrow \hat{\beta}_{ab}\hat{m}_{ab}$
$\hat{\boldsymbol{\pi}} \leftarrow$ proportion of nodes in each block
return $[\hat{C}, \hat{\boldsymbol{\pi}}, \hat{\mu}, \hat{\alpha}, \hat{\beta}]$

3.2 Estimation Procedure

As with many other block models, the maximum-likelihood estimator for the discrete community assignments C is intractable except for extremely small networks (e.g. 20 nodes). We propose a scalable estimation procedure for the CHIP model that has two components as shown in Algorithm 1: community detection and parameter estimation. For the community detection component, we use spectral clustering on the weighted adjacency or count matrix $N(T)$ or simply N with entries $N_{ij}(T)$. Since this is a directed adjacency matrix, we use singular vectors rather than eigenvectors for spectral clustering (see Algorithm A.1 in the supplementary material for details).

For the parameter estimation component, we first consider estimating the Hawkes process parameters $(\mu_{ab}, \alpha_{ab}, \beta_{ab})$ for each block pair (a, b) using only the count matrix N, which discards event timestamps. Even without the event timestamps, we are able to estimate μ_{ab} and the ratio $m_{ab} = \alpha_{ab}/\beta_{ab}$, but not the parameters α_{ab} and β_{ab} separately. Define the following terms, which are the sample mean and (unbiased) sample variance of the pairwise event counts within each block pair:

$$\bar{N}_{ab} = \frac{1}{n_{ab}} \sum_{i,j:C_{ia}=1,C_{jb}=1} N_{ij}, \quad S^2_{ab} = \frac{1}{n_{ab}-1} \sum_{i,j:C_{ia}=1,C_{jb}=1} (N_{ij} - \bar{N}_{ab})^2, \tag{1}$$

where n_{ab} denotes the number of node pairs in block pair (a, b) and is given by $n_{ab} = |a||b|$ for $a \neq b$ and $n_{ab} = |a||a-1|$ for $a = b$, with $|a|$ denoting the number of nodes in block a. $\bar{N}_{ab}$ and S^2_{ab} are unbiased estimators of the mean and variance, respectively, of the counts of the number of events between all node pairs (i, j) in block pair (a, b). Using $\bar{N}_{ab}$ and S^2_{ab}, we propose the following method of moments estimators (conditioned on the estimated blocks) for m_{ab} and μ_{ab} from the count matrix N:

$$\hat{m}_{ab} = 1 - \sqrt{\frac{\bar{N}_{ab}}{S^2_{ab}}}, \quad \hat{\mu}_{ab} = \frac{1}{T}\sqrt{\frac{(\bar{N}_{ab})^3}{S^2_{ab}}}. \tag{2}$$

Finally, the vector of block assignment probabilities $\boldsymbol{\pi}$ can be easily estimated using the proportion of nodes in each block, i.e. $\hat{\pi}_a = \frac{1}{n}\sum_{i=1}^{n} \hat{C}_{ia}$ for all $a = 1, \ldots, k$.

In some prior work, exponential Hawkes processes are parameterized only in terms of m and μ, with β treated as a known parameter that is not estimated [50–52]. In this case, the estimation procedure is complete. On the other hand, if we want to estimate the values of both α and β rather than just their ratio, we have to use the actual event matrix Y with the event timestamps. To separately estimate the α_{ab} and β_{ab} parameters, we replace $\alpha_{ab} = \beta_{ab} m_{ab}$ in the exponential Hawkes log-likelihood for block pair (a, b) then plug in our estimate $\hat{m}_{ab}$ for m_{ab}. Then the log-likelihood is purely a function of β_{ab} and can be maximized using a standard scalar optimization or line search method.

3.3 Selection of the Number of Blocks

The estimation procedure in Algorithm 1 assumes that the number of blocks k is provided. In many practical settings, k is unknown, and choosing k becomes a model selection problem. Given that CHIP uses spectral clustering on the weighted adjacency matrix N, model selection approaches for static block models can be used to find the optimal k. These range in complexity from the eigengap

heuristic [25] to more sophisticated methods including using eigenvalues of the non-backtracking matrix and Bethe Hessian matrix [53] and network cross validation [54, 55]. Another approach, specific to the timestamped network setting we consider in this paper, is to hold out a portion of the events, e.g. the last 20%, and select the k that maximizes the log-likelihood on the held-out events.

4 Theoretical Analysis of Estimators

4.1 Analysis of Estimated Community Assignments

We define the error of community detection as the misclustering error rate $r = \inf_{\Pi} \frac{1}{n} \sum_{i=1}^{n} 1(c_i \neq \Pi(\hat{c}_i))$, where $\Pi(\cdot)$ denotes the set of all permutations of the community labels. Our proposed CHIP model considers directed events; however, we analyze community detection on undirected networks to better match up with the literature on analysis of spectral clustering for the SBM. The bounds and consistency properties we derive still apply to the directed case with only a change in the constants. We assume that $T \to \infty$, which can be achieved by rescaling the time unit for event times. Under this assumption, the mean and variance of the number of events between nodes (i, j) are [56–58]

$$\nu_{ab} = \frac{\mu_{ab} T}{1 - \alpha_{ab}/\beta_{ab}}, \quad \sigma_{ab}^2 = \frac{\mu_{ab} T}{(1 - \alpha_{ab}/\beta_{ab})^3}. \tag{3}$$

We analyze community detection error in a simplified special case of our CHIP model which is in similar spirit to a commonly-employed case in the stochastic block models literature [28, 31, 32, 35, 59]. We provide analogous results for the general CHIP model in Section B.1 of the supplementary material. In this special case, all communities have roughly equal number of elements $|a| \asymp n/k$, all intra-community processes (diagonal block pairs) have the same set of parameters μ_1, α_1, β_1 and all inter-community processes (off-diagonal block pairs) have the same set of parameters μ_2, α_2, β_2. We use the notation $Y \sim \text{CHIP}(C, n, k, \mu_1, \alpha_1, \beta_1, \mu_2, \alpha_2, \beta_2)$ to denote a relational event matrix Y generated from this simplified model. Define $m_1 = \alpha_1/\beta_1$ and $m_2 = \alpha_2/\beta_2$. Let $\nu_1 = \mu_1/(1 - m_1)$ and $\nu_2 = \mu_2/(1 - m_2)$, while $\sigma_1^2 = \mu_1/(1 - m_1)^3$ and $\sigma_2^2 = \mu_2/(1 - m_2)^3$. Assume $\nu_1 > \nu_2$, $\nu_1 \asymp \nu_2$, and $\sigma_1 \asymp \sigma_2$, where the asymptotic equivalence is with respect to both n and T. These assumptions imply that the expected number of events are higher between two nodes in the same community compared to two nodes in different communities and that the asymptotic dependence on n and T are the same for both set of parameters. This setting is useful to understand detectability limits and has been widely employed in the literature on stochastic block models [32, 35, 36, 59–61]. In this setting, we have the following upper bound on the misclustering error rate.

Theorem 1 *Let $Y \sim$ CHIP$(C, n, k, \mu_1, \alpha_1, \beta_1, \mu_2, \alpha_2, \beta_2)$. The misclustering error rate for spectral clustering on the weighted adjacency matrix N at time $T \to \infty$ is*

$$r \lesssim \frac{T\sigma_1^2 n}{(n/k)^2(\nu_2 - \nu_1)^2 T^2} \asymp \frac{k^2}{nT} \frac{\sigma_1^2}{(\nu_1 - \nu_2)^2}.$$

We note that if the set of parameters μ, α, β remain constant as a function of n and T then the misclustering error rate decreases as $1/T$ with increasing T, decreases as $1/n$ with increasing n, and increases as k^2 with increasing k. Hence, as we observe the process for more time, spectral clustering on N has lower error rate. The rate of convergence with increasing T is the same as one would obtain for detecting an average community structure if discrete snapshots of the network were available over time [38, 39, 59]. The dependence of the misclustering error rate on n and k is what one would expect from the SBM literature.

Theorem 1 applies also in the sparse graph setting. We let $\mu \asymp 1/[f(n)g(T)]$, a function of n and T, and explore various sparsity settings by varying f and g in Section B.1.1 of the supplementary material. Our proofs allow μ to vary with n and T and can be as small as $\log(n)/(nT)$. A key component in the proof of Theorem 1 is a bound from Bandeira and van Handel [62]. In Section B.1 of the supplementary material, we provide the proof of Theorem 1, an analogous theorem for the general CHIP model, as well as theorems for spectral clustering on an unweighted adjacency matrix.

4.2 Analysis of Estimated Hawkes Process Parameters

As discussed in Section 3.2, we are able to estimate $m = \alpha/\beta$ and μ from the count matrix N using (2). We analyze these estimators assuming a growing number of nodes n and time duration

T. We do not put any assumption on the distribution of the counts; we only require that T is large enough such that the asymptotic mean and variance equations in (3) hold. The sample mean $\bar{N}_{ab}$ and sample variance S^2_{ab} of the counts are unbiased estimators of ν_{ab} and σ^2_{ab}, respectively. The following theorem shows that these estimators are consistent and asymptotically normal.

Theorem 2 *Define $n_{\min} = \min_{a,b} n_{ab}$. The estimators for m_{ab} and μ_{ab} have the following asymptotic distributions as $n_{\min} \to \infty$ and $T \to \infty$:*

$$\sqrt{n_{ab}}\left(\hat{m}_{ab} - \left(1 - \sqrt{\frac{\nu_{ab}}{\sigma^2_{ab}}}\right)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4\nu_{ab}}\right), \ \sqrt{n_{ab}}\left(\hat{\mu}_{ab}T - \frac{(\nu_{ab})^{3/2}}{\sigma_{ab}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{9}{4}\nu_{ab}\right).$$

Using Theorem 2, we obtain confidence intervals for μ and m, in Section B.2.1 of the supplementary material. In the simplified special case of Theorem 1, we have equal community sizes so $n_{ab} \asymp (n/k)^2$. Therefore, the condition $n_{\min} \to \infty$ boils down to $(n/k)^2 \to \infty$, which is a reasonable assumption. Theorem 2 guarantees convergence of our estimators for μ and m with the asymptotic mean-squared errors (MSEs) decreasing at the rate $n_{ab} \asymp (n/k)^2$ under the assumption that the community structure is correctly estimated. Next, we provide an "end-to-end" guarantee for the convergence of the asymptotic MSE to 0 for estimating the mean number of events in each block pair ν_{ab} using the sample mean $\bar{N}_{ab}$ incorporating the error in estimating communities using spectral clustering from Theorem 1.

Theorem 3 *Assume $n_{ab} \asymp (n/k)^2$. The weighted average of asymptotic MSEs in estimating ν_{ab} using the estimator $\bar{N}_{ab}$ with communities estimated by spectral clustering is*

$$\frac{\sum_{ab} n_{ab} E[(\bar{N}_{ab} - \nu_{ab})^2]}{\sum_{ab} n_{ab}} \lesssim \frac{kT}{n} \max\left\{\sigma_1^2, \frac{k^2\sigma_1^2\nu_2^2}{(\nu_1 - \nu_2)^2}\right\}.$$

For comparison, under the assumption that the community structure is correctly estimated, the weighted average of asymptotic MSEs in estimating ν_{ab} using the estimator $\bar{N}_{ab}$ is

$$\frac{\sum_{ab} n_{ab} E[(\bar{N}_{ab} - \nu_{ab})^2]}{\sum_{ab} n_{ab}} = \frac{k^2 T \sigma_1^2}{n^2}.$$

Theorem 3 guarantees that the MSE for estimating Hawkes process parameters decreases at least at a linear rate with increasing (n/k) when the error from community detection is taken into account instead of the quadratic rate when the error is not taken into account. The proofs for Theorems 2 and 3 are provided in Section B.2.2 of the supplementary material.

5 Experiments

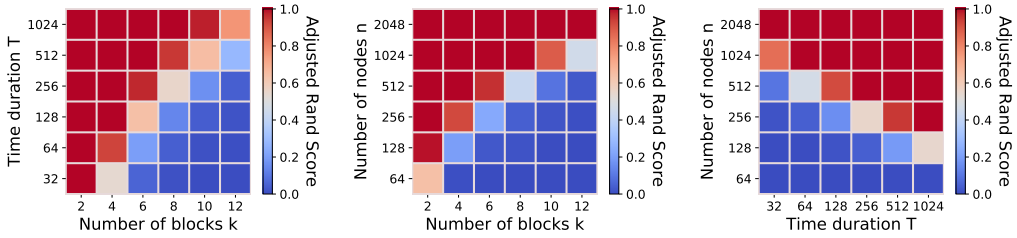
We begin with a set of simulation experiments to assess the accuracy of our proposed estimation procedure and verify our theoretical analysis. We then present several experiments on real data involving both prediction and model-based exploratory analysis. Additional experiments are provided in Section C of the supplementary material, along with the code[1] to replicate all experiments.

5.1 Community Detection on Simulated Networks with Varying T, n, and k

We simulate networks from the simplified CHIP model while varying two of T, n, and k simultaneously. We choose parameters $\mu_1 = 0.085$, $\mu_2 = 0.065$, $\alpha_1 = \alpha_2 = 0.06$, and $\beta_1 = \beta_2 = 0.08$. The upper bounds on the error rates in Theorem 1 involve all three parameters n, k, T simultaneously, making it difficult to interpret the result. To better observe the effects of n, k, T and their relationship with respect to each other, we perform three separate simulations each time varying two and fixing the other one. The community detection accuracy averaged over 30 simulations using the weighted adjacency matrix N as two of T, n, and k are varied is shown in Figure 1. Since the estimated community assignments will be permuted compared to the actual community labels, we evaluate the community detection accuracy using the adjusted Rand score [63], which is 1 for perfect community detection and has an expectation of 0 for a random assignment.

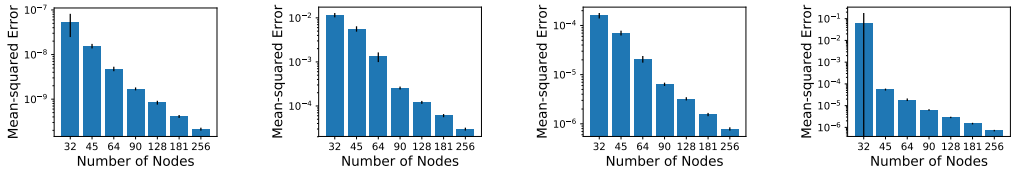
Note that Theorem 1 predicts that the misclustering error rate varies as $k^2/(nT)$ if all three parameters are varied. Figure 1(a) shows the accuracy to be low for small T and large k. As we simultaneously

[1] Code available on GitHub: `https://github.com/IdeasLabUT/CHIP-Network-Model`.



(a) Fixed $n = 256$ (b) Fixed $T = 64$ (c) Fixed $k = 8$

Figure 1: Heat map of adjusted Rand score of spectral clustering on weighted adjacency matrix, with varying T, n, and k, averaged over 30 simulated networks.



(a) μ: base intensity (MSE decay rate: 2.00) (b) m: α to β ratio (MSE decay rate: 2.05) (c) α: intensity jump size (MSE decay rate: 2.01) (d) β: intensity decay rate (MSE decay rate: 2.08)

Figure 2: Mean-squared errors (MSEs) of CHIP's Hawkes parameter estimators averaged over 100 simulations ($\pm$ 2 standard errors) on a log-log plot. MSEs for all four parameters decreases as the number of nodes increases, with the estimated decay rate (exponent) beginning at 90 nodes listed.

increase T and decrease k the accuracy improves until the adjusted Rand score reaches 1. We also note that it is possible to obtain high accuracy either with increasing T or decreasing k or with both even when n is fixed. This is in line with the prediction from Theorem 1 that the misclustering error rate varies as k^2/T if n remains fixed. We observe a similar effect of increasing accuracy with increasing n and decreasing k when T is kept fixed in Figure 1(b). Finally, Figure 1(c) verifies the prediction that accuracy increases with both increasing n and T for a fixed k.

5.2 Hawkes Process Parameter Estimation on Simulated Networks

Next, we examine the estimation accuracy of the Hawkes process parameter estimates as described in Section 4.2. We simulate networks from the simplified CHIP model with $k = 4$ blocks, duration $T = 10{,}000$ and parameters $\mu_1 = 0.0011$, $\mu_2 = 0.0010$, $\alpha_1 = 0.11$, $\alpha_2 = 0.09$, $\beta_1 = 0.14$, and $\beta_2 = 0.16$ so that each parameter is different between block pairs. We then run the CHIP estimation procedure: spectral clustering followed by Hawkes process parameter estimation.

Figure 2 shows the mean-squared errors (MSEs) of all four estimators decay quadratically as n increases. Theorem 2 states that $\hat{m}$ and $\hat{\mu}$ are consistent estimators with MSE decreasing at a quadratic rate for growing n with known communities. Here, we observe the quadratic decay even with communities estimated by spectral clustering, where the mean adjusted Rand score is increasing from 0.6 to 1 as n grows. We observe that α and β are also accurately estimated for growing n even though β is estimated using a line search for which we have no guarantees.

5.3 Comparison with Other Models on Real Networks

We perform experiments on three real network datasets. Each dataset consists of a set of events where each event is denoted by a sender, a receiver, and a timestamp. The MIT Reality Mining [64] and Enron [65] datasets were loaded and preprocessed identically to DuBois et al. [3] to allow for a fair comparison with their reported values. On the Facebook wall posts dataset [66], we use the largest connected component of the network excluding self loops ($43,953$ nodes). Additional details about the datasets and preprocessing are provided in Section C.2.1 of the supplementary material.

Table 1: Mean test log-likelihood per event for each real network dataset across all models. Larger (less negative) values indicate better predictive ability. Bold entry denotes best fit for a dataset. Results for REM are reported values from DuBois et al. [3]. Poisson denotes spectral clustering followed by estimating a Poisson process baseline model. *The BHM local search does not scale up to the Facebook network, so we report results using the (less accurate) spectral clustering procedure.

Dataset	Statistics	Model	$k=1$	$k=2$	$k=3$	$k=10$	Best k
Reality	$n=70$ $l_{\text{train}}=1{,}500$ $l_{\text{test}}=661$	CHIP	-4.83	-4.88	-5.06	-6.69	$\mathbf{-4.83}$ $(k=1)$
		REM	-6.78	-7.42	-6.11	-6.61	-6.11 $(k=3)$
		BHM	-9.05	-7.56	-7.60	-5.74	-5.37 $(k=50)$
		Poisson	-10.3	-10.4	-9.63	-9.38	-8.51 $(k=32)$
Enron	$n=142$ $l_{\text{train}}=3{,}000$ $l_{\text{test}}=1{,}000$	CHIP	-5.63	-5.61	-5.65	-7.15	$\mathbf{-5.61}$ $(k=2)$
		REM	-7.02	-6.86	-6.84	-7.26	-6.84 $(k=3)$
		BHM	-8.72	-8.43	-8.39	-7.93	-7.49 $(k=8)$
		Poisson	-11.9	-11.4	-11.5	-12.0	-11.4 $(k=4)$
Facebook	$n=43{,}953$ $l_{\text{train}}=682{,}266$ $l_{\text{test}}=170{,}567$	CHIP	-9.54	-9.58	-9.58	-9.61	$\mathbf{-9.46}$ $(k=9)$
		BHM*	-16.0	-15.7	-16.2	-14.7	-14.4 $(k=22)$
		Poisson	-20.8	-21.1	-21.1	-20.6	-19.2 $(k=55)$

We fit our proposed Community Hawkes Independent Pairs (CHIP) model as well as the Block Hawkes Model (BHM) [9] to all three real datasets and evaluate their fit. We also compare against a simpler baseline: spectral clustering with a homogeneous Poisson process for each node pair. For each model, we also compare against the case $k=1$, where no community detection is being performed. We do not have ground truth community labels for these real datasets so we cannot evaluate community detection accuracy. Instead, we use the mean test log-likelihood per event as the evaluation metric, which allows us to compare against the reported results in DuBois et al. [3] for the relational event model (REM). Since the log-likelihood is computed on the test data, this is a measure of the model's ability to *forecast future events* rather than detect communities.

As shown in Table 1, CHIP outperforms all other models in all three datasets. Note that test log-likelihood is maximized for CHIP at relatively small values of k compared to the BHM. This is because CHIP assumes independent node pairs whereas the BHM assumes all node pairs in a block pair are dependent. Thus, the BHM needs a higher value for k in order to model independence. This difference is particularly visible for the Reality Mining data, where CHIP with $k=1$ is the best predictor of the test data, while the best BHM has $k=50$ on a network with only 70 nodes! These both suggest a weak community structure that is not predictive of future events in the Reality Mining data, whereas community structure does appear to be predictive in the Enron and Facebook data.

In addition to the improved predictive ability of CHIP compared to the BHM, the computational demand is also significantly decreased. Fitting the CHIP model for each value of k took on average 0.15 s and 0.3 s on the Reality Mining and Enron datasets, respectively, while the BHM took on average 250 s and 30 m, mostly due to the time-consuming local search[2]. We did not implement the MCMC-based inference procedure for the REM and thus do not have results for REM on the Facebook data or computation times. The approach of holding out a test set of events and evaluating test log-likelihood can also be used for selection of the number of blocks k. As shown in Figure 3, on the Facebook data, there is hardly any increase in the runtime of CHIP for $k<100$, and it is manageable even for $k=1{,}000$.

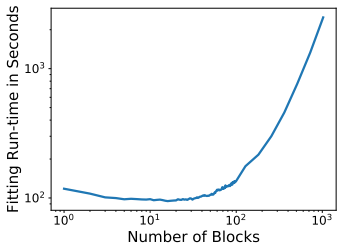
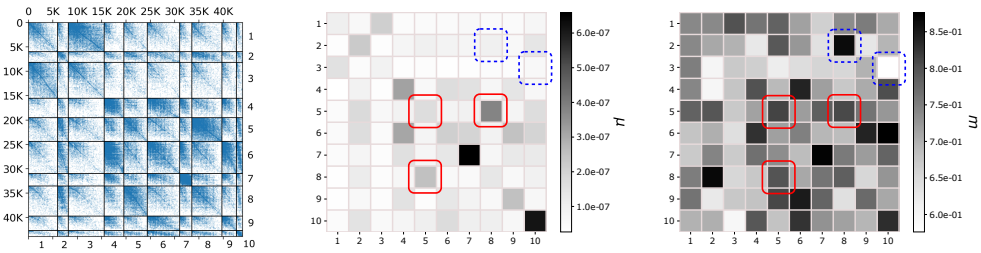


Figure 3: CHIP's fitting runtime on the Facebook data on a log-log scale with increasing k.

5.4 Model-Based Exploratory Analysis of Facebook Wall Post Network

We use CHIP to perform model-based exploratory analysis to understand the behavior of different groups of users in the Facebook wall post network. We consider all $852{,}833$ events and choose $k=10$ blocks using the eigengap heuristic [25], which required 141 s to fit. Note that the CHIP

[2]Experiments were run on a workstation with 2 Intel Xeon 2.3 GHz CPUs with a total of 36 cores.



(a) Adjacency matrix with blocks (b) μ: base intensity (c) m: branching ratio α/β

Figure 4: Inferred CHIP parameters on the largest connected component of the Facebook Wall Posts dataset with $k = 10$. Axis labels denote block numbers. Each tile corresponds to a block pair where (a, b) denotes row a and column b. Boxed block pairs in the heatmap are discussed in the body text.

estimation procedure can scale up to a much higher number of communities also—fitting CHIP to the Facebook data with $k = 1{,}000$ communities took just under 50 minutes! The adjacency matrix permuted by the block structure is shown in Figure 4(a), and heatmaps of the fitted CHIP parameters are shown in Figures 4(b) and 4(c). Diagonal block pairs on average have a base intensity μ of 2.8×10^{-7}, which is higher compared to 9.5×10^{-8} for off-diagonal block pairs, indicating an underlying assortative community structure. However, not all blocks have higher rates of within-block posts, e.g. $\mu_{5,8} > \mu_{5,5}$ and $\mu_{8,5} > \mu_{5,5}$, as shown in red boxes in Figure 4(b), which illustrates that the CHIP model does not discourage inter-block events. These patterns often occur in social networks, for instance, if there are communities with opposite views on a particular subject.

While the structure of μ reveals insights on the baseline rates of events between block pairs, the structure of the branching ratio $m = \alpha/\beta$ shown in Figure 4(c) reveals insights on the burstiness of events between block pairs. For some block pairs, such as $(3, 10)$, there are very low values of α and β indicating the events are closely approximated by a homogeneous Poisson process, while some block pairs such as $(2, 8)$ are extremely bursty despite low baseline rates. Both block pairs are shown in blue dashed boxes. The different levels of burstiness of block pairs cannot be seen from aggregate statistics such as the the count matrix N.

6 Conclusion

We introduced the Community Hawkes Independent Pairs (CHIP) model for timestamped relational event data. The CHIP model has many similarities with the Block Hawkes Model (BHM) [9]; however, in the CHIP model, events among any two node pairs are independent, which enables both tractable theoretical analysis and scalable estimation. We demonstrated that an estimation procedure using spectral clustering followed by Hawkes process parameter estimation provides consistent estimates of the communities and Hawkes process parameters for a growing number of nodes and time duration. Lastly, we showed that CHIP also provides better fits to several real networks compared to the Relational Event Model [3] and the BHM. It also scales to considerably larger data sets, including a Facebook wall post network with over $40{,}000$ nodes and $800{,}000$ events.

There are several limitations to the CHIP model and our proposed estimation procedure. Assuming all node pairs to have independent Hawkes processes simplifies analysis and increases scalability but also reduces the flexibility of the model compared to multivariate Hawkes process-based models that specifically encourage reciprocity [2, 7]. Additionally, our estimation procedure uses unregularized spectral clustering to match our theoretical analysis in Section 4. We note that regularized versions of spectral clustering [26, 27, 30, 34, 37] have been found to perform better empirically and would likely improve the community detection accuracy in the CHIP model. Methods that jointly estimate the community structure and Hawkes process parameters, such as the local search and variational inference approaches explored in Junuthula et al. [9] for the Block Hawkes Model could also improve estimation accuracy of both. Also, methods that integrate change point detection with estimation for continuous-time block models could be used to allow for community structure to change over time [8], resulting in more flexible models.

Broader Impact

Our proposed CHIP model can be applied to analyze any type of timestamped relational event data. In this paper, we considered analysis of mobile phone calls, emails, and user interactions on on-line social networks. However, timestamped relational event data is used in a variety of other disciplines, including financial mathematics, e.g. transactions between traders in financial markets [19]; political science, e.g. military deployments between countries [2, 67]; and sociology, e.g. homicides between gangs in a city [43, 68]. Thus, our CHIP model can have broader impact to society through the advancement of multiple research disciplines.

The CHIP model, like other generative models for dynamic networks, can be used for forecasting, e.g. to predict which nodes are likely to have an event, as well as the number of events during a specified time interval. For some applications, the forecasts may themselves be used to affect decision making. For example, in public policy, crime forecasting can be used for predictive policing, which affects the allocation of police resources to different locations over time. This can have societal benefits, as a recent randomized controlled field trial for predictive policing using Hawkes process models for prediction demonstrated a 7% reduction in crime [69], but also potential for negative consequences like arrests that are biased with respect to minority communities, although such consequences were not observed in the randomized trial [70]. In this paper, we analyzed a publicly available anonymized Facebook on-line social network dataset, so we are not aware of negative consequences that may result from our proposed model.

Acknowledgments and Disclosure of Funding

This material is based upon work supported by the National Science Foundation grants DMS-1830412 and IIS-1755824.

References

[1] Christopher DuBois and Padhraic Smyth. Modeling relational events via latent classes. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 803–812, 2010.

[2] Charles Blundell, Jeff Beck, and Katherine A. Heller. Modelling reciprocating relationships with Hawkes processes. In *Advances in Neural Information Processing Systems 25*, pages 2600–2608, 2012.

[3] Christopher DuBois, Carter T. Butts, and Padhraic Smyth. Stochastic blockmodeling of relational event dynamics. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pages 238–246, 2013.

[4] Lu Xin, Mu Zhu, and Hugh Chipman. A continuous-time stochastic block model for basketball networks. *The Annals of Applied Statistics*, 11(2):553–597, 2017.

[5] Catherine Matias, Tabea Rebafka, and Fanny Villers. A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, 105(3):665–680, 2018.

[6] Jiasen Yang, Vinayak Rao, and Jennifer Neville. Decoupling homophily and reciprocity with latent space network models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2017.

[7] Xenia Miscouridou, François Caron, and Yee Whye Teh. Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data. In *Advances in Neural Information Processing Systems 31*, pages 2343–2352, 2018.

[8] Marco Corneli, Pierre Latouche, and Fabrice Rossi. Multiple change points detection and clustering in dynamic networks. *Statistics and Computing*, 28(5):989–1007, 2018.

[9] Ruthwik R. Junuthula, Maysam Haghdan, Kevin S. Xu, and Vijay K. Devabhaktuni. The Block Point Process Model for continuous-time event-based dynamic networks. In *Proceedings of the World Wide Web Conference*, pages 829–839, 2019.

[10] Eric P. Xing, Wenjie Fu, and Le Song. A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, 4:535–566, 2010.

[11] Tianbao Yang, Yun Chi, Shenghuo Zhu, Yihong Gong, and Rong Jin. Detecting communities and their evolutions in dynamic social networks—a Bayesian approach. *Machine Learning*, 82 (2):157–189, 2011.

[12] Kevin S. Xu and Alfred O. Hero III. Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):552–562, 2014.

[13] Kevin S. Xu. Stochastic block transition models for dynamic networks. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 1079–1087, 2015.

[14] Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141, 2017.

[15] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

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

[17] David Marsan and Olivier Lengline. Extending earthquakes' reach through cascading. *Science*, 319(5866):1076–1079, 2008.

[18] Paul Embrechts, Thomas Liniger, and Lu Lin. Multivariate Hawkes processes: an application to financial data. *Journal of Applied Probability*, 48(A):367–378, 2011.

[19] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(1):1550005, 2015.

[20] Ke Zhou, Hongyuan Zha, and Le Song. Learning triggering kernels for multi-dimensional Hawkes processes. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1301–1309, 2013.

[21] Peter F. Halpin and Paul De Boeck. Modelling dyadic interaction with Hawkes processes. *Psychometrika*, 78(4):793–814, 2013.

[22] Naoki Masuda, Taro Takaguchi, Nobuo Sato, and Kazuo Yano. Self-exciting point process modeling of conversation event sequences. In *Temporal networks*, pages 245–264. Springer, 2013.

[23] Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1513–1522, 2015.

[24] Anna Goldenberg, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

[25] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[26] Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 35.1–35.23, 2012.

[27] Arash A. Amini, Aiyou Chen, Peter J. Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.

[28] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[29] Daniel L. Sussman, Minh Tang, Donniell E. Fishkind, and Carey E. Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

[30] Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems 26*, pages 3120–3128, 2013.

[31] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

[32] Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Proceedings of The 28th Conference on Learning Theory*, pages 391–423, 2015.

[33] Qiuyi Han, Kevin S. Xu, and Edoardo Airoldi. Consistent estimation of dynamic and multi-layer block models. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1511–1520, 2015.

[34] Antony Joseph and Bin Yu. Impact of regularization on spectral clustering. *The Annals of Statistics*, 44(4):1765–1791, 2016.

[35] Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou. Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, 18(1):1980–2024, 2017.

[36] Van Vu. A simple SVD algorithm for finding hidden partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, 2018.

[37] Yilin Zhang and Karl Rohe. Understanding regularized spectral clustering via graph conductance. In *Advances in Neural Information Processing Systems 31*, pages 10631–10640, 2018.

[38] Sharmodeep Bhattacharyya and Shirshendu Chatterjee. Spectral clustering for multiple sparse networks: I. *arXiv preprint arXiv:1805.10594*, 2018.

[39] Marianna Pensky and Teng Zhang. Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, 13(1):678–709, 2019.

[40] Emily M. Jin, Michelle Girvan, and M. E. J. Newman. Structure of growing social networks. *Physical Review E*, 64(4):046132, 2001.

[41] Walid Ahmad, Mason A. Porter, and Mariano Beguerisse-Díaz. Tie-decay temporal networks in continuous time and eigenvector-based centralities. *arXiv preprint arXiv:1805.00193*, 2018.

[42] Xinzhe Zuo and Mason A. Porter. Models of continuous-time networks with tie decay, diffusion, and convection. *arXiv preprint arXiv:1906.09394*, 2019.

[43] Scott W. Linderman and Ryan P. Adams. Discovering latent network structure in point process data. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1413–1421, 2014.

[44] Scott W. Linderman and Ryan P. Adams. Scalable Bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*, 2015.

[45] Mehrdad Farajtabar, Yichen Wang, Manuel Gomez Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. COEVOLVE: A joint point process model for information diffusion and network co-evolution. In *Advances in Neural Information Processing Systems 28*, pages 1945–1953, 2015.

[46] Long Tran, Mehrdad Farajtabar, Le Song, and Hongyuan Zha. NetCodec: Community detection from individual activities. In *Proceedings of the SIAM International Conference on Data Mining*, pages 91–99, 2015.

[47] Xinran He, Theodoros Rekatsinas, James Foulds, Lise Getoor, and Yan Liu. HawkesTopic: A joint model for network inference and topic modeling from text-based cascades. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 871–880, 2015.

[48] Eric C. Hall and Rebecca M. Willett. Tracking dynamic point processes on networks. *IEEE Transactions on Information Theory*, 62(7):4327–4346, 2016.

[49] Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

[50] Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pages 641–649, 2013.

[51] Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Mean-field inference of hawkes point processes. *Journal of Physics A: Mathematical and Theoretical*, 49 (17):174006, 2016.

[52] Emmanuel Bacry, Martin Bompaire, Philip Deegan, Stéphane Gaïffas, and Søren V Poulsen. Tick: a Python library for statistical learning, with an emphasis on Hawkes processes and time-dependent models. *The Journal of Machine Learning Research*, 18(1):7937–7941, 2017.

[53] Can M. Le and Elizaveta Levina. Estimating the number of communities in networks by spectral methods. *arXiv preprint arXiv:1507.00827*, 2015.

[54] Kehui Chen and Jing Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113(521):241–251, 2018.

[55] Tianxi Li, Elizaveta Levina, and Ji Zhu. Network cross-validation by edge sampling. *Biometrika*, 107(2):257–276, 2020.

[56] Alan G. Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.

[57] Alan G. Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 438–443, 1971.

[58] P. A. W. Lewis. Asymptotic properties and equilibrium conditions for branching Poisson processes. *Journal of Applied Probability*, 6(2):355–371, 1969.

[59] Subhadeep Paul and Yuguo Chen. Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *The Annals of Statistics*, 48(1):230–250, 2020.

[60] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, pages 670–688. IEEE, 2015.

[61] Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[62] Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506, 2016.

[63] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2: 193–218, 1985.

[64] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36): 15274–15278, 2009.

[65] Bryan Klimt and Yiming Yang. The Enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning*, pages 217–226. Springer, 2004.

[66] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, pages 37–42, 2009.

[67] Zeev Maoz, Paul L. Johnson, Jasper Kaplan, Fiona Ogunkoya, and Aaron P. Shreve. The dyadic militarized interstate disputes (MIDs) dataset version 3.0: Logic, characteristics, and comparisons to alternative datasets. *Journal of Conflict Resolution*, 63(3):811–835, 2019.

[68] Andrew V. Papachristos. Murder by structure: Dominance relations and the social structure of gang homicide. *American Journal of Sociology*, 115(1):74–128, 2009.

[69] George O. Mohler, Martin B. Short, Sean Malinowski, Mark Johnson, George E. Tita, Andrea L. Bertozzi, and P. Jeffrey Brantingham. Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, 110(512):1399–1411, 2015.

[70] P. Jeffrey Brantingham, Matthew Valasik, and George O. Mohler. Does predictive policing lead to biased arrests? results from a randomized controlled trial. *Statistics and Public Policy*, 5(1): 1–6, 2018.

Supplemental Material: CHIP: A Hawkes Process Model for Continuous-time Networks with Scalable and Consistent Estimation

Makan Arastuie
EECS Department, University of Toledo
makan.arastuie@rockets.utoledo.edu

Subhadeep Paul
Department of Statistics, The Ohio State University
paul.963@osu.edu

Kevin S. Xu
EECS Department, University of Toledo
kevin.xu@utoledo.edu

A Additional Details on Estimation Procedure

A.1 Community Detection

The spectral clustering algorithm for directed networks that we consider in this paper is shown in Algorithm A.1. It can be applied either to the weighted adjacency (count) matrix N or the unweighted adjacency matrix A, where $A_{ij} = 1\{N_{ij} > 0\}$ and $1\{\cdot\}$ denotes the indicator function of the argument. This algorithm is used for the community detection step in our proposed CHIP estimation procedure. For undirected networks, which we use for the theoretical analysis in Section 4, spectral clustering is performed by running k-means clustering on the rows of the *eigenvector* matrix of N or A, not the rows of the concatenated singular vector matrix.

A.2 Estimation of Hawkes process parameters

Ozaki (1979) derived the log-likelihood function for Hawkes processes with exponential kernels, which takes the form:

$$\log \mathcal{L} = -\mu T + \sum_{q=1}^{l} \frac{\alpha}{\beta}\{e^{-\beta(T-t_q)} - 1\} + \sum_{q=1}^{l} \log(\mu + \alpha w(q)) \tag{A.1}$$

where $w(q) = \sum_{q':t_{q'}<t_q} e^{-\beta(t_q - t_{q'})}$. Moreover, $w(q)$ can be computed recursively using $w(q) = e^{-\beta(t_q - t_{q-1})}(1 + w(q-1))$, with the added base case of $w(1) = 0$, which drops the double summation in the last term and decreases the computational complexity of the log-likelihood from $\mathcal{O}(l^2)$ to $\mathcal{O}(l)$ (Laub et al., 2015). The three parameters μ, α, β can be estimated by maximizing (A.1) using standard numerical methods for non-linear optimization (Nocedal & Wright, 2006).

In our CHIP model, we have separate (μ, α, β) parameters for each block pair (a, b). We provide closed-form equations for estimating $m_{ab} = \alpha_{ab}/\beta_{ab}$ and μ_{ab} in (2). To separately estimate the α_{ab}

Algorithm A.1 Spectral clustering algorithm for community detection in directed networks

Input: Adjacency Matrix N, number of blocks k
Result: Estimated block assignments $\hat{C}$

1: Compute singular value decomposition of N
2: $\hat{\Sigma} \leftarrow$ diagonal matrix of k largest singular values of N
3: $\hat{U}, \hat{V} \leftarrow$ left and right singular vectors of N corresponding to k largest singular values
4: $\hat{Z} \leftarrow$ concatenate($\hat{U}$, $\hat{V}$)
5: Normalize the magnitude of each row of $\hat{Z}$ to 1
6: $\hat{C} \leftarrow$ k-means clustering on rows of $\hat{Z}$
7: **return** $\hat{C}$

and β_{ab} parameters, we replace $\alpha_{ab} = \beta_{ab} m_{ab}$ in the exponential Hawkes log-likelihood (A.1) for block pair (a, b) to obtain

$$\log \mathcal{L}(\beta_{ab} | C, [\mathbf{t}_{ij}]_{i,j=1}^{n}) = \sum_{i,j: C_{ia}=1, C_{jb}=1} \Bigg\{ -\mu_{ab} T + \sum_{q=1}^{N_{ij}} m_{ab} \{ e^{-\beta_{ab}(T - t_{ij}^{q})} - 1 \} + \sum_{q=1}^{N_{ij}} \log(\mu_{ab} + \beta_{ab} m_{ab} w_{ij}(q)) \Bigg\} \tag{A.2}$$

where $w_{ij}(q) = \sum_{q': t_{ij}^{q'} < t_{ij}^{q}} e^{-\beta_{ab}(t_{ij}^{q} - t_{ij}^{q'})}$ for $q \geq 2$ and $w_{ij}(1) = 0$. We substitute in the estimates for m_{ab} and μ_{ab} from (2). Then the log-likelihood (A.2) is purely a function of β_{ab} and can be maximized using a standard scalar optimization or line search method. In our experiments, we perform the line search using SciPy's function `minimize_scalar(method="bounded")`.

B Additional Theoretical Analysis of Estimators

We present additional results on estimation of community assignments in Section B.1 along with proofs and discussions. We then provide proofs of our results for estimated Hawkes process parameters in Section B.2 along with discussions on obtaining confidence intervals for the estimated Hawkes process parameters.

B.1 Estimated Community Assignments

We define the notation $Y \sim \text{CHIP}(C, n, k, \mu, \alpha, \beta)$ to denote that relational event matrix Y is generated from a CHIP model with n nodes, k blocks, community assignment matrix C and Hawkes process parameter matrices (μ, α, β). To characterize the misclustering rate of a spectral clustering algorithm applied to N, we define the following quantities. Let $\lambda_{\min}(E[N])$ denote the minimum in absolute value non-zero eigenvalue of the matrix $E[N]$. Define

$$s = \sqrt{T} \max_{a} \sqrt{\sum_{b} |b| \frac{\mu_{ab}}{(1 - \alpha_{ab}/\beta_{ab})^3}}, \tag{B.1}$$

$$s_1 = \sqrt{T} \max_{a,b} \sqrt{\frac{\mu_{ab}}{(1 - \alpha_{ab}/\beta_{ab})^3}}. \tag{B.2}$$

Then we have the following upper bound on the misclustering error rate.

Theorem B.1. *Let $Y \sim CHIP(C, n, k, \mu, \alpha, \beta)$. Then, with probability at least $1 - 1/n$, the misclustering error rate for spectral clustering on the weighted adjacency matrix N at time $T \to \infty$ is*

$$r \leq 64(2+\epsilon_1)|a|_{\max} k \frac{\left\{(1+\epsilon)(2s + \frac{6}{\log(1+\epsilon)} s_1\sqrt{\log n}) + s_1\sqrt{\log n}\right\}^2}{n(\lambda_{\min}(E[N])^2},$$

where $0 < \epsilon < 1/2$ and $\epsilon_1 > 0$ are constants.

Theorem B.1 provides an upper bound to the error rate of spectral clustering on the weighted adjacency matrix N in the setting $T \to \infty$. Note that the assumption of $T \to \infty$ does not preclude us from being able to analyze scenarios where the network is sparse since the expected number of events between a pair of nodes ν_{ab} can be made constant or even $o(1)$ by setting $\frac{\mu_{ab}}{1-\alpha_{ab}/\beta_{ab}} = O(1/T)$ and $\frac{\mu_{ab}}{1-\alpha_{ab}/\beta_{ab}} = o(1/T)$ respectively.

It is also possible to obtain an expression for the mean as a function of T without the assumption of $T \to \infty$ using stochastic differential equations (Laub et al., 2015; Da Fonseca & Zaatour, 2014). In particular, if we substitute the starting intensity $\lambda_0 = \mu$, i.e., the process starts with baseline intensity as we have assumed throughout, and the starting number of events $N_0 = 0$, then from the result of Da Fonseca & Zaatour (2014) and Daw & Pender (2018),

$$E[N_{ij}] = \frac{\mu_{ab} T}{1 - \alpha_{ab}/\beta_{ab}} - \frac{\mu_{ab}\alpha_{ab}\left[1 - e^{-(\beta_{ab}-\alpha_{ab})T}\right]}{(\beta_{ab} - \alpha_{ab})^2}.$$

We note that there is a small negative correction term to the asymptotic mean, since $\mu_{ab}, \alpha_{ab}, \beta_{ab}, T$ are all non-negative. The effect of this term is negligible as $T \to \infty$, so we ignore it.

We now present an upper bound on the error rate for communities (analogous to Theorem B.1) estimated from the unweighted adjacency matrix A. For a pair of nodes (i, j) such that $c_i = a$ and $c_j = b$, we have $E[A_{ij}] = E[1\{N_{ij} > 0\}] = P(N_{ij} > 0) = 1 - e^{(-\mu_{ab}T)}$. Now A is a $n \times n$ symmetric matrix whose elements A_{ij} are independent Bernoulli random variables with mean $E[A_{ij}]$. Let $\Delta = \max\{n \max_{i,j} E[A_{ij}], c_0 \log n\}$ for some constant c_0, and note that $n \max_{i,j} E[A_{ij}] = n \max(1-\exp(-\mu_{ab}T)) = n(1-\exp(-\mu_{\max}T))$, where $\mu_{\max} = \max_{a,b} \mu_{ab}$. Further, let $\lambda_{\min}(E[A])$ denote the minimum in absolute value non-zero eigenvalue of the matrix $E[A]$ and $|a|_{\max}$ denote the size of the largest community. Then we have the following upper bound on the error rate of spectral clustering performed on A.

Theorem B.2. *Let $Y \sim CHIP(C, n, k, \mu, \alpha, \beta)$. Then, with probability at least $1 - n^{-r}$, the misclustering error rate for spectral clustering on the binary adjacency matrix A at time T is*

$$r \leq 64(2+\epsilon)\frac{|a|_{\max} kc\Delta}{n(\lambda_{\min}(E[A]))^2},$$

where $\epsilon > 0$ is a constant and $c > 0$ is a constant dependent on c_0 and r.

B.1.1 Simplified Special Case

The upper bounds on the error rates in Theorems B.1 and B.2 are not very informative in terms of their dependencies on key model parameters. In Section 4.1, we considered a simplified special case that allowed us to simplify the constants in Theorem B.1, resulting in Theorem 1, which bounds the misclustering error rate on the weighted adjacency matrix N. Similarly, we have the following result for spectral clustering using the unweighted adjacency matrix A.

Theorem B.3. *Let $Y \sim$ CHIP$(C, n, k, \mu_1, \alpha_1, \beta_1, \mu_2, \alpha_2, \beta_2)$. The misclustering error rate for spectral clustering on the binary adjacency matrix A at time T is*

$$r \lesssim \frac{k^2}{n} \frac{1 - \exp(-\mu_1 T)}{(\exp(-\mu_2 T) - \exp(-\mu_1 T))^2}. \tag{B.3}$$

If further we assume $\mu_1 \asymp \mu_2 \asymp o(1/T)$, such that $\mu_1 T = o(1)$ and $\mu_2 T = o(1)$, then we have

$$r \lesssim \frac{nT\mu_1}{(n/k)^2(\mu_1 - \mu_2)^2 T^2} \asymp \frac{k^2}{nT} \frac{\mu_1}{(\mu_1 - \mu_2)^2}, \tag{B.4}$$

whereas, for $\mu_1 \asymp \mu_2 \asymp \omega(1/T)$, such that $\mu_1 T \to \infty$ and $\mu_2 T \to \infty$, then the upper bound for the misclustering rate in Theorem B.2 goes to 1.

We note that if the parameters are kept constant as a function of T, then $\mu_1 T \to \infty$ and $\mu_2 T \to \infty$. Consequently, without k and n changing the upper bound on the error rate for the unweighted adjacency matrix in Theorem B.2 explodes and becomes close to 1, making the upper bound guarantee useless. While this result might be a drawback of the upper bound result itself, we note that unbounded error makes sense because in this regime almost all node pairs have at least one communication with high probability. Hence the unweighted adjacency matrix has a 1 in almost all entries, and the community structure cannot be detected from this matrix. In that case, we predict that using the weighted adjacency matrix N can lead to smaller error. Theorem 1 provides the corresponding upper bound for error rate for N.

The density of the aggregate adjacency matrix is governed by the parameters of the CHIP model. Hence, to further characterize the dependence of the μ parameters on the number of nodes n and time T in the network, assume $\mu_1 = c_1 \frac{1}{f(n)g(T)}$ and $\mu_2 = c_2 \frac{1}{f(n)g(T)}$, where c_1 and c_2 are constants that do not depend on n or T. Also assume $1 - \alpha_1/\beta_1$ and $1 - \alpha_2/\beta_2$ do not depend on n and T. Then the upper bound on the error rate becomes $r \lesssim \frac{k^2 f(n)g(T)}{nT(c_1 - c_2)^2}$. Now we note that consistent community detection is possible as long as $k = o\left(\frac{\sqrt{nT}|c_1 - c_2|}{f(n)g(T)}\right)$. For example, if we set $g(T) \asymp T$ and $f(n) = \frac{n}{\log n}$, such that $\mu_1 \asymp \mu_2 \asymp \frac{\log n}{nT}$, then the expected number of events between a node pair is $O(\frac{\log n}{n})$. In that case, $r(T) \lesssim \frac{k^2}{\log n(c_1 - c_2)^2}$, and consistent community detection is possible as long as $k = o(\sqrt{\log n}|c_1 - c_2|)$.

A second example is where we set $g(T) \asymp 1$ and $f(n) = \frac{n}{\log n}$, such that $\mu_1 \asymp \mu_2 \asymp \frac{\log n}{n}$. The expected number of events between a vertex pair is then $O(\frac{T \log n}{n})$ and total expected number of events in the whole network is $O(nT \log n)$. In that case $r \lesssim \frac{k^2}{T \log n(c_1 - c_2)^2}$, and consistent community detection is possible as long as $k = o(\sqrt{T \log n}|c_1 - c_2|)$.

B.1.2 Comparison Between Weighted and Unweighted Adjacency Matrices

We compare the bounds on the error rates in unweighted and weighted adjacency matrices in Theorems B.3 and 1 in the sparse regime where $\mu_1 T$ and $\mu_2 T$ are small such that we can apply the Taylor series approximation. From Theorem B.3, we have the error rate using the unweighted adjacency matrix is upper bounded by $\frac{k^2}{nT} \frac{\mu_1}{(\mu_1 - \mu_2)^2}$, while the error rate for the weighted adjacency matrix is upper bounded by

$$\frac{k^2}{nT} \frac{\frac{\mu_1}{(1-m_1)^3} + \frac{\mu_2}{(1-m_2)^3}}{(\frac{\mu_1}{(1-m_1)} - \frac{\mu_2}{(1-m_2)})^2}.$$

We can make the following comparison comments on the basis of these upper bounds.

1. If $m_1 = m_2 = m$ such that the community structure is expressed only through μ_1 and μ_2, then the error for the weighted adjacency matrix is bounded by $\frac{k^2}{nT}\frac{\mu_1+\mu_2}{(\mu_1-\mu_2)^2}\frac{1}{1-m}$. This upper bound is higher than the corresponding upper bound for spectral clustering in unweighted adjacency matrix indicating a possible advantage of using the unweighted adjacency matrix.

2. If $\mu_1 = \mu_2$ such that the community structure is expressed purely through α, β, then the error for the unweighted case is unbounded. However, the error for the weighted case can still be bounded, indicating a possible advantage of the weighted adjacency matrix.

B.1.3 Proofs

We begin with the proofs of Theorems B.1 and B.2 for spectral clustering applied to the weighted and unweighted adjacency matrices, respectively, in the general CHIP model. We then present the proofs of Theorems 1 and B.3 for the simplified special case.

Proof of Theorem B.1

Proof. We start with the following result.

Lemma B.1. *Let $Y \sim CHIP(C, n, k, \mu, \alpha, \beta)$. Let N denote the weighted adjacency matrix obtained by aggregating Y at time $T \to \infty$. Then, with probability at least $1 - 1/n$, we have*

$$\|N - E[N]\|_2 \leq (1+\epsilon)\left\{2s + \frac{6}{\log(1+\epsilon)} s_1\sqrt{\log n}\right\} + 2s_1\sqrt{\log n}, \tag{B.5}$$

where $0 < \epsilon < 1/2$ is a constant, and the terms s and s_1 are as defined in (B.1) and (B.2), respectively.

We present the proof of this lemma following the proof of this theorem.

Since $E[N]$ can also be written in the form of a stochastic block model as $E[N] = C\nu C^T$, we can use the same arguments as in the proof of the previous result. Using the Davis-Kahan Theorem (Davis & Kahan, 1970; Stewart & Sun, 1990), we have the following bound:

$$\begin{aligned} r &\leq \frac{1}{n}|a|_{\max} 8(2+\epsilon_1)\|\hat{U} - C(C^TC)^{-1/2}\mathcal{O}\|_F^2 \\ &\leq 64(2+\epsilon_1)\frac{|a|_{\max} k\|N - E[N]\|_2^2}{n(\lambda_{\min}(E[N])^2}, \end{aligned} \tag{B.6}$$

Combining (B.5) and (B.6), we arrive at the desired result.

□

Proof of Lemma B.1

Proof. We note that N_{ij} is asymptotically normal (Theorem 4 of Hawkes & Oakes (1974)) as $T \to \infty$, i.e.

$$N_{ij}|(C_{ia} = 1, C_{jb} = 1) \sim \mathcal{N}(\nu_{ab}, \sigma_{ab}^2).$$

Then $(N - E[N])$ is a $n \times n$ symmetric matrix with elements $(N - E[N])_{ij} = g_{ij}\sigma_{ij}$, where $g_{ij}; i \geq j$ are i.i.d $\mathcal{N}(0,1)$ and σ_{ij} is the standard deviation of N_{ij} given before.

We will use Corollary 3.9 in Bandeira & van Handel (2016). In the notation of Bandeira & van Handel (2016), we set $\sigma = s$, $\sigma^* = s_1$ and let $t = 2s_1\sqrt{\log n}$. Then for any $0 < \epsilon < 1/2$, we have

$$P\left(\|N - E[N]\|_2 \geq (1+\epsilon)\{2s + \frac{6}{\log(1+\epsilon)} s_1\sqrt{\log n}\} + 2s_1\sqrt{\log n}\right) \leq \exp(-\log n).$$

Since $\exp(-\log n) = 1/n$, one can then take the probability of the complement, which completes the proof. □

Proof of Theorem B.2

Proof. We note that the matrix A is an adjacency matrix with independent entries. Further $n \max_{ij} E[A_{ij}] \leq \Delta$ and $\Delta \geq c_0 \log n$ by definition. Then by Theorem 5.2 of Lei & Rinaldo (2015), we have with probability at least $1 - n^{-r}$,

$$\|A - E[A]\|_2 \leq c\sqrt{\Delta}, \tag{B.7}$$

where c is a constant dependent on c_0 and r.

Since $E[A]$ can be written in the form of a stochastic block model as $E[A] = C(1 - \exp(\mu T))C^T$, we can use known results in the SBM literature. Let $\hat{U}_{n\times k}$ denote the $n \times k$ matrix whose columns are the top k eigenvectors of the matrix A. By Lemma 3.1 of Rohe et al. (2011), the matrix of eigenvectors corresponding to the largest k non-zero eigenvalues of the matrix $E[A]$ is $C(C^TC)^{-1/2}\mathcal{O}$ for some $k \times k$ orthogonal matrix $\mathcal{O}$. Then we have the following relationship for the difference between matrices of population eigenvectors (those of $E[A]$) and sample eigenvectors (those of A) and the misclustering error rate of community detection by applying $(1 + \epsilon)$ approximate k-means algorithm to those matrices (Pensky & Zhang, 2019):

$$r \leq \frac{1}{n}|a|_{\max} 8(2 + \epsilon)\|\hat{U} - C(C^TC)^{-1/2}\mathcal{O}\|_F^2. \tag{B.8}$$

Next we use the Davis-Kahan Theorem (Davis & Kahan, 1970; Stewart & Sun, 1990) that relates perturbation of matrices to perturbation of eigenspaces of those matrices. Then we have the following bound on the misclustering rate (also see Lemma 5.1 of Lei & Rinaldo (2015)):

$$r \leq 64(2 + \epsilon)\frac{|a|_{\max} k\|A - E[A]\|_2^2}{n(\lambda_{\min}(E[A])^2}. \tag{B.9}$$

Combining (B.7) and (B.9), we arrive at the desired result.

□

Next, we present the proofs of the theorems for the simplified special case with k equivalent communities.

Proof of Theorem 1

Proof. Under the simplified model we have

$$E[N] = C\left((\nu_1 - \nu_2)TI_k + \nu_2 T\mathbf{1}_k\mathbf{1}_k^T\right)C^T.$$

As before all communities have the same number of nodes, i.e., $|a| = \frac{n}{k}$ for all a, and $|a|_{\max} = \frac{n}{k}$. Then by Rohe et al. (2011), $\mathbf{1}_k$ is an eigenvector corresponding to the eigenvalue $\frac{n}{k}(\nu_1 - \nu_2)T + n\nu_2 T$, and the remaining non-zero eigenvalues are of the form $\frac{n}{k}(\nu_1 - \nu_2)T$. Since $n\nu_2 > 0$, the smallest non-zero eigenvalue

$$\lambda_{\min}(E[N]) = \frac{n}{k}(\nu_1 - \nu_2)T.$$

The upper bound from Theorem B.1 can also be simplified further under this model. We have

$$s = \sqrt{T}\sqrt{\frac{n}{k}\sigma_1^2 + \frac{(k-1)n}{k}\sigma_2^2} \asymp \sqrt{\frac{nT}{k}}\sqrt{\sigma_1^2 + (k-1)\sigma_2^2} \asymp \sqrt{nT}\sigma_1,$$

and

$$s_1 = \sqrt{T}\sigma_1,$$

and consequently,

$$(1+\epsilon)\left(2s + \frac{6}{\log(1+\epsilon)} s_1\sqrt{\log n}\right) + 2s_1\sqrt{\log n} \asymp \sqrt{T}\sigma_1\left(\sqrt{n} + \sqrt{\log n}\right)$$
$$\lesssim \sqrt{T}\sigma_1\sqrt{n}.$$

Substituting these quantities into Theorem B.1 completes the proof.

□

Proof of Theorem B.3

Proof. Under the simplified model all communities have the same number of nodes, i.e., $|a| = \frac{n}{k}$ for all a, and consequently $|a|_{\max} = \frac{n}{k}$. Further, we can write

$$E[A] = C\left((\exp(-\mu_2 T) - \exp(-\mu_1 T))I_k + (1 - \exp(-\mu_2 T)\mathbf{1}_k\mathbf{1}_k^T\right)C^T,$$

where I_k is the k-dimensional identity matrix, and $\mathbf{1}_k$ is the k-dimensional vector of all 1's. Then by Rohe et al. (2011), $\mathbf{1}_k$ is an eigenvector corresponding to the eigenvalue $\frac{n}{k}(\exp(-\mu_2 T) - \exp(-\mu_1 T)) + n(1 - \exp(-\mu_2 T))$, and the remaining non-zero eigenvalues are of the form $\frac{n}{k}(\exp(-\mu_2 T) - \exp(-\mu_1 T))$. Since $n(1 - \exp(-\mu_2 T)) > 0$, the smallest in absolute value non-zero eigenvalue of $E[A]$ is then,

$$\lambda_{\min}(E[A]) = \frac{n}{k}(\exp(-\mu_2 T) - \exp(-\mu_1 T)).$$

Also, under this setting, the numerator in the upper bound from Theorem B.2 becomes

$$\Delta = n(1 - \exp(-\mu_1 T)).$$

Substituting these quantities into Theorem B.2, we arrive at (B.3), the first statement of the theorem.

If we further assume that μT is small then we can make some further simplifications using the Taylor series expansion of $\exp(-x)$ near $x = 0$. In this case,

$$\lambda_{\min} \asymp \frac{n}{k}(\mu_1 - \mu_2)T,$$

and

$$\Delta \asymp n\mu_1 T.$$

Substituting these quantities into Theorem B.2, we arrive at (B.4), the second statement of the theorem, which completes the proof. □

B.2 Estimated Hawkes Process Parameters

B.2.1 Confidence Intervals

We derive confidence intervals for m using Theorem 2 and the following result readily obtained using the Law of Large numbers: $\bar{N}_{ab} \xrightarrow{p} \mu_{ab}$. A $(1-\theta)*100\%$ Bonferroni-corrected (due to multiple comparisons) simultaneous confidence interval for all k^2 parameters m_{ab} is

$$\hat{m}_{ab} \pm z_{(1-\frac{\theta}{2k^2})}\sqrt{\frac{1}{4n_{ab}\bar{N}_{ab}}}. \tag{B.10}$$

The confidence intervals on m_{ab} are particularly appealing to detect the "burstiness" of the network dynamics by testing the hypothesis $m_{ab} > 0$ for a block pair (a, b).

For the μ parameters, we are more interested in confidence intervals for pairwise differences between block pairs to identify whether the block pairs differ in their baseline event rates. Therefore, we build the following pairwise confidence intervals for all $2k(k-1)$ pairwise differences:

$$(\hat{\mu}_{ab} - \hat{\mu}_{ac}) \pm z_{(1-\frac{\theta}{4(k-1)k})} \frac{1}{T} \sqrt{\frac{9}{4}\left(\frac{\bar{N}_{ab}}{n_{ab}} + \frac{\bar{N}_{ac}}{n_{ac}}\right)}. \tag{B.11}$$

Note even though the random variables $\hat{m}_{ab}$ and $\hat{\mu}_{ab}$ are dependent across block pairs due to the spectral clustering step, the Bonferroni correction is still going to give a conservative (wide) interval with a simultaneous confidence coverage at least $1 - \theta$.

B.2.2 Proofs

Proof of Theorem 2

Proof. First, using the Central Limit Theorem and Law of Large Numbers, we have

$$\bar{N}_{ab} \xrightarrow{d} \mathcal{N}\left(\nu_{ab}, \frac{\sigma_{ab}^2}{n_{ab}}\right) \text{ and } S_{ab}^2 \xrightarrow{p} \sigma_{ab}^2, \qquad \text{as } n_{ab} \to \infty.$$

Then by Slutsky's theorem (Lehmann, 2004) we have,

$$\frac{\bar{N}_{ab}}{S_{ab}^2} \xrightarrow{d} \mathcal{N}\left(\frac{\nu_{ab}}{\sigma_{ab}^2}, \frac{1}{\sigma_{ab}^2 n_{ab}}\right) \Leftrightarrow \sqrt{n_{ab}}\left(\frac{\bar{N}_{ab}}{S_{ab}^2} - \frac{\nu_{ab}}{\sigma_{ab}^2}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\sigma_{ab}^2}\right).$$

Finally, we will apply the delta method (See Theorem 2.5.2 of Lehmann (2004)) on the random variable $X = \frac{\bar{N}_{ab}}{S_{ab}^2}$ with the function $g(x) = 1 - \sqrt{x}$. Note that $g'(x) = \frac{1}{2\sqrt{x}}$. Then we can compute $g'\left(\frac{\nu_{ab}}{\sigma_{ab}^2}\right) = \frac{\sigma_{ab}}{2\sqrt{\nu_{ab}}}$. Then we have

$$\sqrt{n_{ab}}\left(\hat{m}_{ab} - \left(1 - \sqrt{\frac{\nu_{ab}}{\sigma_{ab}^2}}\right)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4\nu_{ab}}\right).$$

Next we derive the asymptotic distribution for $\hat{\mu}_{ab}$. We first apply the delta method to the random variable $\bar{N}_{ab}$ with the function $g(x) = x^{3/2}$. Clearly, $g'(x) = \frac{3}{2}\sqrt{x}$, such that $g'(\nu_{ab}) = \frac{3}{2}\sqrt{\nu_{ab}}$. Then we have

$$\sqrt{n_{ab}}((\bar{N}_{ab})^{3/2} - (\nu_{ab})^{3/2}) \xrightarrow{d} \mathcal{N}\left(0, \frac{9}{4}\nu_{ab}\sigma_{ab}^2\right).$$

Applying Slutsky's theorem, we then have

$$\sqrt{n_{ab}}\left(\frac{(\bar{N}_{ab})^{3/2}}{S_{ab}} - \frac{(\nu_{ab})^{3/2})}{\sigma_{ab}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{9}{4}\nu_{ab}\right).$$

□

Proof of Theorem 3

Let $\bar{C}$ and $\hat{C}$ denote the true and estimated community assignment matrices respectively. Define $\bar{H} = \bar{C}(\bar{C}^T\bar{C})^{-1/2}$ and $\hat{H} = \hat{C}(\hat{C}^T\hat{C})^{-1/2}$, such that $\bar{H}^T\bar{H} = \hat{H}^T\hat{H} = I$.

We have

$$E[N] = \bar{C}\nu\bar{C}^T$$

Then

$$(\bar{C}^T\bar{C})^{1/2}\nu(\bar{C}^T\bar{C})^{1/2} = (\bar{C}^T\bar{C})^{-1/2}\bar{C}^T E[N]\bar{C}(\bar{C}^T\bar{C})^{-1/2} = \bar{H}^T E[N]\bar{H}.$$

Instead, the estimate for ν we get using estimated community assignment matrix $\hat{C}$ applied to N is

$$(\hat{C}^T\hat{C})^{1/2}\hat{\nu}(\hat{C}^T\hat{C})^{1/2} = \hat{H}^T N\hat{H}$$

Note that $(\hat{C}^T\hat{C})$ and $(\bar{C}^T\bar{C})$ are $k \times k$ diagonal matrices whose qth diagonal element represents the number of vertices that are part of the qth community. Next we make a key assumption—the sizes of the communities from the estimated community partition are similar to the true community sizes. In particular, we assume that the size of each of the k communities in the true and estimated partition is $O(\frac{n}{k})$. Therefore, the difference

$$(\hat{C}^T\hat{C})^{1/2}\hat{\nu}(\hat{C}^T\hat{C})^{1/2} - (\bar{C}^T\bar{C})^{1/2}\nu(\bar{C}^T\bar{C})^{1/2} \asymp \frac{n}{k}(\hat{\nu} - \bar{\nu}).$$

Now we have

$$\begin{aligned}
\frac{n}{k}(\hat{\nu} - \nu) &= \hat{H}^T N\hat{H} - \bar{H}^T E[N]\bar{H} \\
&= \hat{H}^T N\hat{H} - \hat{H}^T E[N]\hat{H} + \hat{H}^T E[N]\hat{H} - \bar{H}^T E[N]\bar{H} \\
&= \hat{H}^T(N - E[N])\hat{H} + \{\hat{H}^T E[N](\hat{H} - \bar{H}) + (\hat{H} - \bar{H})^T E[N]\bar{H}\}
\end{aligned}$$

We also note that

$$\|\bar{H}\|_2 \leq \sqrt{\lambda_{\max}(\bar{H}^T\bar{H})} = 1,$$

Note by assumption, $\sqrt{n_{ab}} \asymp \frac{n}{k}$. Now,

$$\begin{aligned}
\sum_{ab} n_{ab}(\hat{\nu} - \nu)_{ab}^2 &\asymp \left(\frac{n}{k}\right)^2 \|\hat{\nu} - \nu\|_F^2 \\
&\leq (\|\hat{H}^T(N - E[N])\hat{H}\|_F + 2\|\hat{H}^T E[N](\hat{H} - \bar{H})\|_F)^2 \\
&\leq 2k\left(\|N - E[N]\|_2^2 + 4\|E[N]\|_2^2 \frac{\|N - E[N]\|_2^2}{\lambda_{min}^2(N)}\right)
\end{aligned}$$

In the notation of Theorem 1,

$$\lambda_{\min}(N) = \frac{n}{k}(\nu_1 - \nu_2)T.$$

Also, using the upper bound in terms of expectation (instead of the in probability upper bound) from Theorem 1 we have

$$E[\|N - E[N]\|_2] \lesssim \sqrt{nT}\sigma_1, \quad \|E[N]\|_2 \lesssim n\nu_1 T.$$

Therefore,

$$E\left[\sum_{ab} n_{ab}(\hat{\nu} - \nu)_{ab}^2\right] \lesssim knT\sigma_1^2 + k\frac{n^2\nu_1^2T^2nT\sigma_1^2k^2}{n^2(\nu_1 - \nu_2)^2T^2} \lesssim knT\sigma_1^2 + \frac{k^3nT\sigma_1^2\nu_1^2}{(\nu_1 - \nu_2)^2}.$$

And consequently the sum of the weighted mean squared errors is,

$$\sum_{ab} n_{ab} E[(\hat{\nu} - \nu)_{ab}^2] \lesssim knT \max\left\{\sigma_1^2, \frac{k^2\sigma_1^2\nu_1^2}{(\nu_1 - \nu_2)^2}\right\}$$

Noting that $\sum_{ab} n_{ab} \asymp n^2$, the average MSE of estimating ν_{ab} is then asymptotically

$$\frac{kT}{n} \max\left\{\sigma_1^2, \frac{k^2\sigma_1^2\nu_1^2}{(\nu_1 - \nu_2)^2}\right\} \text{ or } \frac{T}{\sqrt{n_{ab}}} \max\left\{\sigma_1^2, \frac{k^2\sigma_1^2\nu_1^2}{(\nu_1 - \nu_2)^2}\right\}$$

For comparison, the weighted sum of MSEs in estimating ν_{ab}, using the estimator $\bar{N}_{ab}$ when the community structure is known (from Theorem 2) is

$$\sum_{ab} n_{ab} E[(\bar{N}_{ab} - \nu_{ab})^2] = \sum_{ab} \sigma_{ab}^2 = k^2 T \sigma_1^2,$$

and average MSE is asymptotically

$$\frac{k^2 T \sigma_1^2}{n^2} \text{ or } \frac{T\sigma_1^2}{n_{ab}}.$$

C Additional Experiments

We present two additional simulation experiments to analyze the effects of various parameters of the CHIP model on the accuracy of spectral clustering and to compare spectral clustering using weighted and unweighted adjacency matrices in detecting the ground truth community structure in simulated networks. We then present additional details and analyses for our real network dataset experiments.

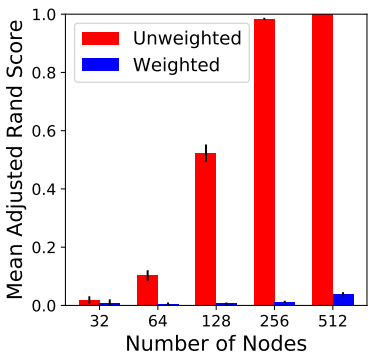
C.1 Simulation Experiments

C.1.1 Community Detection with Varying n

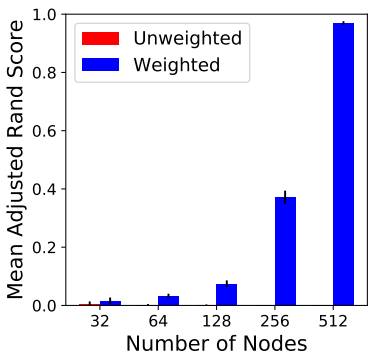
We simulate networks from the simplified CHIP model described in Section B.1.1 with $k = 4$ communities, duration $T = 400$, and a growing number of nodes n. We estimate community assignments of nodes using both the weighted adjacency (count) matrix N and unweighted adjacency matrix A.

First, we choose parameters $\mu_1 = 0.002$, $\mu_2 = 0.001$, $\alpha_1 = \alpha_2 = 7$, and $\beta_1 = \beta_2 = 8$ so that only μ is informative. The upper bound on the misclustering error rate using N is worse by a factor of $(1 - m)^{-1} = 8$ compared to using A as discussed in Section B.1.2. The adjusted Rand scores for spectral clustering on both A and N over 100 simulated networks for varying n are shown in Figure C.1(a). The accuracy on A approaches 1 for growing n, as expected. The accuracy on N is significantly worse, as predicted by the comparison of the respective upper bounds on the misclustering error rates, and no better than a random community assignment until $n = 512$ nodes.

Next, we choose parameters $\mu_1 = \mu_2 = 0.001$, $\alpha_1 = 0.006$, $\alpha_2 = 0.001$, and $\beta_1 = \beta_2 = 0.008$. so that only α is informative. The error for A is unbounded, while the error for N still follows the upper bound in Corollary 1. As shown in Figure C.1(b), the accuracy on N approaches 1 as n increases, while the accuracy on A is no better than random even for growing n, as expected.



(a) Only μ is informative ($\mu_1 \neq \mu_2$ and $\alpha_1 = \alpha_2$)



(b) Only α is informative ($\mu_1 = \mu_2$ and $\alpha_1 \neq \alpha_2$)

Figure C.1: Mean adjusted Rand scores of spectral clustering on weighted and unweighted adjacency matrices over 100 simulated networks ($\pm$ 2 standard errors). $\beta_1 = \beta_2$ in both cases. Both sets of results agree with upper bounds from Section B.1.2.

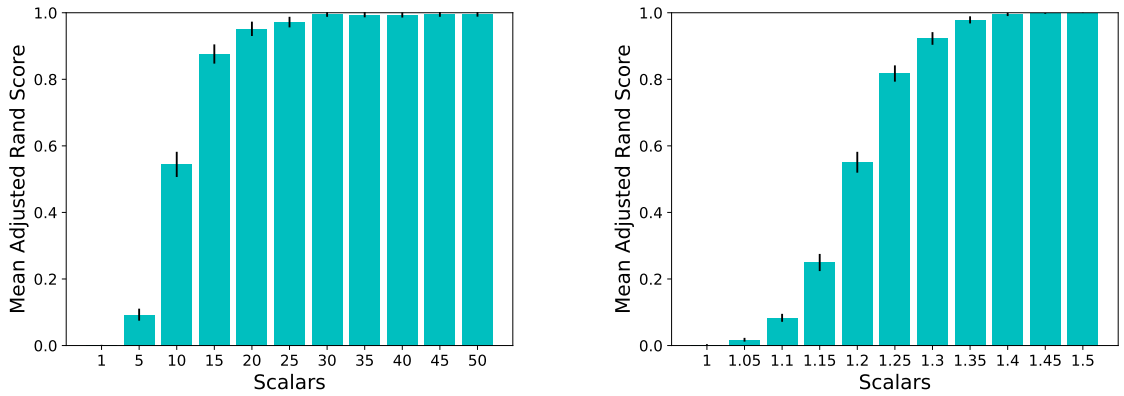
C.1.2 Effects of Diagonal and Off-diagonal μ's on Community Detection

In Section C.1.1 we observed that community detection will be easier if μ is informative ($\mu_1 \neq \mu_2$). In this experiment, we will explore two different ways of encoding community information into simulated networks by

1. Scaling up both μ_1 and μ_2, while keeping a fixed $\mu_1 : \mu_2$ ratio.
2. Only scaling up μ_1, allowing for $\mu_1 : \mu_2$ ratio to increase.

Both settings share the same base parameters of $\mu_1 = 0.075$ and $\mu_2 = 0.065$, with $k = 4$ communities and $n = 128$ nodes, a duration of $T = 50$, where $\alpha_1 = \alpha_2 = 0.05$ and $\beta_1 = \beta_2 = 0.08$. These parameters are chosen to create a base network that is nearly impossible for spectral clustering to accurately detect communities. The objective is similar to that of Section C.1.1, where in both settings we perform community detection using spectral clustering on the weighted adjacency of simulated networks, while increasing μ_1 and μ_2 or their ratio. Lastly, we average over the adjusted Rand score of 100 simulations.

As shown in Figure C.2, community detection accuracy increases in both settings as the scalars increase; however, we find that the increase in accuracy occurs for different reasons. In the first setting, Figure C.2(a), where both μ's are scaled up with a fixed ratio, community detection becomes easier simply because the networks are becoming denser, as shown in the numbers above the bars in Figure C.3, and more information is available. Furthermore, although we keep the $\mu_1 : \mu_2$ ratio fixed, as the scalars increase the difference between the two starts to magnify. On the other hand, as networks become denser and most node pairs start to have at least one interaction, it is only the number of interactions among node pairs that becomes informative. Therefore, spectral clustering on the weighted adjacency matrix continues to result in a high adjusted Rand score, while the adjusted Rand score of spectral clustering on the unweighted adjacency matrix decreases with increasing density, as it is illustrated in Figure C.3. This observation confirms the theoretical prediction made in Theorem B.3. The opposite also holds to some degree. For really sparse networks spectral clustering is more accurate on the unweighted adjacency matrix; however, in Figure C.3 we observe that it loses its advantage as the proportion of node pairs with at least one interaction approaches 0.5 and starts impairing community detection as it passed 0.8.



(a) Scaling both μ's up, with a fixed $\mu_1 : \mu_2$ ratio

(b) Scaling μ_1 up only, keeping μ_2 fixed

Figure C.2: Adjusted Rand score of spectral clustering on weighted adjacency matrix, averaged over 100 simulated networks ($\pm$ 2 standard errors), while multiplying μ_1 and μ_2 or their ratio by scalars. C.2(a) Scaling up both μ_1 and μ_2, keeping their ratio fixed. C.2(b) Only scaling up μ_1, while keeping μ_2 fixed.

In the second setting, Figure C.2(b), by only scaling up μ_1, the difference between the baseline rate of occurrence of an event between the diagonal and the off-diagonal blocks increases. This increases the signal-to-noise ratio and is a more effective way of encoding community information into a network. This can be observed by comparing the scalars of Figures C.2(a) and C.2(b). Starting from the same base network, a perfect adjusted Rand score is achieved when only μ_1 is scaled up by a factor of 1.4, compared to scaling both μ's up by a factor of 30.

C.2 Real Data

C.2.1 Dataset Descriptions

We consider three real network datasets consisting of timestamped relational events. For each dataset, we normalize the event times to the range $[0, 1{,}000]$.

- MIT Reality Mining (Eagle et al., 2009): Consists of 2,161 phone calls where the start time of each call was used as the event timestamp. This dataset has a "core-periphery" structure, where there is a core group for whom we have all of their communication data and a much larger group of people in the periphery who had contact with the core. We consider calls between pairs of the core 70 callers and recipients. We use the last 661 phone calls as the test set[1].

- Enron (Klimt & Yang, 2004): Consists of 4,000 emails exchanged among 142 individuals. We use the last 1,000 emails as the test set.

- Facebook Wall Posts (Viswanath et al., 2009): Consists of a total of 876,993 wall posts from 46,952 users from September 2004 to January 2009. We consider only posts from a user to

[1]We found some inconsistencies between the actual dataset used and its description in DuBois et al. (2013). For a fair comparison, we loaded and preprocessed this dataset using their code available on GitHub: https://github.com/doobwa/blockrem/blob/master/process/reality.r.

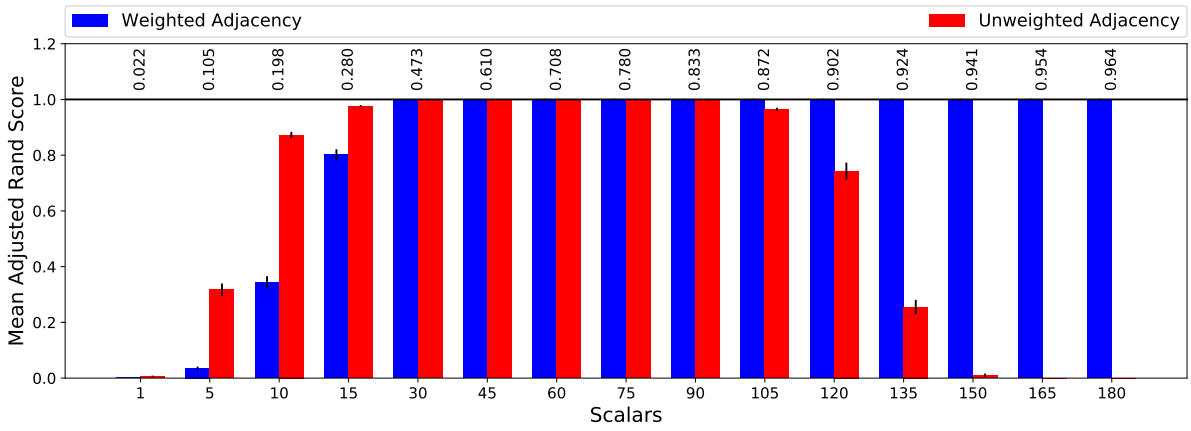


Figure C.3: Adjusted Rand score of spectral clustering on weighted vs. unweighted adjacency matrices, averaged over 100 simulations ($\pm$ 2 standard errors), while multiplying both μ_1 and μ_2 by scalars. The numbers above each bar indicate the average density of simulated networks as the proportion of non-zero entries to the total number of elements in the adjacency matrix. Base model parameters are: $\mu_1 = 7.5 \times 10^{-4}$, $\mu_2 = 3.5 \times 10^{-4}$, $k = 4$, $T = 50$, $n = 256$, $\alpha_1 = \alpha_2 = 0.05$, and $\beta_1 = \beta_2 = 0.08$.

another user's wall so that there are no self-edges. We analyze the largest connected component of the network excluding self loops: 43,953 nodes and 852,833 events. We divide the dataset into train and test sets using a 80%/20% split on the number of events.

C.2.2 Comparison with Other Models

We find that our proposed CHIP model achieves higher test log-likelihood than the relational event model (REM) (DuBois et al., 2013), block Hawkes model (BHM) (Junuthula et al., 2019), and the spectral clustering with homogeneous Poisson process baseline on the Reality Mining and Enron datasets as shown in Table 1. CHIP and the Poisson baseline were able to scale to the Facebook network, which was two orders of magnitude larger. The local search procedure in the BHM does not scale to such a large network, so we provide fits using only the faster but less accurate spectral clustering procedure. We did not implement the REM so we compare against the reported results in DuBois et al. (2013), which did not include the Facebook data. We note that since all the three models assume the same Poisson process for arrival of events with different rates (which are governed by different set of parameters), the joint distribution of event times has the same form for all three models. Hence the likelihood function of the models are directly comparable. Therefore, the test log-likelihood is a reasonable metric for comparing the fits of the models to the data.

To compute the test log-likelihood for CHIP and BHM, we use the following process. First, we use the estimation procedure explained in Section 3.2 to estimate all CHIP's Hawkes process parameters using the training set (the entire dataset excluding the test set). Next, we calculate the model log-likelihood on the entire dataset and subtract the training log-likelihood from it. The result is then divided by the total number of events in the test set to evaluate the mean log-likelihood per test event, which is the metric used in DuBois et al. (2013). Lastly, if a node in the test set did not appear in the training data, it was automatically assigned to the largest block.

We implemented the BHM by using spectral clustering followed by local search (Junuthula et al., 2019), which they found to achieve the highest adjusted Rand score in simulations compared to just spectral clustering and variational EM. We allowed the local search to converge to a local maximum

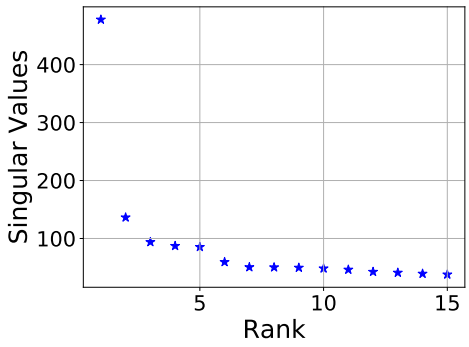


Figure C.4: 15 largest singular values of spectral clustering on the weighted adjacency matrix of the Enron dataset. The gap between the 2nd and the 3rd largest singular values led us to select $k = 2$ blocks.

Table C.1: Number of node pairs and events in each block pair of the CHIP model with $k = 2$ in the Enron dataset.

Block Pair (a, b)	(1, 1)	(1, 2)	(2, 1)	(2, 2)
Node Pair Count	5700	5016	5016	4290
Event Count	965	572	1038	1425

for all values of k.

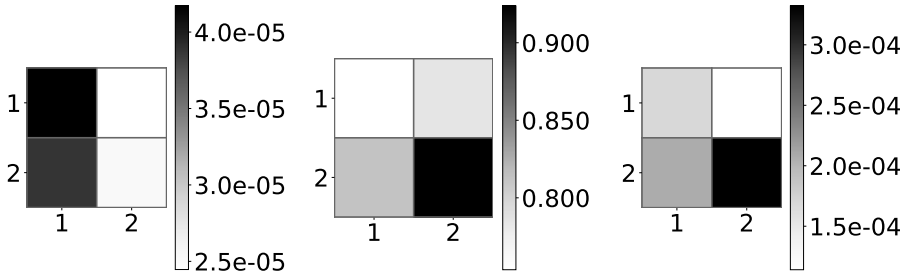
C.2.3 Exploratory Analysis of Enron Network

Next, we perform model-based exploratory analysis of the Enron network using CHIP. We find a large gap between the 2nd and the 3rd largest singular values of the weighted adjacency matrix as shown in Figure C.4 so we choose a fit with $k = 2$ blocks. The number of node pairs and events in each block pair are shown in Table C.1.

Figure C.5(a) shows the estimated baseline intensity of each block pair. This can be thought of as the rate at which email conversations get started. We observe that $\hat{\mu}_{11}$ is much larger than $\hat{\mu}_{22}$; however block pair $(1, 1)$ only accounts for 965 emails as opposed to 1,425 for block pair $(2, 2)$. Thus, the community structure is not evident only from the differences in the baseline rates μ.

It is only after we consider how bursty interactions are in each block pair, as shown in Figure C.5(b), that we can explain the dynamics of this network. In particular, $\hat{m}_{22}$ is much higher than $\hat{m}_{11}$. In other words, once an email conversation is started in block-pair $(2, 2)$ we can expect more emails to follow, as opposed to more frequent conversations starting in $(1, 1)$, but with less follow-ups. Hence, the combination of μ and m allows us to observe the community structure, with more edges within block pairs than between, as shown by the values of $\hat{\mu}/(1 - \hat{m})$ in Figure C.5(c).

Table C.2 shows the numerical values for $\hat{m}$ along with their 95% confidence intervals obtained using (B.10), indicating that all block pairs exhibit highly bursty behavior. As previously mentioned, the baseline rates $\hat{\mu}$ are not by themselves indicative of the community structure due to the burstiness of events in all of the blocks. Indeed, when we examine the 95% confidence intervals for pairwise differences between the μ values for different block pairs using (B.11) shown in Table C.3, all of the confidence intervals include 0.



(a) $\hat{\mu}$ (b) $\hat{m}$ (c) $\hat{\mu}/(1-\hat{m})$

Figure C.5: Estimated CHIP parameters on Enron data, where axis labels of each heatmap denote block index. Each tile corresponds to a block pair where (a, b) denotes row a and column b.

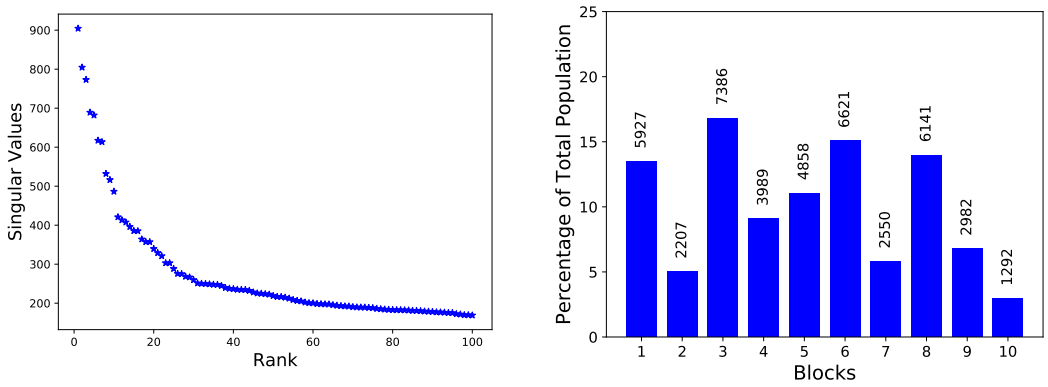
Table C.2: Estimated $\hat{m}_{ab} \pm 95\%$ confidence interval from CHIP on the Enron dataset with $k = 2$. All values of $\hat{m}_{ab}$ are statistically significant at the 5% level for the test $m_{ab} > 0$. The high values of $\hat{m}_{ab}$ indicate that interactions in all block pairs are quite bursty.

Block Pair	1	2
1	0.7536 ± 0.0440	0.7855 ± 0.0572
2	0.8126 ± 0.0424	0.9237 ± 0.0362

Table C.3: Pairwise difference for unique pairs of diagonal vs. off-diagonal $\hat{\mu}_{a,a} - \hat{\mu}_{a,b} \pm 95\%$ confidence interval of the CHIP model fitted to the Enron dataset with $k = 2$. None of the differences are statistically significant at the 5% level for the test $\hat{\mu}_{a,a} - \hat{\mu}_{a,b} \neq 0$, suggesting that the community structure is not evident from differences in the baseline rates μ.

Pairwise Differences in $\hat{\mu}$

$\hat{\mu}_{1,1} - \hat{\mu}_{1,2}$	$1.724 \times 10^{-5} \pm 2.970 \times 10^{-5}$
$\hat{\mu}_{1,1} - \hat{\mu}_{2,1}$	$2.929 \times 10^{-6} \pm 3.455 \times 10^{-5}$
$\hat{\mu}_{2,2} - \hat{\mu}_{1,2}$	$8.650 \times 10^{-7} \pm 4.105 \times 10^{-5}$
$\hat{\mu}_{2,2} - \hat{\mu}_{2,1}$	$-1.345 \times 10^{-5} \pm 4.468 \times 10^{-5}$



(a) Singular values

(b) Block sizes

Figure C.6: Results of spectral clustering on the weighted adjacency matrix of the largest connected component of the Facebook Wall Posts dataset. C.6(a) 100 largest singular values. There is a large gap between the 10^{th} and the 11^{th} largest singular values that leads us to select $k = 10$ blocks. C.6(b) Size of each formed block. Numbers on top of each bar indicate the actual number of nodes in that block.

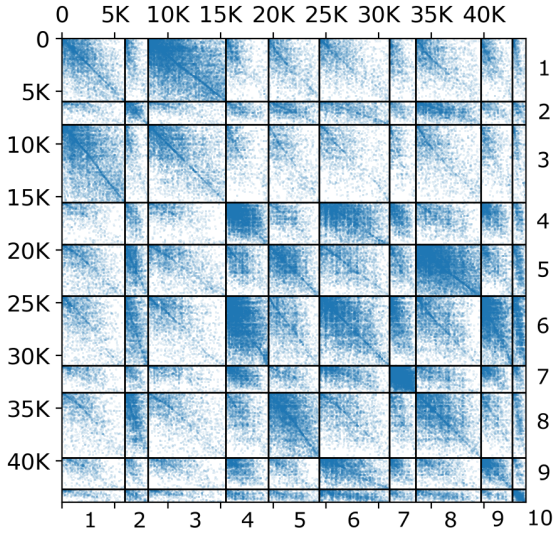
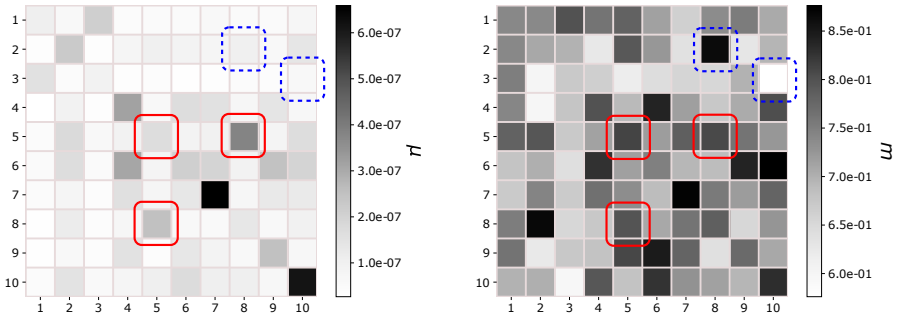


Figure C.7: Adjacency matrix for Facebook wall post network with rows and columns rearranged to show block structure.

C.2.4 Exploratory Analysis of Facebook Wall Post Network

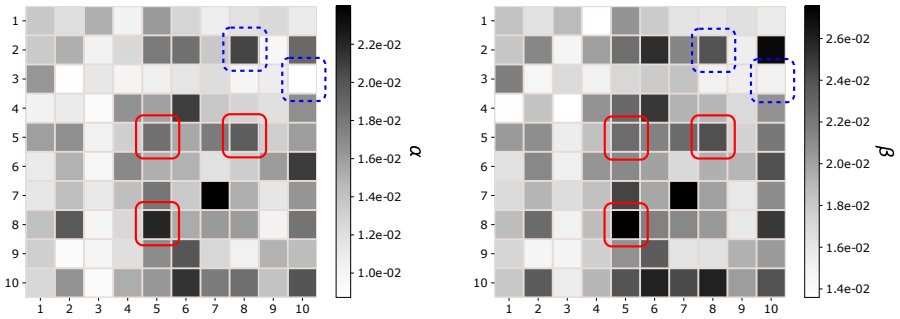
Fitting the CHIP model to the largest connected component of the network (excluding self loops) consisting of 43,953 nodes and 852,833 edges required only 141.4 s. Considering the gap between the 10^{th} and the 11^{th} largest singular values of the weighted adjacency matrix of the network as shown in Figure C.6(a), we choose a model with $k = 10$ blocks, resulting in the block sizes depicted in Figure C.6(b).

Figure C.8 shows heatmaps of the fitted CHIP parameters. Although diagonal block pairs have a higher base intensity on average, indicating an underlying assortative community structure, there



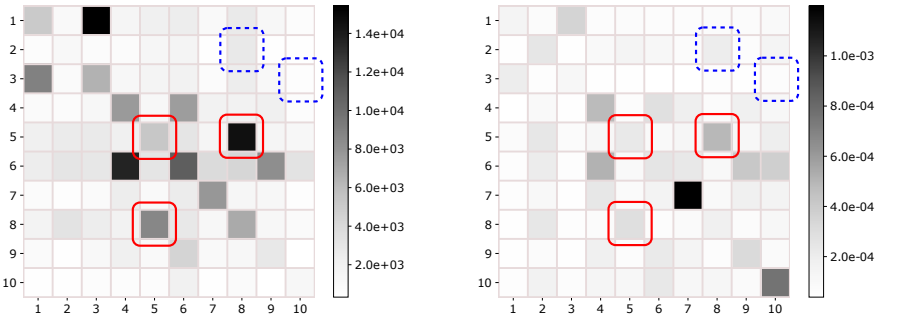
(a) μ: base intensity

(b) m: α to β ratio



(c) α: intensity jump size

(d) β: intensity decay rate



(e) Total number of events

(f) Mean number of events per node pair

Figure C.8: Inferred CHIP parameters on the largest connected component of the Facebook Wall Posts dataset with $k = 10$. Axis labels denote block numbers. Each tile corresponds to a block pair where (a, b) denotes row a and column b. Boxed block pairs in the heatmap are discussed in the body text.

are some off-diagonal block pairs with a high μ such as $(5, 8)$ and $(8, 5)$, as shown in the red boxes in Figure C.8. This illustrates that the CHIP model does not discourage inter-block events. These patterns often occur in social networks, for instance, if there are communities with opposite views on a particular subject.

While the structure of μ reveals insights on the baseline rates of events between block pairs, the structure of α (Figure C.8(c)) and β (Figure C.8(d)) reveal insights on the burstiness of events between block pairs. Note that the structure of the α to β ratio m (Figure 4(c)) affects the asymptotic mean number of events in (3). For some block pairs, such as $(3, 10)$, there are very low values of α and β indicating the events are closely approximated by a homogeneous Poisson process. There are some block pairs, such as $(2, 8)$ that have a low baseline rate of events but are extremely bursty, which relatively increases the mean number of events per node pair. Both of these block pairs are shown in the blue dashed boxes in Figure C.8. The different levels of burstiness of block pairs cannot be seen from aggregate statistics such as the total number of events (Figure C.8(e)) or even the mean number of events per node pair (Figure C.8(f)).

Unlike the findings of Junuthula et al. (2019), who studied only a subset of the network containing $3,582$ nodes using $k = 2$ blocks, we find that α is not necessarily higher for diagonal blocks as shown in Figure C.8(c). Additionally, even though we do not explicitly model reciprocity between node pairs in our CHIP model, we can nevertheless empirically observe certain reciprocities through the patterns of the estimated α and β parameters. We note that the high reciprocity present in social networks is captured by CHIP through the symmetry in all Hawkes process parameters about block pairs. This can be observed in block pairs $(8, 5)$ and $(5, 8)$. In the context of this dataset, a symmetric α and β corresponds to the notion that wall posts posted by the people in block 5 on the wall of the people in block 8 will urge people in block 8 to respond, which in turn promotes more wall posts by people in block 5.

Lastly, it is worth noting that fitting the CHIP model to this data set using the unweighted adjacency matrix resulted in a per event log-likelihood of -10.04 compared to -9.61 for the weighted adjacency matrix on the test data set when using a 80%/20% train and test split on the events. Thus, this was another reason to use the weighted adjacency matrix besides its aforementioned advantages in previous sections. We note that running spectral clustering on the unweighted adjacency matrix, compared to the weighted adjacency matrix, seemed to detect communities with larger number of intra-block events, while inter-block events were a lot less common.

References

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

Da Fonseca, J. and Zaatour, R. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6):548–579, 2014.

Davis, C. and Kahan, W. M. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Daw, A. and Pender, J. Queues driven by Hawkes processes. *Stochastic Systems*, 8(3):192–229, 2018.

DuBois, C., Butts, C. T., and Smyth, P. Stochastic blockmodeling of relational event dynamics. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pp. 238–246, 2013.

Eagle, N., Pentland, A. S., and Lazer, D. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

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

Junuthula, R. R., Haghdan, M., Xu, K. S., and Devabhaktuni, V. K. The Block Point Process Model for continuous-time event-based dynamic networks. In *Proceedings of the World Wide Web Conference*, pp. 829–839, 2019.

Klimt, B. and Yang, Y. The Enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning*, pp. 217–226. Springer, 2004.

Laub, P. J., Taimre, T., and Pollett, P. K. Hawkes processes. *arXiv preprint arXiv:1507.02822*, 2015.

Lehmann, E. L. *Elements of large-sample theory*. Springer Science & Business Media, 2004.

Lei, J. and Rinaldo, A. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

Nocedal, J. and Wright, S. *Numerical optimization*. Springer Science & Business Media, 2006.

Ozaki, T. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979.

Pensky, M. and Zhang, T. Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, 13(1):678–709, 2019.

Rohe, K., Chatterjee, S., and Yu, B. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Stewart, G. W. and Sun, J.-G. *Matrix perturbation theory*. Academic Press, Boston, MA., 1990.

Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, pp. 37–42, 2009.