Multivariate Hawkes Processes for Incomplete Biased Data

Zihan Zhou
Division of Computer Science and Engineering
Louisiana State University
Baton Rouge, Louisiana
zzhou23@lsu.edu

Mingxuan Sun
Division of Computer Science and Engineering
Louisiana State University
Baton Rouge, Louisiana
msun11@lsu.edu

***Abstract*—Multivariate Hawkes processes have been widely used in many applications such as crime detection and disaster rescue forecast to model events that exhibit self-exciting properties. One of the biggest challenges is that data collected from real world is usually incomplete, and even biased. The training of a machine learning model using such data can introduce biased predictions. For example, event hotspot predictions using biased data can make the visibility of minority groups (e.g., communities of racial minorities) more apparent. While there have been some explorations in developing Hawkes processes for event data, none of those methods deals with incomplete biased data where events of certain markers (e.g., events reported from racial minorities) may be missing or heavily underrepresented. In this paper, we propose a novel Multivariate Hawkes model to tackle the incomplete biased data challenge. First, we assume that there is possibility that events can be missing between any two observed events and we define a novel likelihood function integrating missing window probabilities. A Markov Chain Monte Carlo (MCMC) sampling framework is used to generate virtual event data probabilistically in missing windows. Second, we propose to incorporate event marker features such as geographic information to regularize the infectivity kernel matrix between markers. In such a way, the MCMC sampler is encouraged to generate more virtual events with markers that are biased. Both observed and virtual events will contribute to the model estimation through maximizing the log-likelihood. We carry on experiments over several real-world datasets, and our model improves prediction accuracy in comparison with the state-of-arts.**

***Index Terms*—Multivariate point process, Hawkes processes, incomplete biased data.**

I. Introduction

Temporal event sequences associated with different markers (e.g., location indices, patient ids) are observed in various applications such as criminology, epidemiology, and social-network studies. For example, rescue request events with time-stamps and geo-tags are collected from social media such as Twitter, which can be used for rapid flood mapping, damage assessment, and situation awareness. Similarly, burglary event data collect the time and location of each reported event. In epidemiology, the time stamps along with the patient ids who are infected by a certain disease are recorded. Such time series usually exhibit mutually exciting patterns of diffusion. For example, the occurrence of a rescue event is likely to increase the likelihood of another in nearby space and time. Similarly, patients infected by an infectious disease might also spread the disease to their neighbors. We are interested in discovering correlations among events and predicting future events, e.g., answering questions such as when an event is likely to happen and where it will take place.

Multivariate Hawkes processes (MHP) are widely used for capturing mutually exciting patterns between event sequences. Specifically, Multivariate Hawkes processes assume that an event triggers a set of subsequent offspring events correlated in time and markers. Temporal kernels are used to model self-exciting effects. An infectivity kernel matrix is used in multivariate Hawkes to specify the correlation between different markers. A larger similarity between two markers indicates that events from those two markers can excite each other more. Multivariate Hawkes processes have been used to model many different applications such as predictive finance [1], predictive crimes [2]–[4], and predictive online user behaviors [5]–[9].

One of the biggest challenges for event modeling is that data collected from real word is usually incomplete, and even biased. For example, in rescue event predictions, it has been reported that higher disaster-related Twitter-use communities tend to be of higher socioeconomic status [10]. Some regions of lower socioeconomic status may be severely flooded. However, Twitter rescue data collected from these regions might be heavily under-represented. It is known that MHPs suffer from the amount of training data, and the excitation patterns learned by MHPs from incomplete sequences can be unreliable [11]. Moreover, MHPs trained on incomplete and biased data can lead to inaccurate and unfair predictions. For example, by using a Hawkes model trained on biased Twitter data, the regions that are at risks but have fewer Twitter events may be predicted as lower risk regions.

While there have been some explorations in enhancing Hawkes processes for incomplete data, most of them ignore the data bias and assume events are missing in certain time windows regardless of event markers. For example, a data synthesis method [11] is proposed to enhance Hawkes processes for double-censored data, which assumes events before or after the observed time are missing. A recent attempt [4] assumes that events from hidden nodes (e.g., unknown markers) are

978-1-6654-3902-2/21/$31.00 ©2021 IEEE

unobserved during a given time period, which can excite or be excited by events of observed markers. However, none of those methods deals with incomplete biased data, where events associated with any marker can be missing in any time window and events of certain markers may be heavily underrepresented.

In this paper, we proposed a novel method to learn multivariate Hawkes processes from incomplete biased data. First, we relax the assumption in the existing work that data are only missing in a certain time window. Instead, we assume that there is possibility that events can be missing between any two observed events. We define a missing probability for each interval, which depends on the time elapsed between two observed events and the feature associated with each marker. We proposed a novel likelihood function integrating missing window probabilities. A Markov Chain Monte Carlo (MCMC) sampling framework is used to generate virtual event data probabilistically in missing windows. Second, to mitigate the biased problem where events of some markers are more incomplete than others, we propose to incorporate marker features such as geographic information to regularize the infectivity kernel matrix between markers. In such a way, the MCMC sampler is encouraged to generate more virtual events with markers that are biased. In summary, both observed and virtual events will contribute to the model estimation through maximizing the log-likelihood.

II. Related Work

Hawkes processes, which are capable of modeling self-excited event sequences, have been widely used in various applications including earthquake prediction [12], predictive policing [13], and hazard rate prediction [3], [14]. In comparison with traditional spatio-temporal modeling approaches, Hawkes processes show better prediction accuracy for predicting event hazard rates and ranking event hotspots [15]. Some neural temporal point processes such as recurrent marked temporal point process (RMTTP) [8] and neurally self-modulating multivariate point process (N-SM-MPP) [16] are proposed to model event sequences. Recently, a particle smoothing method by a bidirectional continuous-time LSTM is proposed in [17] to impute missing events for neural temporal point processes. However, these models do not explicitly recover the excitation influences between markers.

We focus on improving Multivariate Hawkes processes, which explicitly recover mutually exciting patterns between different markers. There are some existing methods for improving multivariate Hawkes with missing data. Tucker et al. [18] use a Bayesian estimation procedure to get the conditional probability of missing data, given observed data. A stitching method is proposed in [11] to deal with the double-censored type of missing data. It assumes sequences of events of any type can be missing either at beginning or at the end, or at both ends. This type of missing data can happen in the sequences of an individual's disease history during the lifetime. However, it is a very specific case of the missing patterns. Recent work [4] assumes there are events from unknown markers unobserved during a given time period, which can excite or can be excited by events of observed markers. Our assumption is different with that in the stitching method [11] assuming data can be missing only at the beginning and end of the observed sequence. Our assumption is also different with that in [4] assuming data are complete on observed nodes (markers) and all the missing events are on hidden nodes (markers). Also none of the aforementioned methods deals with incomplete and biased data.

Standard re-sampling approaches such as oversampling and undersampling are used to reweight data based on certain classes, which are mostly for classification and regression tasks. Existing methods such as bootstrapping [19] have been used for time series to improve learning results when observations are imperfect. For point processes, Mariana Oliveira et al. [20] propose to use a sampling weight to represent the probability that a certain event will be selected to stay or be replicated in the training set. This weight is a linear combination of spatial and temporal information of an event. This is a different re-sampling and learning strategy from ours. First, they re-balance the training data before the model learning stage. However, we propose a new Hawkes model that can oversample events in underrepresented groups during the model learning stage. Second, they use a weight to decide whether or not an original event should be replicated or deleted. We do not modify original data; instead we encourage the sampler to generate more events of rare markers based on Hawkes process parameters such as the kernel matrix and base rates.

III. Background

A. Multivariate Hawkes Process

Multivariate Hawkes process is used to model sequences of mutually exciting events associated with different markers such as locations where events happened. A collection of events with markers and time stamps during a time window $[0, T)$ can be represented as a sequence $x = \{(t_1, l_1), (t_2, l_2), \cdots, (t_n, l_n)\}$, where $0 < t_{i-1} < t_i < T$, n is the total number of events, and l_i is the marker of the event at time t_i. In many event prediction applications such as the analysis of crime reports or disaster-related rescue requests, an area (e.g., a city) is discretized into square grid cells, geographic block regions, or political boundaries (e.g., ZIP codes). Let L be the number of distinct markers and l_i the index to the grid cell or zip codes, that is $l_i \subset \{1, 2, .., L\}$.

Multivariate Hawkes process can be characterized by its conditional intensity function $\lambda_l(t)$, which is the expected instantaneous rate of the event of marker l at time t given the history of all the previous events up to time t [21]. That is $\lambda_l(t) = \lim_{\Delta t \to 0}(E[N_l(t, t + \Delta t)|H_t]/(\Delta t))]/dt$, where $N_l(t, t+\Delta t)$ is the count of type l events during time $[t, t+\Delta t]$, and $H_t = \{(t_i, l_i)|t_i < t, l_i \in L\}$ is the set of historical events up to time t. The probability of a sequence x is given by

$$p(x) = \exp(-\sum_l \int_0^T \lambda_l(s, h_s)ds) \prod_{i=1}^n \lambda_{l_i}(t_i, h_{t_i}). \qquad (1)$$

Specifically, in a linear multivariate Hawkes process, the conditional intensity takes the form

$$\lambda_l(t, h_t) = \mu_l + \sum_{i:t_i<t} \phi_{l_i,l}(t - t_i), \tag{2}$$

where μ_l is the base rate of events of marker l, and $\phi_{l_i,l}$ is the kernel function, which captures the increase in the rate of marker l triggered by the event of marker l_i that occurs $(t-t_i)$ time units ago. We can simplify (2) by introducing a special root event (t_0, l_0)=$(0, 0)$. Let the kernel for this new event be $\phi_{l,0} = 0, \forall l\ \phi_{0,l} = \mu_l, \forall l > 0$. This event "causes" the base rate of events for each marker. Let $\mathcal{I}_t^0 = \{i|t_i < t\} \cup \{0\}$. The intensity rate can be simplified as:

$$\lambda_l(t, h_t) = \sum_{i\in\mathcal{I}_t^0} \phi_{l_i,l}(t - t_i). \tag{3}$$

The kernel $\phi_{l_i,l}$ can be decomposed into two components: an $L \times L$ non-negative matrix M and a base kernel $\phi(t)$. That is:

$$\phi_{l_i,l} = M_{l_i,l}\phi(t). \tag{4}$$

The entry $M_{l_i,l}$ in matrix M quantifies the parent-offspring excitation rate at any time. The base kernel $\phi(t)$ captures the temporal decay. Some common choices of base kernel include exponential $\phi(t) = \exp(-\beta t)$ with $\beta > 0$, and power law $\phi(t) = (t + \gamma)^{-(1+\beta)}$ with $\beta > 0$ and $\gamma > 0$. We use the exponential kernel in our model.

We denote $\Phi_{l_i,l}(t) = \int_0^t \phi_{l_i,l}(s)ds$ and $\Phi_{l_i,\star}(t) = \sum_l \phi_{l_i,l}(t)$. Equations (1) and (3) can be combined and the likelihood of sequence x is

$$p(x) = \exp(-\sum_{i:t_i<t} \Phi_{l_i,\star}(T - t_i)) \prod_{i=1}^{n} \sum_{j:t_j<t_i} \phi_{l_j,l_i}(t_i - t_j). \tag{5}$$

Given the observed historical event sequences, the model parameters can be estimated by maximizing the joint log-likelihood.

B. Sampling Method

A Hawkes process can be viewed as generation processes of parent events and offspring events, which forms a latent branching structure. The first generation of parents follows an inhomogeneous Poisson process. Each parent event generates a set of children events independently. At time t, the rate of an event with marker l is the sum of the rates of any previous event generating a child with marker l at time t. That is, each parent (t_i, l_i) generates off-springs (t, l) with intensity λ_l according to (2).

An unconditional sampler generates events with markers (e.g., 1, 2, ..., L) recursively. It starts with a root event of marker 0 at time 0, and samples "children" events from base rates until time T. Each of these children recursively generates its own "children" events based on the kernel function.

A recent study [4] builds a more effective Metropolis-Hastings sampler based on the unconditional sampler with two sets of auxiliary variables: "parent structure" and "virtual events". Firstly, denote parent auxiliary variable as $a = \{a_1, \cdots, a_n\}$, where a_i is the index of the parent of event with index i. These variables indicate the "parent-children" relationship in the branching structure of Hawkes process. The joint distribution of the sequence x and the variable a is

$$p(x, a) = \prod_{i=1}^{n} \phi_{l_{ai},l_i}(t_i - t_{ai}) \exp(-\Phi_{l_i,\star}(T - t_i)). \tag{6}$$

It is obvious that the equation $\sum_a p(x, a) = p(x)$ (from (5)) holds.

The second set of auxiliary variables consist of a set of virtual events, which are children sampled from observed and root events and unable to generate their own children events. They can be represented by $\tilde{x} = \{(\tilde{t}_1, \tilde{l}_1), (\tilde{t}_2, \tilde{l}_2), \ldots, (\tilde{t}_{\tilde{n}}, \tilde{l}_{\tilde{n}})\}$. Let $\tilde{a}_i$ denote the index of the parent (a real event) of the i^{th} virtual event. The complete joint distribution over all auxiliary variables is

$$\begin{aligned} p(x, a, \tilde{x}, \tilde{a}) = &\prod_{i=1}^{n} \exp[-(\kappa + 1)\Phi_{l_i,\star(T-t_i)}] \\ &\times \prod_{i=1}^{n} \phi_{l_{ai},l_i}(t_i - t_{ai}) \prod_{i=1}^{\tilde{n}} \kappa \cdot \Phi_{l_{\tilde{a}_i},\tilde{l}_i}(\tilde{t}_i - t_{\tilde{a}_i})), \end{aligned} \tag{7}$$

where κ is a parameter, which determines the rate of generating a virtual event.

The MCMC sampler maintains a state of $\{x, a, \tilde{x}, \tilde{a}\}$. Among all types of events, the root event (marker 0), observed events, and sampled events are members of $\{x\}$. Virtual events are members of $\{\tilde{x}\}$. At each iteration, an event is randomly picked from the set $x \cup \tilde{x}$ and the states of the sampler are updated by three moves. The first move is named "virtual children". If the picked event is from x, it can generate virtual children and increase the pool of $\tilde{x}$. The second move is called "virtualness". If the picked event is from $\tilde{x}$, it will change the event from state $\tilde{x}$ to x, and call these events sampled events. Sampled events can generate its own "children events" in later interactions. The third move is "parent", which updates parent-child relationship of non-root event from x to maximize the joint likelihood. Both sets of auxiliary variables contribute to the estimation of Hawkes processes with unobserved events.

IV. Our Proposed Approach for Incomplete and Biased Data

We tackle the incomplete and biased data challenge from two perspectives. First, we assume that there is possibility that events can be missing between any two consecutive observed events. The possibility depends on the previous history, the time elapse between the two observed events, and the association between their markers. For example, if the estimated probability of observing an event in a time window is high but there are no observations, the probability of some events missing in the window is high. The proposed model with missing window probability is defined in Section IV-A. Second, to mitigate the bias issue where events of some markers are more incomplete than others, we propose to incorporate marker features such as geographic information to regularize the infectivity kernel matrix between markers. In such a way, a MCMC sampler is encouraged to generate more virtual events with markers that are biased. Infectivity matrix regularization is described in Section IV-B1.

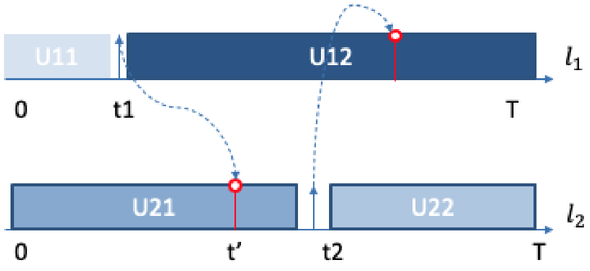


Fig. 1. Missing Window. Each blue arrow represents an observed event; each red dot represents a virtual event; the darkness of shade represents how likely an event can be missing in the corresponding time window. (The lightest colored U_{11} is with the smallest possibility to have missing data). Dash lines represent parent-child relationship.

A. Missing Window Probability

A collection of n events $x = \{(t_1, l_1), (t_2, l_2), \cdots, (t_n, l_n)\}$ are observed during a fixed time period $[0, T)$, where events are ordered by time $0 < t_{i-1} < t_i < T$ and l_i is the marker (e.g., location) of the event at time t_i. For each marker l_i, we assume events can be missing at any time window between two observed events. We define $U_{ij} = [s_{ij}, e_{ij}]$ as the j-th unobserved time window associated with marker l_i, where s_{ij} is the start time of U_{ij} and e_{ij} is the end time of U_{ij}.

For example, in Fig. 1, two events $x = \{(t_1, l_1), (t_2, l_2)\}$ of different markers l_1, l_2 are observed during the time period $[0, T)$. Take rescue request events as an example, event markers are the ZIP codes where requests are sent out. One event with marker l_1 occurs at time t_1 and another event with marker l_2 happens later at time t_2. In this example, $U_{11} = [0, t_1]$ and $U_{12} = [t_1, T)$ are the first and second missing windows associated with marker l_1. Similarly, $U_{21} = [0, t_2]$ and $U_{22} = [t_2, T)$ are the first and second missing windows associated with marker l_2. Suppose event (t_1, l_1) is likely to trigger another event of marker l_2 at $t' \in [0, t_2] = U_{21}$ but the event is unobserved. We say that the missing window U_{21} covers the event and the probability of this event occurring in U_{21} is $P_{U_{21}}$. Similarly, the event (t_2, l_2) may trigger other events that fall in the missing window U_{12} with probability $P_{U_{12}}$.

Intuitively, the probability of event (t', l_i) being missing should depend on the historical events before U_{21}, the length of missing window U_{21}, and some prior knowledge of its marker l_2, if available. Note that conditional intensity function λ is defined as the limit of the expected number of events in $[t, t + \Delta t]$ given the history H_t of all the events up to time t [21], that is $\lambda_l(t) = \lim_{\Delta t \to 0}(E[N_l(t, t + \Delta t)|H_t]/(\Delta t))]/dt$. We propose that the missing window probability should be proportional to the expected number of events in missing window U_{ij}, given the historical observed events before the end of the missing window.

Formally, let t' be the time of a possible missing event. Denote $H(e_{ij}) = \{(t_k, l_k)|t_k < e_{ij}, l_k \in L\}$ as the historical events before the end of missing window e_{ij}. The probability $P_{U_{ij}}$ of an event missing in a time window U_{ij} is defined as the probability of next arrival time in U_{ij} given historical events before U_{ij}:

$$\begin{aligned} P_{U_{ij}} &= P(t_{s_{ij}} < t' < t_{e_{ij}} | H(e_{ij})) \\ &= 1 - P(No\ events\ happen\ in\ U_{ij} | H(e_{ij})) \\ &= 1 - \exp[-\int_{t_k}^{t_{e_{ij}}} \lambda_{l_i}(s) ds] \\ &= 1 - \exp[-\int_{t_k}^{t_{e_{ij}}} \sum_k \Phi_{l_k, l_i}(s - t_k) ds] \\ &= 1 - \exp[-\Phi_{\star, l_i}(t_{e_{ij}} - t_k)]. \end{aligned} \tag{8}$$

With this definition, a larger value of $P_{U_{ij}}$ suggests a higher probability of a next arrival event missing in U_{ij} than in other time windows.

We adopt a MCMC sampler similar to the one described in Section III-B to estimate model parameters but integrate $P_{U_{ij}}$ into the likelihood of event sequences. The sampler is also based on auxiliary variables such as $\tilde{x}$, $\tilde{a}$. The joint likelihood function will guide the sampler in three moves. Since virtual events $\tilde{x}$ and sampled events (part of x, changed from virtual events by "virtualness" move) are children sampled from either root events or observed events, they must fall into some unobserved time windows. We propose to weigh these events by the probabilities of missing windows where they fall into. In such cases, those events that fall into the window U_{ij} that are too short or too early to have a parent event (close to 0) will be down-weighted. For example, U_{11} in Fig. 1 is the first missing window on l_1, there is no historical observed event before U_{11} except for the root event. Therefore, no potential parents could generate children events during this window.

Formally, let $\mathbb{1}(\tilde{t}_i \in U_{\tilde{l}_i, j})$ be an indicator function, which takes value 1 if sampled events $\tilde{t}_i$ fall into the j-th missing window of marker $\tilde{l}_i$, and 0 otherwise. Now we multiply the corresponding $P_{U_{ij}}$ to each sampled event, and update the likelihood as:

$$\begin{aligned} p(x, a, \tilde{x}, \tilde{a}) = & \prod_{i=1}^{n} \exp[-(\kappa + 1)\Phi_{l_i, \star(T - t_i)}] \times \prod_{i=1}^{n} \phi_{l_{ai}, l_i}(t_i - t_{ai}) \\ & \times \prod_{i=1}^{\tilde{n}} [\kappa \cdot \Phi_{l_{\tilde{a}_i}, \tilde{l}_i}(\tilde{t}_i - t_{\tilde{a}_i})) \cdot \sum_j P_{U_{\tilde{l}_i, j}} \cdot \mathbb{1}(\tilde{t}_i \in U_{\tilde{l}_i, j})]. \end{aligned} \tag{9}$$

B. Learning from Biased Data

Events associated with different markers (e.g., location indices) can be highly biased. For example, rescue events mostly appear in the top one or two hotspots, and events from regions of lower socioeconomic status are underrepresented. In the previous section, virtual events are generated in missing windows to improve model parameter estimation. To further tackle the bias challenge, we propose to enforce our sampler to generate more virtual events with markers that are short of data, so that the model learned using augmented data is more accurate in prediction. To correct event data with underrepresented markers, we propose to incorporate marker features such as geographic and demographic information to guide the learning of the Hawkes process parameters. Intuitively, we can define markers with similar features as neighbors. For example, in flood rescue events, markers (e.g., location indices) with similar flooding levels may be considered neighbors. If events of marker l_i are rare but we observe more events of markers that are neighbors to l_i, we consider that marker l_i is likely biased. We propose to utilize such neighbor information to regularize the parameters of the Hawkes processes.

1) Regularization: As described in (4), the infectivity matrix M characterizes the excitation rates between markers at any time. Ideally, if data are observed fully with no bias, the infectivity matrix can be learned to fit the event excitation patterns. However, in case of incomplete biased data, the matrix M may not fully capture the underlying infectivities. In such cases, we would like to regularize M from the prior knowledge such as the similarities between markers. Markers are defined to be neighbors if they have similar features. The details are described in Section IV-B2. If markers are neighbors, the events associated with the markers should have similar behaviors in exiting children events or be excited by parents. Specifically, if markers l and l' are neighbors, they should have similar impacts on other markers (i.e., $M_{l\cdot}$ is close to $M_{l'\cdot}$). They should be influenced by other markers in a similar way (i.e., $M_{\cdot l}$ is close to $M_{\cdot l'}$). This regularization helps to learn clusters in events. By integrating the regularization into Hawkes model estimation, our sampler in Section IV is able to generate more events for a marker that is underrepresented but is a neighbor to some markers with high volumes of events, and to generate fewer events for a marker that is less biased and is similar to markers with low event densities.

Formally, we define our pairwise regularization term as:

$$\mathbf{E}(M) = \sum_{l=1}^{L} \sum_{l' \in \mathbb{N}(l)} ||M_{\cdot l} - M_{\cdot l'}||_F^2 + ||M_{l\cdot} - M_{l'\cdot}||_F^2, \quad (10)$$

where $\mathbb{N}(l)$ is the set of neighbors of marker l, $M_{\cdot l} \in \mathbb{R}^L$ is the l-th column vector of infectivity matrix M, and $M_{l\cdot} \in \mathbb{R}^L$ is the l-th row vector. We name it our L2 regularization. We also include L1 regularization on M defined as $||M||_1 = \sum_{l,l'} |M_{l,l'}|$ to enforce sparsity.

Let Θ denote all the parameters in the Hawkes process (i.e., μ_l, M, β, etc.), the learning problem is

$$\min_{\Theta} -Log(p(x, a, \tilde{x}, \tilde{a}))_{\Theta} + \alpha_1 ||M||_1 + \alpha_2 \mathbf{E}(M). \quad (11)$$

After model initialization, the learning process includes the following: we sample data with the MCMC sampler given current model parameters. Then we calculate the negative log-likelihood in (11). We use the gradient descent method to update model parameters to minimize the negative log-likelihood. We repeat this process until (11) is minimized and stable.

2) Define Neighbors: We use features of markers to help define their neighbors. The choice of features depends on domain applications. For example, in Houston rescue data, each marker is a location index, we can use geographic information and population information to define neighbors of markers. Assume there are n total features, each marker can be represented as an n-dimensional vector. We can calculate the similarity scores between different markers. Then for each marker l_i, we pick the ones with highest k similarity scores to be the neighbors of l_i. The neighbor size can be chosen via cross validation.

We adopt cosine similarity to be the similarity measurement for defining the neighbors in our experiments. Cosine similarity is one of the most popular metrics. Given two n-dimensional instance $\boldsymbol{a} = (a_1, ..., a_n) \in \mathbb{R}^n$ and $\boldsymbol{b} = (b_1, ..., b_n) \in \mathbb{R}^n$, the similarity is defined as:

$$\mathcal{S}_{cos}(\boldsymbol{a}, \boldsymbol{b}) = \frac{\sum_{i=1}^{n} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}}. \quad (12)$$

Integrating marker features into model regularization also helps us learn whether or not specific features are associated with data bias. If the prediction results get better and there are more predictions on minority classes, we can conclude that using those features can help reduce data bias. Otherwise, those features may be not correlated with data bias.

V. Experiments

A. Data

We evaluate our methods on three real-world datasets and provide the summary in table I.

The first dataset "Chicago" contains gang-related events from 1965 through 1995 filtered from a homicide report provided by the Chicago Police Department [22]. The details of "gang-related" event data are described in existing studies [4], [23]. Each gang-related homicide event is associated with a timestamp and a marker (e.g., a location index) to indicate the community where the event occurs. There are 77 Chicago communities and 2195 events. A unit time t is a day. Similar to [4], we use data in year 1993 and 1994 as training data and data in 1995 for testing. There are no available features for Chicago communities.

The second dataset "Dallas" is obtained from Kaggle[1], which contains different types of crime incidences collected for around 3 years from the Dallas Police Department. We focus on "ROBBERY OF BUSINESS" events for a length of 380 days. Each robbery incident is an event with the corresponding ZIP code as its marker. For demographic features associated

[1] https://www.kaggle.com/carrie1/dallaspolicereportedincidents

with each marker, we count the number of events reported by three races (i.e., Black, White, and Hispanic/Latino). There are 2058 events and we use 80% data for training and the rest for testing.

The third dataset is "Houston" rescue data[2], which contains social media rescue requests from locations around Harris county in Houston area during the Hurricane Harvey. Each rescue request contains a time stamp and a corresponding ZIP code as the event marker. There are 1182 events during 26 hours. We use the first 21 hour data (about 70% events) as the training data and the last 5 hour data as the testing data. We collect the race population statistics from American Fact Finder[3]. We get geographic and social vulnerability variables such as flooding zone, flood depth, civilian unemployed, per capital GDP, minority, population as features for each marker (i.e., ZIP code) from U.S. census data.

Specifically, Houston flood rescue data are highly biased. For example, the adjacent regions with ZIP codes 77044 and 77049 have similar mean flood depth and population. However, there are about 230 rescue tweets from ZIP code 77044 in one day and only 68 tweets from 77049. Those two regions are in the same level of flooding vulnerability but have imbalanced representations in Twitter events.

B. Biased Training Data

We create biased datasets to demonstrate that our method can improve prediction accuracy even if the training data are highly biased. Specifically, we consider the original data as full data and split it into a training set and a test set. Then we follow different strategies to create biased training sets from the full training set. We will train different models on the biased training sets, and evaluate the models on the same complete test set.

We design a top-K strategy to create a biased training set based on the total number of events per marker. Specifically, we rank markers by the number of events associate with them. In some cases, top-ranked markers (e.g., hotspots) contain a large percentage of events. For example, in the Houston data, top 1 hotspot contains 50% of total events. Using a top-K strategy, we first select all events of top k markers (e.g., events from hotspots), and then sub-sample events of the rest of markers randomly with a given rate (e.g., 60%). As the selection rate decreases from 60% to 20%, the events become more imbalanced among different markers. Therefore, we can evaluate models on data with different degrees of bias.

C. Evaluation Metrics

We evaluate the accuracy of predicted markers l_i of the testing event at time t based on historical events $H(t_i)$. The Hawkes model parameters are estimated using the training data before time t. At each unique time unit t in the test event set, the model can sample a set of events associated with different markers. The most probable marker is the one with the highest probability.

Precision, Recall and F-1 We calculate the average precision, recall and F-1 score for all time units in the testing cases. At each testing time unit, we consider a prediction is correct, only if the highest ranked marker (by probability) is the same with ground truth. For datasets (Houston and Dallas) with multiple ground-truth markers (e.g., multiple ground-truth ZIP codes in the same time unit), we consider the prediction is correct as long as the highest ranked marker is one of the ground-truth markers.

1) Precision: Precision is the number of correct predictions divided by the number of event markers sampled.
2) Recall: Recall is the number of correct predictions divided by the number of ground-truth markers.
3) F-1 score: F-1 score is harmonic mean of precision and recall.

MAP, MAR and MRR Markers are ranked from the highest to the lowest by the predicted probabilities. Since accuracy at the top are more appropriate for imbalanced data to address or alleviate the minority bias, we use top-k mean average precision (MAP), mean average recall (MAR) and mean reciprocal rank (MRR) to evaluate the prediction accuracy. Rank k can be chosen to fit different applications. For example, considering that the rescue resources are usually limited during each short time unit, we can predict locations that need the most help. In our case, Houston data has multiple ground-truth markers at one time unit (i.e., requests from multiple locations at the same time). We choose top-2 MAP, MAR and MRR in the experiment.

1) Mean Average Precision@2: We calculate precision at top 2 of the predicted markers, then average over all test cases. We claim if a predicted marker is any of the multiple ground-truth ones, it is a correct prediction.
2) Mean Average Recall: We calculate recall at top 2 of the predicted markers and average over all test cases.
3) Mean Reciprocal Rank: We compute the rank of each ground-truth markers and calculated the reciprocal rank score by $rr = \frac{1}{rank}$ and average over all test cases.

D. Baselines

We compare our method with two baseline methods and their variations:

Stitch Method The work [11] assumes an observed sequence is double censored. It is proposed to stitch multiple short sequences with probabilities proportional to the similarities between those short sequences. The similarity depends on the proximity between the start of a sequence and the end of another sequence. It also depends on features associated with sequence markers.

Hidden Nodes We use "Hidden Nodes" to refer to the sampling method [4], in which it is assumed there exists hidden nodes (unobserved markers) that affect distribution of events and therefore introduce extra nodes and generate sample events from and onto these hidden nodes. We set 5 hidden nodes, the same with [4] for Chicago crime data. For Houston rescue data, we choose the size of hidden nodes that can give the best prediction result.

[2] https://data.world/sya/harvey-rescue-doc

[3] https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

TABLE I
DATASET DESCRIPTION.

Dataset	Events	Geo-Type	Unique-IDs	Time	Neighborhood Features
Chicago	2195	ZIP Code	77	$901d$	NA
Dallas	2058	ZIP Code	74	$380d$	Demographics (e.g., race)
Houston	1182	ZIP Code	106	$26h$	Demographics, Geographical (e.g., flood zone)

TABLE II
PARAMETERS INITIALIZATION.

Dataset	μ	κ	max init M	α_1	α_2
Chicago	0.011	2	0.0056	150000	200000
Dallas	0.03	2	0.005	5000	100000
Houston	0.011	2	0.0056	1500	4500

TABLE III
PREDICTION OF NEXT EVENT LOCATION FOR HOUSTON RESCUE.

					Methods	
Top K	Metrics	Biased	Stitch	Hidden Nodes	WithL2 (ours)	Full
Top 1	Precision	0.2239	0.2027	0.2696	**0.5383** (99.6%)	0.2346
	Recall	0.1883	0.2804	0.1711	**0.3624** (29.2%)	0.2102
	F-1	0.1794	0.1771	0.1801	**0.4033** (123.9%)	0.1940
Top 2	Precision	0.2708	0.3622	0.2643	**0.5294** (46.1%)	0.2346
	Recall	0.1898	0.2061	0.1363	**0.3588** (74.1%)	0.2102
	F-1	0.1930	0.3092	0.2143	**0.3990** (29.1%)	0.1940
Top 3	Precision	0.2767	0.3027	0.2653	**0.4489** (48.3%)	0.2346
	Recall	0.1580	0.1850	0.1376	**0.3112** (68.3%)	0.2102
	F-1	0.2386	0.2729	0.2156	**0.3353** (22.9%)	0.1940
Top 5	Precision	0.2272	0.2810	0.2505	**0.4777** (70.0%)	0.2346
	Recall	0.1768	0.2164	0.1499	**0.3287** (51.9%)	0.2102
	F-1	0.1670	0.2166	0.1529	**0.3602** (66.3%)	0.1940

Full We assume the original event sequences are complete (no missing data) and train multivariate Hawkes models without sampling events.

Biased We train multivariate Hawkes models with biased data created by Section V-B as it is complete (no missing data).

We compare those methods with our model. Specifically, our model has two variations. **NoL2** is our re-balanced method without pairwise-similarity regularization on infectivity matrix. **WithL2** is our re-balanced method with pairwise-similarity regularization on infectivity matrix.

In general, model "Full" is the reference method. Model "Biased" is expected to be the worst since the model can be misled by biased data. Other methods such as "Stitch", "Hidden Nodes", our "NoL2", and our "WithL2" attempt to either fill incompleteness or re-balance the data. Those methods should be better than "Biased" or even "Full".

For each dataset and method, we tune the parameters to achieve the best. For example, for our model, we tune regularization parameters α_1 and α_2, the initialization of infectivity matrix M, and virtual rate κ, etc.

E. Results

1) Top-1 with different subsampling rates: In this experiment, for each dataset, we would like to compare the improvements of each method for three different levels of bias. We use top-1 strategy where we keep all the events of top-1 marker and select 20%, 40% and 60% of the rest. We plot rank 2 MAP, rank 2 MAR and MRR for Chicago data in Fig. 2, Houston data in Fig. 3 and Dallas data in Fig. 4.

For all three datasets, our method "WithL2" outperforms almost all the others in most of the metrics. The second best model is our method "NoL2" without pairwise-similarity regularization. This means that our sampler alone without the help of regularizer can re-balance the biased data and generate better results than baseline methods. With the help of the neighborhood information and regularization, our sampler can generate more useful data and learn a better model. Method "Full" stays the same across different bias levels as it is trained on the same complete data. For other methods, as training data become less biased (e.g., 20% to 60%), the prediction accuracy on the test set improves in general.

Specifically, for "Chicago" in Fig. 2, "WithL2" outperforms all other methods in all metrics for 40% and 60% sampled data. It is bellow "Full" for 20% sampled data, but still the highest one, in comparison with other baseline methods. This suggests that our method can mitigate extreme biased data.

Our method "WithL2" performs the best in Houston data for all metrics and all levels of bias except for MRR at 20% (Fig. 3). Stitch method is the best in MRR at 20% bias level and the same as our method "WithL2" at 40% and 60% levels. The model "Full" trained by original data performs not even as good as the one trained on biased data. It suggests that Houston data may be noisier and may contain redundant data. Methods such as "WithL2", "NoL2" and "Hidden Nodes" generate more balanced data from biased data and achieve better prediction performance than original data.

In Dallas data, our method again outperforms other methods. Similar trends are observed in Fig. 4. Specially in MAR graph, "WithL2" is the only method that can compete with "Full".

2) Top-K with different K: We keep events from top 1 to top 5 locations and randomly sample 60% of the rest of data as biased training data. We compare prediction accuracy in terms of precision, recall and F-1 score.

We compare methods on Houston rescue data. Because events happen in multiple locations in each hour, a prediction is correct if it is any one of the ground-truth locations. The results are shown in table III. First, "Biased" performs worst for most of Top K settings in all three metrics. This suggests when data are biased, the model trained on biased data is distorted. Second, our method "WithL2" performs the best in all Top K settings of all three metrics. We also show the percentage of improvement of our method in comparison with the second best baseline. For example, our method has a precision of 0.5383 in Top 1 setting. In comparison with the

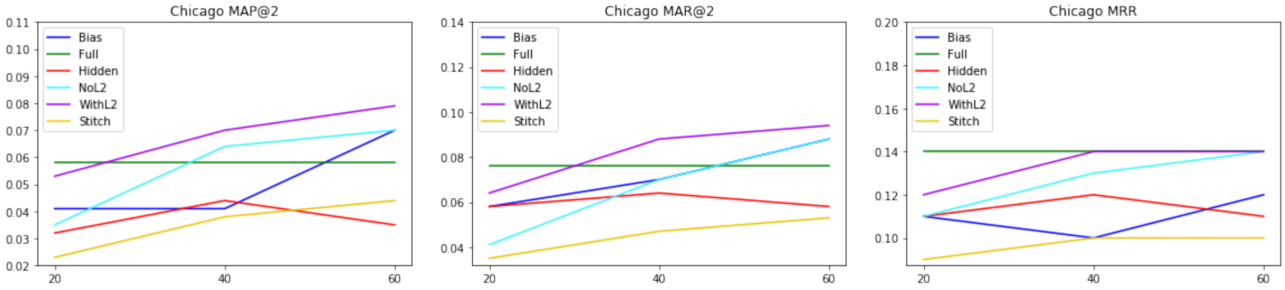


Fig. 2. Chicago Rank 2 MAP, MAR and MRR.

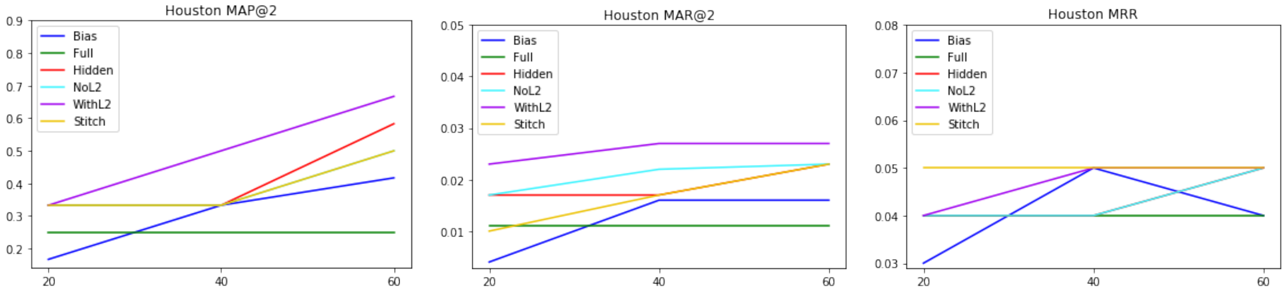


Fig. 3. Houston Rank 2 MAP, MAR and MRR.

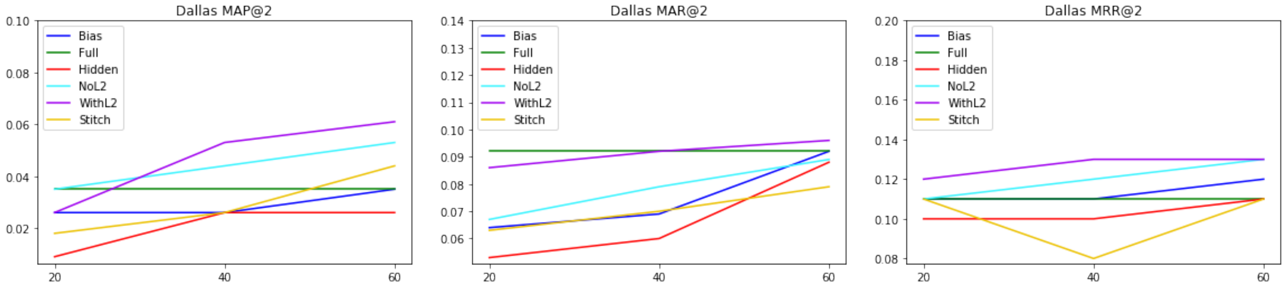


Fig. 4. Dallas Rank 2 MAP, MAR and MRR.

baseline "Hidden node", we improve 99.6%.

We try different sets of parameter initialization and choose the set that makes gradient descent efficient and stable. We set parameters as shown in table II for our method. Our code is available at Github[4].

VI. Case Study

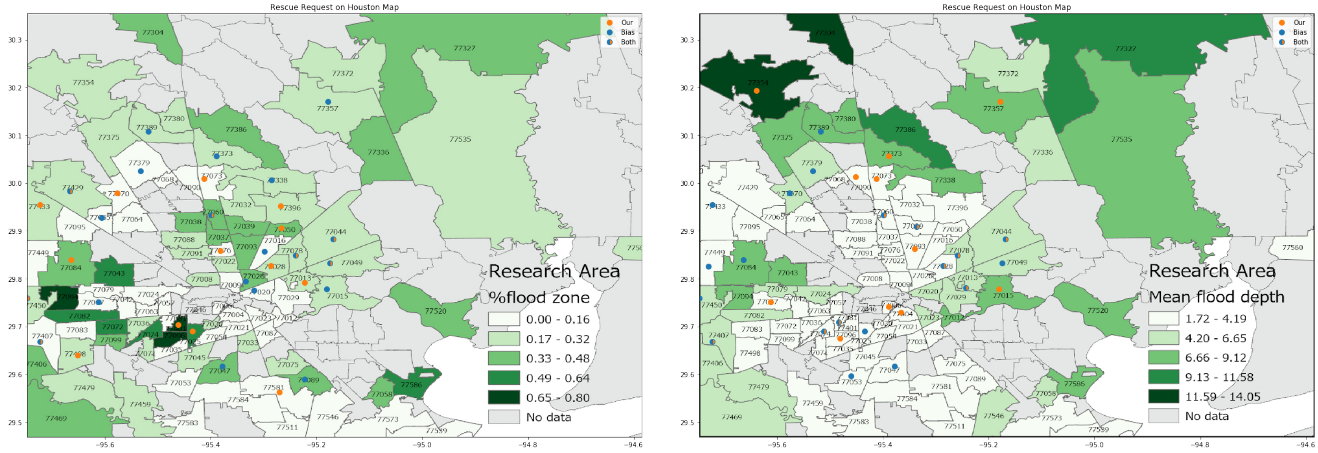
In this section, we explore the effect of different choices of features in defining "neighborhood" of markers on Houston rescue data. In Houston rescue data, we have side information for each ZIP code, including "% flood zone", "Mean flood depth", "Civilian unemployed", "Per capita GDP", "Minority", and "Population".

First, we use only flood related information to define neighbors and regularize the model. That is assuming that ZIP codes with similar flood areas and depths should have similar rescue requests. Pairwise regularization will force the sampler to generate more events in a location that is short of data but has a similar flood situation with locations with many request events.

The second choice of features is "Civilian unemployed" and "Per capita GDP". Using these two features is assuming locations with similar economic status should have similar rescue requests. This assumption is based on the intuition that if two ZIP codes are similar in economic status, similar numbers of people should send rescue requests. This way, it can mitigate the bias due to the fact that some people may be not used to send rescue requests through social media.

The third choice of features is the combination of "Minority" and "Population". "Minority" is the percentage of minority in a ZIP code, and "Population" is the total number of people in a ZIP code. Using these two features to define neighbor assumes that if two ZIP codes have similar population with similar percentages of minority people, they should have similar numbers of request events. This way we can test if the data bias is associated with minority groups.

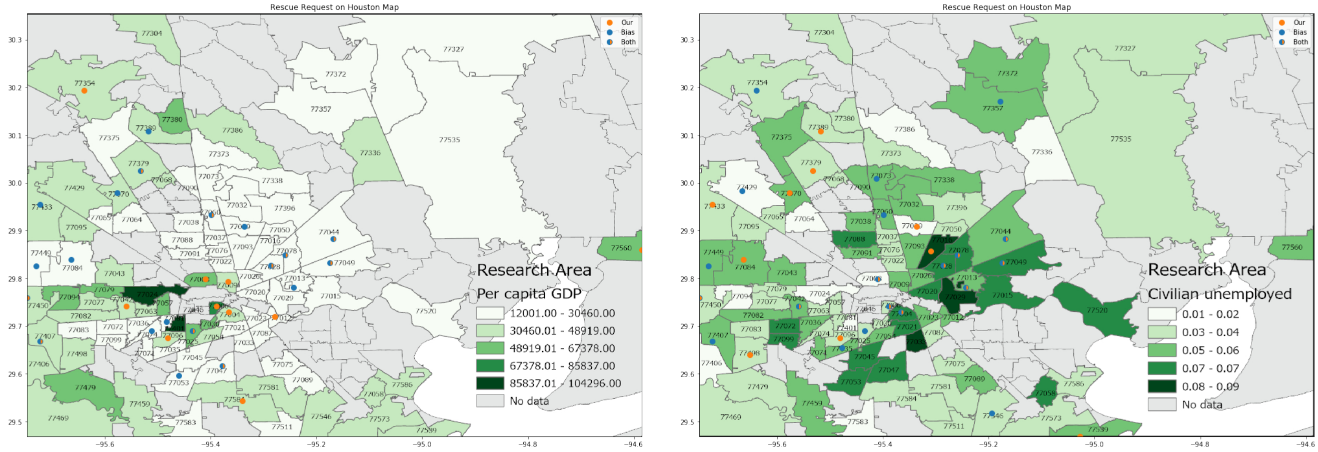
[4] https://github.com/Zihan-Zhou/BiasedIncomp_MHP



(a)

(b)

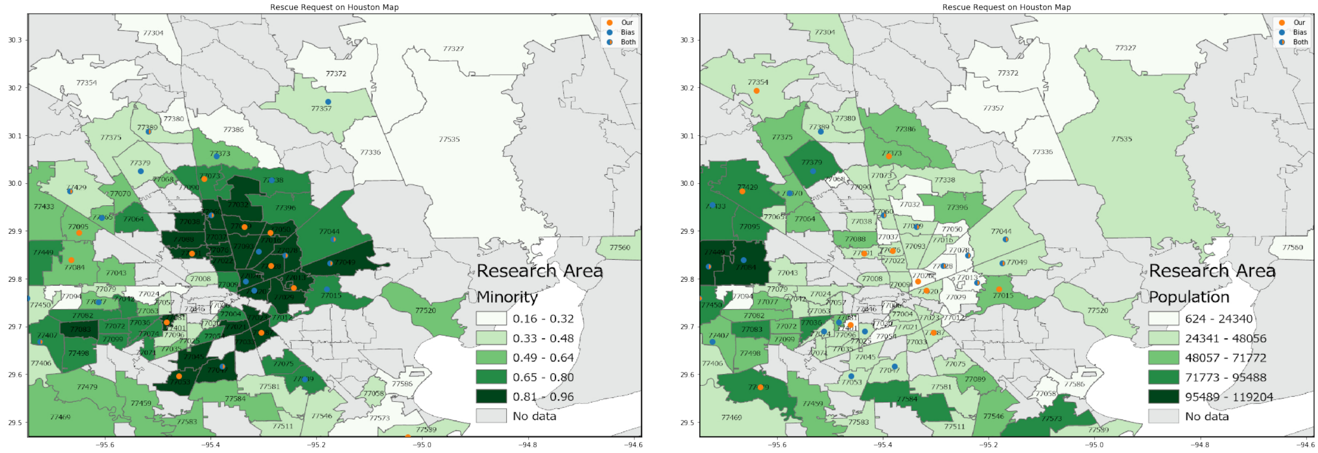
Fig. 5. Houston map with flood level.



(a)

(b)

Fig. 6. Houston map with economic situation.



(a)

(b)

Fig. 7. Houston map with minority population.

For each choice of features, we visualize the top-20 detected hotspots using baseline "Biased" and ours "WithL2" with the set of features. Blue dots represent the predicted top ZIP codes using baseline "biased". Orange dots represent the predicted top ZIP codes using our method. The dots of both blue and red indicate that the locations are captured in both ranking lists.

The background in Fig. 5(a) shows the map of Houston area with ZIP codes colored by "% of flood zone". Our prediction has 31% locations with high flooding level, compared to 15% using "biased" baseline. This indicates that our method can guide the sampler to sample more data in high-level flood zone areas. Similar trends are observed in Fig. 5(b). The original data distribution shows that some regions in high "% flood zone" and high "mean flood depth" are short of data. Our method mitigates the bias issue and predicts more events in these regions.

Fig. 6(a) shows the predicted ZIP codes by integrating feature "Per capita GDP". 10% of ZIP codes predicted by our method belong to regions of top-2 level of GDP, compared to 0% by "Biased" baseline. Original data distribution by five levels of GDP shows that the data are short in high level GDP. Our method mitigates this issue and predicts more on high level GDP. Similar trends are observed in Fig. 6(b), where we have more predictions on low unemployment regions.

Fig. 7(a) shows Houston map with different levels of minority rates and Fig. 7(b) is the map with different population levels. Our predictions concentrate more on regions with higher minority rates.

VII. Conclusion

To our best knowledge, we are the first to propose a novel learning method for multivariate Hawkes processes from incomplete biased data. We carry on experiments over several real-word datasets, and our model improves prediction accuracy in comparison with the state-of-arts. In case studies, we demonstrate the results of incorporating different event marker features, which provides valuable discussions for the adaptation of Hawkes processes in fairness sensitive domains such as crime and disaster management. Our proposed method can help government agencies design a fair allocation of public sources, especially for the disadvantaged classes of individuals or communities.

VIII. Acknowledgement

This work was supported in part by the National Science Foundation under Grant No. 1927513 and No. 1943486.

References

[1] E. Bacry, I. Mastromatteo, and J.-F. Muzy, "Hawkes processes in finance," *Market Microstructure and Liquidity*, vol. 1, no. 01, p. 1550005, 2015.

[2] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, "Self-exciting point process modeling of crime," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 100–108, 2011.

[3] S. Linderman and R. Adams, "Discovering latent network structure in point process data," in *Proc. of the International Conference on Machine Learning (ICML)*, 2014, pp. 1413–1421.

[4] C. R. Shelton, Z. Qin, and C. Shetty, "Hawkes process inference with missing data," in *Proc. of the AAAI Conference on Artificial Intelligence*, 2018.

[5] K. Zhou, H. Zha, and L. Song, "Learning triggering kernels for multi-dimensional hawkes processes," in *International Conference on Machine Learning*. PMLR, 2013, pp. 1301–1309.

[6] J. R. Zipkin, F. P. Schoenberg, K. Coronges, and A. L. Bertozzi, "Point-process models of social network interactions: Parameter estimation and missing data recovery," *European journal of applied mathematics*, vol. 27, no. 3, pp. 502–529, 2016.

[7] M.-A. Rizoiu, Y. Lee, S. Mishra, and L. Xie, "Hawkes processes for events in social media," in *Frontiers of Multimedia Research*, 2017, pp. 191–218.

[8] N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song, "Recurrent marked temporal point processes: Embedding event history to vector," in *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1555–1564.

[9] J. Shang and M. Sun, "Local low-rank Hawkes processes for temporal user-item interactions," in *Proc. of the IEEE International Conference on Data Mining (ICDM)*, Nov. 2018, pp. 427–436.

[10] L. Zou, N. Lam, S. Shams, H. Cai, M. A. Meyer, S. Yang, K. Lee, S.-J. Park, and M. A. Reams, "Social and geographical disparities in Twitter use during Hurricane Harvey," *International Journal of Digital Earth*, pp. 1–19, 2018.

[11] H. Xu, D. Luo, and H. Zha, "Learning Hawkes processes from short doubly-censored event sequences," in *Proc. of the International Conference on Machine Learning (ICML)*, 2017, pp. 3831–3840.

[12] M. J. Werner, A. Helmstetter, D. D. Jackson, and Y. Y. Kagan, "High-resolution long-term and short-term earthquake forecasts for California," *Bulletin of the Seismological Society of America*, vol. 101, no. 4, pp. 1630–1648, 2011.

[13] G. Mohler, R. Raje, J. Carter, M. Valasik, and J. Brantingham, "A penalized likelihood method for balancing accuracy and fairness in predictive policing," in *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018, pp. 2454–2459.

[14] E. W. Fox, M. B. Short, F. P. Schoenberg, K. D. Coronges, and A. L. Bertozzi, "Modeling e-mail networks and inferring leadership using self-exciting point processes," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 564–584, 2016.

[15] G. Mohler, J. Carter, and R. Raje, "Improving social harm indices with a modulated hawkes process," *International Journal of Forecasting*, vol. 34, no. 3, pp. 431–439, 2018.

[16] H. Mei and J. M. Eisner, "The neural hawkes process: A neurally self-modulating multivariate point process," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6754–6764.

[17] H. Mei, G. Qin, and J. Eisner, "Imputing missing events in continuous-time event streams," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4475–4485.

[18] J. D. Tucker, L. Shand, and J. R. Lewis, "Handling missing data in self-exciting point process models," *Spatial statistics*, vol. 29, pp. 160–176, 2019.

[19] Y. Guan and J. M. Loh, "A thinned block bootstrap variance estimation procedure for inhomogeneous spatial point patterns," *Journal of the American Statistical Association*, vol. 102, no. 480, pp. 1377–1386, 2007.

[20] M. Oliveira, N. Moniz, L. Torgo, and V. S. Costa, "Biased resampling strategies for imbalanced spatio-temporal forecasting," in *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2019, pp. 100–109. [Online]. Available: https://doi.org/10.1109/dsaa.2019.00024

[21] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Springer Science & Business Media, 2007.

[22] C. R. Block, R. L. Block, and I. C. J. I. Authority, "Homicides in chicago, 1965-1995," *Inter-university Consortium for Political and Social Research* [*distributor*], vol. 32, 2005.

[23] A. Papachristos, "Murder by structure: Dominance relations and the social structure of gang homicide," *American Journal of Sociology*, vol. 115, no. 1, pp. 74–128, 2009. [Online]. Available: http://www.jstor.org/stable/10.1086/597791