List-wise Fairness Criterion for Point Processes

Jin Shang, Mingxuan Sun, and Nina S. N. Lam
Louisiana State University
jshang2@lsu.edu,msun@csc.lsu.edu,nlam@lsu.edu

ABSTRACT

Many types of event sequence data exhibit triggering and clustering properties in space and time. Point processes are widely used in modeling such event data with applications such as predictive policing and disaster event forecasting. Although current algorithms can achieve significant event prediction accuracy, the historic data or the self-excitation property can introduce biased prediction. For example, hotspots ranked by event hazard rates can make the visibility of a disadvantaged group (e.g., racial minorities or the communities of lower social economic status) more apparent. Existing methods have explored ways to achieve parity between the groups by penalizing the objective function with several group fairness metrics. However, these metrics fail to measure the fairness on every prefix of the ranking. In this paper, we propose a novel list-wise fairness criterion for point processes, which can efficiently evaluate the ranking fairness in event prediction. We also present a strict definition of the unfairness consistency property of a fairness metric and prove that our list-wise fairness criterion satisfies this property. Experiments on several real-world spatial-temporal sequence datasets demonstrate the effectiveness of our list-wise fairness criterion.

KEYWORDS

Fairness; Ranking; List-wise; Spatial-Temporal Point Process

ACM Reference Format:
Jin Shang, Mingxuan Sun, and Nina S. N. Lam. 2020. List-wise Fairness Criterion for Point Processes. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3394486.3403246

1 INTRODUCTION

Many types of event sequence data exhibit triggering and clustering properties in space and time. For example, after a large earthquake, events of after-shocks usually occur in the following days or weeks near the epicenter of the main shock [8]. Similarly, criminologists have reported that 25% to 50% crime events are observed in a few areas of a city [17]. They have also demonstrated that certain types of crime events such as burglaries are often reported repetitively from the same neighborhood [2]. The time interval and spatial distance among events carry important information about the underlying dynamics of a specific type of events.

Predicting and ranking the rate of events as a function of space and time enables important applications. Typically, space is divided into regions, time is divided into short intervals, and regions are ranked based on the predicted event rates over a time window. For example, in a predictive policing system, a city is divided into geographic sub-regions such as grid cells or political boundaries. A predictive algorithm is used to forecast the rates of crime events for each region at each day based on historical crime events. According to the predicted rates, police daily patrol activities can be adjusted so that more resources are allocated to the regions with higher risks. In practice, due to limited resources, regions are ranked by the predicted hazard rates in each day and police activities are directed to top-k regions, also known as hotspots [17].

Variations of point-process models [22, 23] have become very popular for modeling event rates based on historic events. They assume different forms of dependencies on the history. For example, the Hawkes process [7, 14] assumes that the influences from previous events are linearly additive towards the current event. Such models are able to capture the temporal correlations between events and are well-suited for inhomogeneous inter-event time modeling. Spatial-temporal Hawkes models extend temporal models to predict the rate of events at a specific location and time. Spatial heterogeneity in hazard rates can be characterized as base intensities and the self-exciting effects can be modeled with a variety of temporal kernels. Model parameters can be estimated using standard maximum likelihood estimators given training data, e.g., events observed before a specific time.

Although those predictive models improve event forecasting accuracy, biased predictions may be introduced and amplified due to factors such as data bias and the feedback loop of algorithms. For example, time-stamped geo-tagged event data from Twitter have been used for rapid flood mapping, damage assessment, and situation awareness. However, it has been reported that higher disaster-related Twitter-use communities tend to be of higher socioeconomic status [33]. Prediction based on such data may exhibit socioeconomic bias. Moreover, recent studies have focused on the bias problem of event prediction in predictive policing. One potential problem is that if the police only patrol areas with higher estimated risks, there will likely be more arrests than in other areas, and then biased arrests may be further amplified through the feedback loop.

While there have been some early explorations [12, 24, 28, 31] in developing ranking fairness metrics that can be adopted by hazardous event prediction, most of them focus on either measurement of fairness or post-processing ranking list to satisfy a fair condition. Hence, the ranking functions are not influenced by the fairness metrics. A recent work [18] introduces demographic parity into spatial-temporal crime prediction and directly uses it to penalize

KDD '20, August 23–27, 2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3403246

the likelihood function. The fairness metric enforces the amount of police patrol allocated to each demographic group in selected hotspots to be proportional to the percentage of that group in the whole population. However, the fairness metric does not guarantee group parity at any point in ranked regions.

In this paper, we propose a novel list-wise fairness criterion for spatial-temporal point process, which can efficiently evaluate the list ranking fairness. We also present a strict definition of the unfairness consistency property of a fairness metric and prove that our list-wise fairness criterion satisfies this property. We further integrate the fairness criterion into the objective function and then obtain a fairness-aware ranking function that can generate a fair ranking list. We carry on experiments over several real-word spatial-temporal datasets, and the results demonstrate the effectiveness of our list-wise fairness criterion. We also discuss the scalability of our method and propose a smoothed variation, which makes it easier for optimization.

2 RELATED WORK

Spatial-temporal Hawkes processes, which are capable of modeling correlated spatial-temporal event sequences, have been widely used in various applications including earthquake prediction [27], predictive policing [18], and hazard rate prediction [7, 14]. Traditionally, events are aggregated in discrete time intervals over a set of grid cells. A regression model is learned to predict event occurrences in each cell and time interval given spatial and temporal covariates, and previous counts. However, these methods suffer from discrete granularity in both space and time. In comparison with traditional approaches, point processes show better prediction accuracy for predicting event hazard rates and ranking event hotspots [16].

Machine learning and artificial intelligence (AI) systems exhibit bias due to a number of factors including the human bias in training data and the design of algorithm models. It is also well known that machine learning and AI algorithms may reproduce and even amplify human biases and social inequities especially in applications involving feedback loops such as predictive policing [10, 18]. There are many definitions of fairness such as group parity [4], equalized odds [9, 30], individual fairness [5], and counterfactual fairness [21]. Group parity and its variations are widely applied in classification and regression tasks [3, 11, 29].

The impact of imposing fairness constraints to machine learning and AI is dependent on the specific domain datasets, the specific fairness definition, and the prediction algorithms. Most of the models and algorithms proposed to improve fairness fall into three categories: pre-processing [15, 32], optimization at training time [11, 30], and post-processing [6, 9]. Generally, training time optimization, which is domain-specific, can achieve good performance on accuracy and fairness measures and offers the flexibility to balance the trade-off between accuracy and fairness measures.

Recent studies have focused on the fairness problem of ranking. Specifically, a fairness measure is proposed in [28] to compare the distributions between two demographic groups at several prefixes with a discount factor based on an inverse logarithmic function. However, the definition of ranking fairness is heuristic with no rigorous proof and only preliminary results are demonstrated. Fairness constraints on rankings are formulated in [24], which uses linear programming approaches to find rankings that maximize user utility while provably satisfying a specified fairness constraint. However, this approach still needs to sample the final rankings rather than directly optimize the objective, which may be inefficient for large scale industry dataset. An auditing framework is proposed [12] to measure search engine bias. The work focuses on the identification of the sources of bias rather than the generation of a fair ranking list. A recent work [31] presents a ranked group fairness criterion based on the statistical hypothesis testing. The method can adjust a ranking list so that a minimal number of instances in protected groups must appear in the top-k list to guarantee a fair criterion. However, this post-processing algorithm and the training of ranking function are independent and thus the adjustment is limited. A variation of group parity is proposed in [18] for top-K crime hotspots prediction. The fairness loss is integrated into the likelihood of event occurrences and the model parameters are penalized to strike a balance between accuracy and fair loss. However, the fairness metric does not guarantee group parity at any point in the ranked list.

3 MOTIVATION

In this section, we first introduce the background of spatial-temporal point process model for event prediction and then present the fairness concerns on event prediction.

3.1 Ranking Prediction by Spatial-Temporal Point Process

A collection of n events in an area (e.g., a city) during a time window $[0, T]$ is represented as a temporally ordered list $\mathcal{T} = \{e_k = (x_k, t_k)\}_{k=1}^{n}$, where event e_k happens at time t_k and location x_k, e.g., a pair of longitude and latitude. An event e_k may be a crime event reported from a victim or a disaster-related rescue request. In many event prediction applications, an area is divided into grid cells or political boundaries such as ZIP Codes. For example, disaster areas can be discretized into 30m by 30m square grid cells (resolution of TM remote sensing image), 150m by 150m (size of a city block), or larger. Let $\mathcal{G}$ denote the set of grid cells and $g \in \{1, 2, \ldots, m\}$ index all m cells in $\mathcal{G}$. For each location x_k, let g_k denote the index of the cell that covers this location.

A temporal event sequence at the g^{th} grid cell can be modeled as a Hawkes process [16, 18]. The process can be characterized via its conditional intensity $\lambda_g(t)$, which models the expected rate of the event occurrences at the cell given the history of all the previous events up to time t. The conditional intensity function is:

$$\lambda_g(t) = \eta_g + \sum_{t_k < t, g_k = g} \theta \kappa_\omega(t - t_k), \tag{1}$$

where $\eta_g > 0$ is the base intensity of cell g, θ is the self-exciting coefficient, and κ_ω is the kernel function that captures the temporal intensity triggered by recent events. A common choice of kernel functions is an exponential kernel function with a bandwidth ω, i.e., $\kappa_\omega(t) = \omega \exp(-\omega t)$.

The base intensity η_g can be modeled as a function of spatial covariates/features, such as demographics [18], geological and socioeconomic variables. Let $\boldsymbol{f}_g$ denote the d-dimensional feature vector for cell g, i.e., $\boldsymbol{f}_g \in \mathbb{R}^d$. Commonly, the base intensity is

log-linear with the coefficients $\boldsymbol{\alpha}$ and the feature vector $\boldsymbol{f}_g$, that is, $\eta_g = \exp(\boldsymbol{\alpha} \cdot \boldsymbol{f}_g)$. The base intensity can be inhomogeneous through the space, which explains spatial variations of event hazards (e.g., disparate crime rates or flood hazards in different neighborhoods).

Given the observed historic event sequences $\mathcal{T}$, the model parameters can be estimated by maximizing the log-likelihood [18], or equivalently, minimizing the joint negative log-likelihood:

$$\mathcal{L}(\boldsymbol{\alpha}, \theta, \omega) = -\sum_{k=1}^{n} \log(\lambda_{g_k}(t_k)) + \sum_{g \in \mathcal{G}} \int_0^T \lambda_g(t) dt, \qquad (2)$$

where g_k is the cell index for event e_k.

To better capture correlations between multiple processes defined on different grid cells, we can incorporate spatial proximities in the form of graphs into Hawkes processes. Specifically, each grid cell is a node and two nodes are connected by an edge if they are neighborhoods. The graph is a proximity network of spatial cells. Similar to [20, 23], a graph regularization is added to enforce spatial smoothness of the intensities at each cell. Formally, the objective function of the spatial-temporal Hawkes process with graph regularization is:

$$O(\boldsymbol{\alpha}, \theta, \omega) = \min \mathcal{L}(\boldsymbol{\alpha}, \theta, \omega) + \rho\{T^{-1} \sum_{t=1}^{T} tr(\boldsymbol{\Lambda}(t)^\top L \boldsymbol{\Lambda}(t))\}, \qquad (3)$$

where $\boldsymbol{\Lambda}(t) = [\lambda_1(t), \lambda_2(t), ..., \lambda_m(t)]^\top$ is the vector of event rates at m cells during a time window, ρ is the regularization parameter, and L is the Laplacian matrix constructed on the graph.

For event forecasting, the model parameters estimated using training data can be used to compute the intensities in eq. (1) at each cell g and a given time t. A higher intensity means a larger probability that an event will happen at its corresponding location. In practice, the intensities $\boldsymbol{\Lambda}(t)$ in the list are ranked from the highest to the lowest, and top-K hotspots may be selected at time t for informing further activities such as police patrolling.

3.2 Fairness Concerns on Ranking

Our concern is that grid cells ranked by event hazard rates can make the visibility of a disadvantaged group even worse. For instance, the disadvantaged groups can be racial minorities or the communities of lower social economic status. Existing fairness metrics focus on the group fairness averaged over the entire list such that the average amount of attention received by each demographic group should be fair. However, they do not compare the group fairness at every point in the ranked list.

In table 1, we list a simple example to demonstrate that event rate prediction may exhibit bias towards certain groups. Assume there are 10 locations (e.g., grid cells), which are ranked by the predicted event rates during a given time period. Each cell is associated with 2-dimensional demographic feature and each feature indicates the population of one race in the cell. The column "Group" indicates the type of majority race for each cell. For example, 1 means that race 1 is the majority and 2 means race 2 is the majority in that cell. Specifically, for cell 1 in the first row, the predicted hazard rate is 10.0, which is the highest. There are 10.0 persons of race 1 and 1.0 person of race 2 living in the area of cell 1.

If we use the traditional fairness metrics to evaluate the entire list in table 1, it is fair. Specifically, we can see that the total numbers of population for race 1 and race 2 are the same, which is 56.0. Also, there are six cells labeled as group 2 while four labeled as group 1. However, it is not fair at every point of the list. As we can see, most of the locations on the top of the list are labeled as group 2, which means the ranker intends to rank group 2 higher than group 1. Moreover, in reality, only top ranked cells receive sufficient attentions. If we take top-5 locations into consideration, there are 80% of the locations labeled as group 2, which is also unfair for race 1. In this case, a more specific fairness criterion focusing on the entire ranking list is needed.

Table 1: Example for ranking fairness.

Cell by Rank	Intensity	Group	Race 1	Race 2
Cell 1	10.0	1	10.0	1.0
Cell 2	5.0	2	5.0	6.0
Cell 3	4.6	2	5.0	6.0
Cell 4	4.2	2	4.0	7.0
Cell 5	3.8	2	3.0	8.0
Cell 6	3.4	2	2.0	9.0
Cell 7	3.0	2	1.0	10.0
Cell 8	2.8	1	9.0	2.0
Cell 9	2.4	1	8.0	3.0
Cell 10	0.8	1	9.0	4.0

4 LIST-WISE FAIRNESS CRITERION

In this section, we introduce our **List-wise Fairness Criterion** for spatial-temporal point process. We first propose a series of definitions to describe the **unfairness consistency** property of a fairness metric and then prove that our metric satisfies this property.

4.1 Preliminaries

Let $\mathcal{G}$ be the instance space, e.g., a set of all m grid cells. Each instance is associated with some sensitive features such as races or social economic status. For simplicity, we assume there are two sensitive features, such as race type one and race type two. Let $\mathcal{Y}$ be the set of values for one race for all grid cells, and $\tilde{\mathcal{Y}}$ be the set of values for the other, respectively. The feature value of each instance indicates the relevance of the instance with respect to that feature. For example, if we define the feature as the race population in a cell, $y = 10.0$ means 10.0 population of race type one and $y = 0$ means zero population of that type. Also, a larger $y \in \mathcal{Y}$ indicates a larger representation of race type one.

At a specific time t, the intensity function $\lambda_g(t)$ can be considered as a mapping from instances $\mathcal{G}$ to $\mathbb{R}$ and be shortened as λ_g. For every instance $g \in \mathcal{G}$, we rank them by intensity function λ_g. The final ranking list is denoted by $g_{(1)}, ..., g_{(m)}$, which satisfies $\lambda_{g_{(1)}} \geq ... \geq \lambda_{g_{(m)}}$. Let $y_1, ..., y_m (y_i \in \mathcal{Y})$ and $\tilde{y}_1, ..., \tilde{y}_m (\tilde{y}_i \in \tilde{\mathcal{Y}})$ be the sensitive features associated with $g_1, ..., g_m$, respectively. Denote $S_m = \{(g_1, y_1, \tilde{y}_1), ..., (g_m, y_m, \tilde{y}_m)\}$ the set of intensities and features. Following [26], we assume that $(g_i, y_i, \tilde{y}_i)$ are i.i.d. samples taken from a distribution $P_{G\mathcal{Y}\tilde{\mathcal{Y}}}$ over $\mathcal{G} \times \mathcal{Y} \times \tilde{\mathcal{Y}}$.

The Normalized Discounted Cumulative Gain (NDCG) is a widely used list-wise ranking metric to measure the ranking quality. It

is often used to measure if web search engine algorithms rank most relevant documents at top ranks. In our case, we adopt its formula to measure the relevance of a ranked list with respect to sensitive feature values. Thus, we replace the relevance scores with the sensitive feature values in the following definition.

Definition 1. Let $P(r)(r \leq 1)$ denote a discount function on ranking positions. The intensity function λ_g is the ranker. The Discounted Cumulative Gain (DCG) of the ranker λ_g with respect to a sensitive feature $\mathcal{Y}$ using a discount function $P(r)$ is defined as:

$$DCG(\lambda_g, \mathcal{G}, \mathcal{Y}) = \sum_{r=1}^{m} y_{(r)} P(r). \tag{4}$$

We can similarly define DCG for another sensitive feature $\tilde{\mathcal{Y}}$ as $DCG(\lambda_g, \mathcal{G}, \tilde{\mathcal{Y}}) = \sum_{r=1}^{m} \tilde{y}_{(r)} P(r)$.

The ideal DCG (IDCG) is the best DCG value of any possible ranking function with respect to a sensitive feature. Specifically, for the sensitive group $\mathcal{Y}$, we have $IDCG(\mathcal{G}, \mathcal{Y}) = \max_{\lambda'_g} \sum_{r=1}^{m} y'_{(r)} P(r)$. Thus, the Normalized DCG of intensity function λ_g on S_m with discount function $P(r)$ is defined as:

$$NDCG(\lambda_g, \mathcal{G}, \mathcal{Y}) = \frac{DCG(\lambda_g, \mathcal{G}, \mathcal{Y})}{IDCG(\mathcal{G}, \mathcal{Y})}. \tag{5}$$

NDCG are normalized scores ranging from 0.0 to 1.0 and thus are cross-group comparable. A NDCG is a standard NDCG if the discounting function is chosen to be the inverse logarithm decay $P(r) = \frac{1}{\log(r+1)}$. The choice of the base of the logarithm does not affect NDCG since the normalization can cancel out constant scaling. We use the natural logarithm in this paper. It is worth mentioning that even though the discount function $P(r)$ is defined as a function of positive integers r, it can be treated as a function of non-negative real variable in the following sections. Thus, we can also consider the corresponding derivative $P'(r)$ and integral $\int P(r)dr$. In the following section, we leave out the word "standard" and directly use NDCG unless we emphasis the difference.

4.2 List-wise Fairness Criterion

We now propose our **List-wise Fairness Criterion** of ranked list with respect to sensitive features. Intuitively, we can compare the difference of NDCG scores with respect to different sensitive features (e.g., racial groups). A disparity between NDCG scores indicates a larger degree of unfairness between the racial groups. A strict definition of our **List-wise Fairness Criterion** between each pair of groups is:

$$F(\lambda_g, \mathcal{G}, \mathcal{Y}, \tilde{\mathcal{Y}}) = (NDCG(\lambda_g, \mathcal{G}, \mathcal{Y}) - NDCG(\lambda_g, \mathcal{G}, \tilde{\mathcal{Y}}))^2. \tag{6}$$

Note that an ideal **List-wise Fairness Criterion** should substantially distinguish the ranking gain with respect to two groups at any prefix of the ranking. Below we first give the formal definition that a ranker measured by a metric $\mathcal{F}$ is **consistently unfair** between two groups. The definition describe the **unfairness consistency** property of a fairness measure.

Definition 2. Let $(g_1, y_1, \tilde{y}_1), (g_2, y_2, \tilde{y}_2), \ldots$ be i.i.d. instance-label tuples taken from the underlying distribution $P_{GY\tilde{Y}}$ over $\mathcal{G} \times \mathcal{Y} \times \tilde{\mathcal{Y}}$. Given $S_m = \{(g_1, y_1, \tilde{y}_1), \ldots, (g_m, y_m, \tilde{y}_m)\}$ and intensity function λ_g as the ranker. The ranker λ_g measured by a fairness metric $\mathcal{F}$ is said to be **consistently unfair** between two groups if there exists a negligible function [1] $\mu(N)$ such that for every sufficient large N, with probability $1 - \mu(N)$,

$$\mathcal{F}(\lambda_g, \mathcal{G}, \mathcal{Y}, \tilde{\mathcal{Y}}) > 0 \tag{7}$$

holds for all $m \geq N$ simultaneously.

This definition indicates the **unfairness consistency** property by a metric measuring the rank list. We then give a theorem to show that our fairness metric indeed satisfies this property. For simplicity, we present the theorem for features with binary values, i.e., $\mathcal{Y} = \{0, 1\}$ and $\tilde{\mathcal{Y}} = \{0, 1\}$. It can be easily extended to general cases where the values of $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$ are finite sets [26].

To begin with, suppose there exist another intensity function $\hat{\lambda}_g$ that preserves the order [2] as original intensity function λ_g, then we have $\text{NDCG}(\lambda_g, \mathcal{G}, \mathcal{Y}) = \text{NDCG}(\hat{\lambda}_g, \mathcal{G}, \mathcal{Y})$ by definition. Hence, the NDCG is defined on an equivalent class of intensity functions which can preserve the same order. We now introduce the concept of canonical form.

Definition 3. Given an intensity function λ_g, we present a canonical form of λ_g as:

$$\hat{\lambda}_g = \Pr_{G \sim P_G}[\lambda_G \leq \lambda_g]. \tag{8}$$

The benefit of using the canonical form intensity function is that it satisfies the following property, which can be proven by definition.

Proposition 4. *For any intensity function λ_g, its canonical form $\hat{\lambda}_g$ preserves the order of λ_g and has uniform distribution on interval* $[0, 1]$.

Now we give the following theorem:

Theorem 5. *Given the canonical intensity function $\hat{\lambda}_g$, let $y'(s) = \Pr_{G \sim P_G}[Y = 1 \mid \hat{\lambda}_G = s]$ and $\tilde{y}'(s) = \Pr_{G \sim P_G}[\tilde{Y} = 1 \mid \hat{\lambda}_G = s]$. Assume $y'(s)$ and $\tilde{y}'(s)$ are Hölder continuous in s. Then, unless $y'(s) = \tilde{y}'$ almost everywhere on interval* $[0, 1]$*, the ranker λ_g measured by our **List-wise Fairness Criterion** is **consistently unfair** between the groups with sensitive features $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$.*

Proof. We prove our **unfairness consistency** in theorem 5 by adopting the technology provided by [26] which are used to prove the property that a measure can distinguish ranking functions. We first define the pseudo expectation $\mathcal{N}(m)$ and $\tilde{\mathcal{N}}(m)$, which are integrals to approximate the DCG, for the sensitive features $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$ respectively. We start with $\mathcal{Y}$:

Definition 6. Assume $\mathcal{Y} = \{0, 1\}$, and let $y'(s) = \Pr_{G \sim P_G}[Y = 1 \mid \hat{\lambda}_G = s]$, we define the pseudo expectation $\mathcal{N}(m)$ for the unnormalized DCG as:

$$\mathcal{N}(m) = \int_1^m y'(1 - r/m) P(r) dr = m \int_{1/m}^1 y'(1 - s) P(ms) ds, \tag{9}$$

[1] A function $\mu : \mathbb{N} \to \mathbb{R}$ is negligible iff $\forall c \in \mathbb{N}, \exists n_0 \in \mathbb{N}$ such that $\forall n \geq n_0$, $\mu(n) < n^{-c}$.

[2] Preserving the order means for $\forall g_1, g_2 \in \mathcal{G}$, $\lambda_{g_1} \geq \lambda_{g_2}$ implies $\hat{\lambda}_{g_1} \geq \hat{\lambda}_{g_2}$.

with the substitution of integration $r = ms$. Suppose that $F(x) = \int_1^x P(r)dr$ and the probability $p = \Pr[Y = 1] > 0$, we have the normalized pseudo expectation $\mathcal{E}(m)$ as $\mathcal{E}(m) = \mathcal{N}(m)/F(mp)$.

We first prove that the difference between the NDCG and its pseudo expectation is relatively small with high probability by lemma 7.

Lemma 7. *Suppose* $p = \Pr[Y = 1] > 0$ *and* $y'(s) = \Pr_{G\sim P_G}[Y = 1 \mid \hat{\lambda}_G = s]$ *is Hölder continuous*[3] *in* $s \in [0, 1]$ *with constants* $a, C > 0$. *Then*

$$\Pr[|NDCG(\lambda_g, \mathcal{G}, \mathcal{Y}) - \mathcal{E}(m)| \geq 5Cp^{-1}m^{-\min(a/3,1)}] \leq O(e^{-m^{1/4}}). \tag{10}$$

We then prove that the difference between the pseudo expectations for the NDCG of the two groups is much larger by lemma 8.

Lemma 8. *Suppose* $p = \Pr[Y = 1] > 0$ *and let* $y'(s) = \Pr_{G\sim P_G}[Y = 1 \mid \hat{\lambda}_G = s]$ *and* $\tilde{y}'(s) = \Pr_{G\sim P_G}[\tilde{Y} = 1 \mid \hat{\lambda}_G = s]$. *Then, unless* $y'(s) = \tilde{y}'$ *almost everywhere on interval* $[0, 1]$, *there must exist an integer* $K \geq 0$ *and a constant* $B \neq 0$, *so that*

$$|\mathcal{E}(m) - \tilde{\mathcal{E}}(m) - \frac{B}{\log^K m}| \leq O(\frac{1}{\log^{K+1} m}). \tag{11}$$

The proofs of lemma 7 and lemma 8 are in appendix B. Thus, from the two lemmas, and with the observation that $\sum_{m>N} e^{-m^{1/4}} \leq O(N^{3/4}e^{-N^{1/4}}) \leq O(e^{-N^{1/5}})$, the ranker λ_g measured by our **List-wise Fairness Criterion** is **consistently unfair** between two groups with high probability. □

Remark: theorem 5 provides the consistent analysis of our **List-wise Fairness Criterion**. It can consistently differentiate two group in the ranking list provided by the intensity function. Thus, we consider using it to penalize the objective function later in section 5. By minimizing our **List-wise Fairness Criterion**, the penalties affect the final intensity function to generate a fair ranking list.

It is worth mentioning that in the Standard NDCG, the inverse logarithm function is used as the discount function. If other functions such as inverse polynomial $P(r) = r^{-\beta}, \beta > 0$ are used for computing the NDCG, the **unfairness consistency** is not exactly guaranteed. Also, an inverse polynomial decay with $\beta > 1$ might not be appropriate when the list is huge, since the tail of the ranking list may be omitted in calculation.

4.3 Cut-off Version

It is usually computational inhibitive the when calculate all the instance in practice. Thus, we consider a cut-off version of our **List-wise Fairness Criterion** $F(\lambda_g, \mathcal{G}, \mathcal{Y}, \tilde{\mathcal{Y}})@k$ by using the NDCG@k with $k = cm$ for some constant $c \in (0, 1)$ in eq. (6). We also adopt the discount function $\tilde{P}(r) = \frac{1}{\log(r+1)}$ if $r \leq k$ and $\tilde{P}(r) = 0$ otherwise. Note that it is not appropriate to define k as a constant independent with list size m. The reason is the NDCG@k is bounded by the partial summation, which cannot consistently cover the total ranking list. Thus, k must grow unboundedly when m goes to infinity. In addition, by adopting $k = cm$, the **unfairness consistency** of $F(\lambda_g, \mathcal{G}, \mathcal{Y}, \tilde{\mathcal{Y}})@k$ holds under the conditions given in theorem 5. The proof is similar to its full version in theorem 5.

[3] That is, for $\forall s, s' \in [0, 1]$, $|y'(s) - y'(s')| \leq C\|s - s'\|^a$

5 LEARNING

In this section, we develop a penalized likelihood approach to incorporate fairness penalties into point process models. Trade-off between event prediction accuracy and fairness can be achieved by controlling the degree of fairness penalties in objective function.

5.1 Objective Function with List-wise Fairness Criterion

The fairness penalties based on **List-wise Fairness Criterion** for ranking grid cells with respect to sensitive groups over the total training time period $[0, T]$ is defined as follows:

$$F(\boldsymbol{\alpha}, \theta, \omega) = \frac{1}{T}\sum_{t=1}^{T}(NDCG(\lambda_g(t), \mathcal{G}, \mathcal{Y}) - NDCG(\lambda_g(t), \mathcal{G}, \tilde{\mathcal{Y}}))^2. \tag{12}$$

When $F = 0$, the ranking list with respect to the two groups achieves consistently fairness averagely over a time period.

More generally, suppose there are q types of sensitive features and the i-th type of sensitive features $\boldsymbol{f}_i$ contains c_i groups, then for $\forall l_i, l_i' \in c_i$, the total penalty is defined as follows:

$$F(\boldsymbol{\alpha}, \theta, \omega) = \frac{1}{T}\sum_{i=1}^{q}\sum_{l_i > l_i'}\sum_{t=1}^{T}(NDCG(\lambda_g(t), \mathcal{G}, \mathcal{Y}_{l_i}) - NDCG(\lambda_g(t), \mathcal{G}, \mathcal{Y}_{l_i'}))^2, \tag{13}$$

where $\mathcal{Y}_{l_i}$ is the l-th group of the i-th type sensitive features. For example, sensitive features include race and gender. There are multiple types of race and different gender. When $F = 0$, for every type of sensitive features and for each pair of feature groups, the ranker achieves consistently fairness averagely over a time period.

Finally, we add the penalty F weighted by a trade-off parameter γ to the objective function eq. (3) and minimize:

$$OPT = \min \mathcal{L}(\boldsymbol{\alpha}, \theta, \omega) + \rho\{T^{-1}\sum_{t=1}^{T} tr(\boldsymbol{\Lambda}(t)^\top L\boldsymbol{\Lambda}(t))\} + \gamma F(\boldsymbol{\alpha}, \theta, \omega). \tag{14}$$

Once we obtain $\boldsymbol{\alpha}, \theta$ and ω, we can directly calculate the intensities for all grid cells by eq. (1) and present a fair ranking list.

5.2 Optimization and Scalability

The objective function with fairness penalties defined by eq. (14) is non-differentiable since grid cells needs to be ranked by intensities and a threshold is required for selecting top-k cells at each time slot t. Thus, we adopt the Nelder-Mead simplex method [13] to find a local minimum. We show the details how we apply this method in appendix A.

It is well known that simplex method takes polynomial time complexity, i.e., $O(n^k)$ in average [25], which is computational inhibitive when the dataset is huge. Hence, we provide a smoothed variation of our method, which uses a non-linear function to approximate the rank and makes it differentiable.

As we introduced before, for the standard NDCG, we use the inverse logarithm decay $P(r) = \frac{1}{\log(r+1)}$ as the discount function.

We first rewrite the standard DCG in the following form:

$$DCG(\lambda_g, \mathcal{G}, \mathcal{Y}) = \sum_{i=1}^{m} \frac{y_i}{\log(R(i)+1)}, \tag{15}$$

where $R(i)$ is the rank position of the cell g_i by the ranker, intensity function λ_g. The DCG is non-smooth mainly because of the non-continuous mapping from the intensity score λ_{g_i} to the rank position $R(i)$. Specifically, the rank position can be defined in the following form:

$$R(i) = 1 + \sum_{j \neq i} I_{\{\lambda_{g_i} - \lambda_{g_j} < 0\}}. \tag{16}$$

To deal with this problem, we follow [19] to revise the discount function so that it becomes a continuous function of the intensities. Thus, we have the approximate rank position $\tilde{R}(i)$ as:

$$\tilde{R}(i) = 1 + \sum_{j \neq i} \frac{\exp(-\delta(\lambda_{g_i} - \lambda_{g_j}))}{1 + \exp(-\delta(\lambda_{g_i} - \lambda_{g_j}))}, \tag{17}$$

where δ is the hyper-parameter which is often set dynamically like the decay of learning rate. A larger δ leads to a better approximation of rank position. However it increase the difficulty of optimization due to the stronger degree of nonlinearity. When $\lambda_{g_j} \ll \lambda_{g_i}$, the non-linear part approaches zero, thus the position hardly changes. Integrating the approximate rank position eq. (17) into eq. (15), we obtain the smoothed DCG. The smoothed DCG can be optimized by gradient based methods, which makes the computation faster and the model scalable. Nevertheless, we have to mention that this smoothed method is *not* suitable for the cut-off version NDCG@k. In addition, the smoothed method *cannot* guarantee the **unfairness consistency** property we introduced before.

It is worth mentioning that our **List-wise Fairness Criterion** is not limited to spatial-temporal point processes, in fact, it can be extended to other ranking problems. For example, suppose we recommend a candidate list and the candidates may have sensitive features such as gender and race. Our fairness metric can be applied to either binary or finite sets of features. The computational complexity increases when the list is huge (e.g., a million). In this case, our method using the smoothed DCG can tackle the computation challenge and we can obtain an approximately fair ranking list.

6 EXPERIMENT

In this section, we introduce the experiments and results.

6.1 Data

We evaluate our list-wise fairness criterion on three open sourced real-world datasets detailed in table 2. Specifically, the **Portland** dataset [4] [17] is provided by 2017 NIJ Crime Forecasting Challenge [5] that tasks participants to predict the spatial locations with highest numbers of crime related calls in Portland, OR. It contains a list of events with geographic coordinates, timestamps, and the types of events such as burglary, street crime, and auto theft from March, 1, 2012 to February 28, 2017. In our setting, a unit time slot t is a day. Each event is assigned to one of equal sized regular rectangle grids based on the longitude and latitude. In the experiments, we only use the street event data and we simulate the race populations for white and Hispanic/Latino as the sensitive features, which is an extreme case. We first learn the model without fairness penalties to obtain a ranked list of locations, and assign the population for white as 1 to m in the order from high to low and m to 1 for the Hispanic/Latino. The **Dallas** dataset [6] in Kaggle comes from the Dallas Police Department containing detailed incidence reports for around 3 years at Dallas, Texas. We adopt the similar settings for Portland dataset to specify the locations that the events belong to. For the race population feature, we count the number of events for complainants in three races (black, white, and Hispanic/Latino) and regard them as the population of that location grid. The **Houston** dataset [7] is a crowdsourcing dataset obtained from a Google doc which contains rescue requests for 3 days around Harris County in Greater Houston Area during the Hurricane Harvey disaster. In this dataset, a time slot t is an hour and we use the ZIP Code as the location id. We get the race population statistics from American FactFinder [8]. We use the populations of white and Hispanic/Latino as the sensitive features.

Table 2: Dataset description.

Dataset	Events	Geo-Type	Unique-IDs	Time	Groups
Portland	$166K$	Grid	398	$1916d$	2
Dallas	$201K$	Grid	303	$853d$	3
Houston	1182	ZIP Code	106	$26h$	2

For Portland and Dallas datasets, we use the first 200 days for training and the days from 201 to 400 for testing given the huge number of events, while for Houston dataset we adopt the first 14 hours for training and the rest for testing. For geometric settings such as graph regularizers in eq. (14), we assume that each location is a node in a graph and adjacent location nodes are connected. In training, we use cut-off version of our list-wise fairness criterion as the fairness penalties with NDCG@50 to improve computational efficiency. Since the optimization algorithm only converges to local minimal, we run several times with different initialization and present the best results.

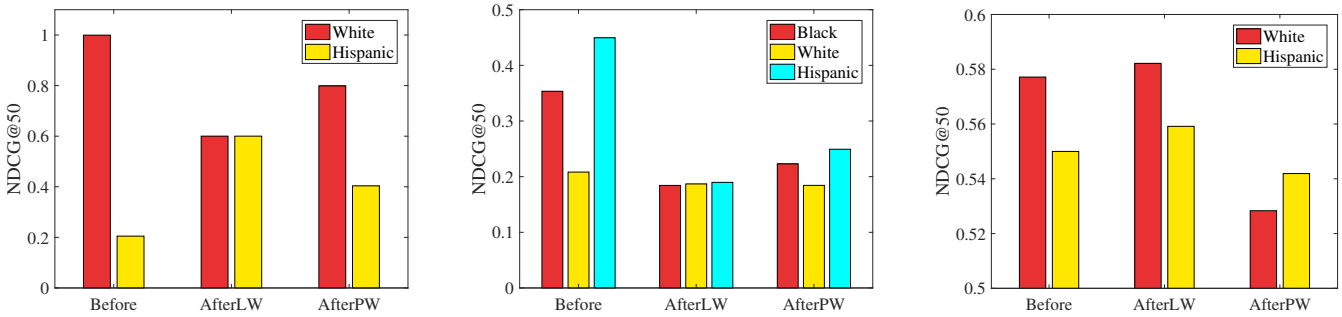
6.2 Evaluation Metrics

We use several metrics to evaluate both prediction performance and fairness influence by adopting our list-wise fairness criterion.

6.2.1 *Fairness Evaluation Metrics.*

- **NDCG@k**: We directly compare the NDCG@50 scores of different groups since we use them in training. A smaller difference indicates a fairer prediction.
- **FairLoss**: We apply the fairness penalties that we define in eq. (14) to the test data. A smaller fairness loss indicates a fairer ranking list.
- **Patrol@k**: We use the fairness metric defined in [18], which is the ratio of the summation of population in the top-k

[4] https://github.com/gomohler/crimerank
[5] https://nij.ojp.gov/funding/real-time-crime-forecasting-challenge
[6] https://www.kaggle.com/carrie1/dallaspolicereportedincidents
[7] https://data.world/sya/harvey-rescue-doc
[8] https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml



(a) Portland (b) Dallas (c) Houston

Figure 1: The NDCG@50 for different racial groups before and after adding list-wise fairness penalties.

locations over that in the total list per race group. Specifically,

$$\text{Patrol@k} = \frac{\sum_{i=1}^{k} y_{(i)}}{\sum_{i=1}^{m} y_m}, \tag{18}$$

and we just replace the NDCG with this metric and the difference between two groups should still be averaged over time slot t and types i in eq. (13). We name it **List-sum Fairness Criterion**, in contrast to our **List-wise Fairness Criterion**. In the experiments, we adopt $k = 50$ to keep uniform standard with the former setting.

6.2.2 *Prediction Performance Evaluation Metrics.*

- **Correlation**: We use the Pearson correlation coefficient between the predicted intensity list $\mathbf{\Lambda}(t)$ and the ground truth, which is the list of numbers of events at time slot t, to evaluate the prediction performance. It is between 1 and -1, where 1 means total positive linear correlation, 0 means no linear correlation, and -1 represents total negative linear correlation according to the Cauchy–Schwarz inequality.
- **TestLoss**: It is the test loss without fairness penalties in eq. (14), which is the log-likelihood of point process in eq. (2) that indicates the probability that existing history event $\mathcal{T}$ has happened and no events happen in $[t_n, t)$.
- **PAI@k**: Predictive Accuracy Index (PAI) is widely used to measure the percentage of crime events in the top-k locations [17, 18] and has the following form:

$$\text{PAI@k} = \frac{\text{events in k locations}}{\text{total events}} \cdot \frac{\text{total area}}{\text{area of k locations}}. \tag{19}$$

Since PAI@k is area normalized, a value of 1 indicates random predictions. We also apply it to the Houston rescue dataset. The value of k is chosen by the police resources or the rescue resources and we provide two choices in the experiments, PAI@15 and PAI@50.

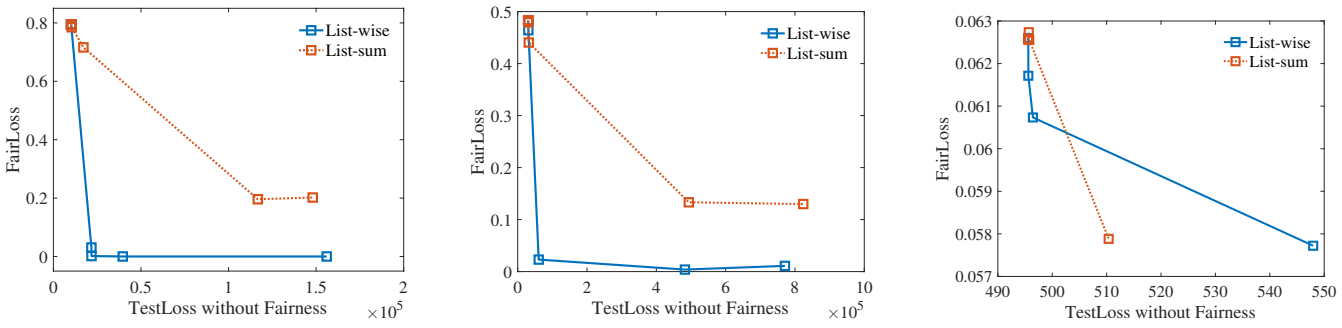
6.3 Fairness over Groups

We plot both the neutral (before adding our list-wise fairness penalties) and fair scenarios of our model by measuring NDCG@50 on test data per group over all three datasets in fig. 1. We can see that in general, this list-wise fairness criterion is effective and the differences between the groups become smaller after adding our list-wise fairness penalties, and all the scores become closer to each other to approach the ideal case with the fairness penalties close to 0. The performance on Houston dataset is not as good as the former two due to data sparsity. In particular, there are much fewer events in a little smaller number of unique ZIP Codes as described in table 2. Besides the time slot t is an hour, and thus the events/time that represents the temporal sparsity is also at a low level. Therefore, the locations most influenced by Hurricane Harvey might have much higher intensities and much more rescue requests than others. As a result, it requires a much larger fairness penalty to change the order of the ranking list. This leads to the weak performance in terms of the fairness metrics and makes it hard to balance the NDCG@50 values between two groups. In addition, for the neutral scenario, the difference of the NDCG@50 values between two groups for Houston data is relatively smaller than others, which indicates the fairness penalty unscaled with γ is relatively small. That also increases the difficulty in obtaining an extremely fair ranking list.

6.4 Fairness vs. Prediction Performance

We measure the prediction performance and present the correlation and PAI@k before and after adding our list-wise fairness penalties in table 3. A higher correlation coefficient indicates stronger correlation between the predicted intensity list and the ground truth of the number of events, which finally represents the point process model's prediction accuracy. A higher PAI@k does not reflect the ranking accuracy of the intensity list according to the definition; however, it represents a higher predicted number of independent events at top-k locations which is useful in practice with limited police and rescue resources.

From the table, we can see that at first, the prediction performance is influenced when we incorporate the fairness in objective functions. The ranking prediction performance represented by the correlation is affected to a large extent. However, either PAI@15 or PAI@50 still keeps a higher level. This demonstrates that most of the hotspots is still on the top of the predicted ranking list. It is worth mentioning that although there is a significant cost in considering the list-wise fairness, the PAI value is not only much higher than the random case, which is 1, but also potentially even more accurate than human analysis.



(a) Portland (b) Dallas (c) Houston

Figure 2: Fairness-accuracy curves for list-wise and list-sum fairness.

Table 3: Average prediction performance before and after adding list-wise fairness penalties.

Dataset	Accuracy Measure	Results	
		Before	After
Portland	PAI@15	**344.0795**	263.9681
	PAI@50	**194.2702**	95.9875
	Correlation	**0.6614**	0.0030
Dallas	PAI@15	**156.3209**	21.1041
	PAI@50	**105.2752**	16.4861
	Correlation	**0.6550**	0.1534
Houston	PAI@15	**455.4241**	400.3325
	PAI@50	**179.1580**	171.0044
	Correlation	**0.3993**	0.3367

6.5 Comparison between List-wise and List-sum Fairness

Similar to [1], we investigate the trade-off between accuracy and fairness for two different types of fairness penalties including the **List-wise Fairness** and **List-sum Fairness**. We apply these two different fairness metrics in the training stage, and adjust the trade-off parameter of the fairness γ in the range $10^s, s = [0, 1, 2, ..., 8]$. The x axis is the test loss without considering fairness penalties and it indicates the model prediction performance. A lower test loss value represents better prediction performance. The y axis is the fairness penalty based on our list-wise fairness criterion and is calculated over test data. A lower value means a fairer ranking list.

According to the results shown in fig. 2, we can see that for all the three datasets, the degree of the fairness of the model increases as the trade-off parameter γ becomes larger, resulting in a worse prediction. Also, with the same level of the fairness loss, our **List-wise Fairness** achieves better prediction performance than the **List-sum Fairness** in general. This indicates that our list-wise fairness criterion is more efficient and less costly in the optimization. In addition, note that the fairness loss for Houston data is relatively smaller than others as we described in section 6.3. List-wise fairness have resulted in consistently more efficient curves with different values of trade-off parameter γ than list-sum fairness.

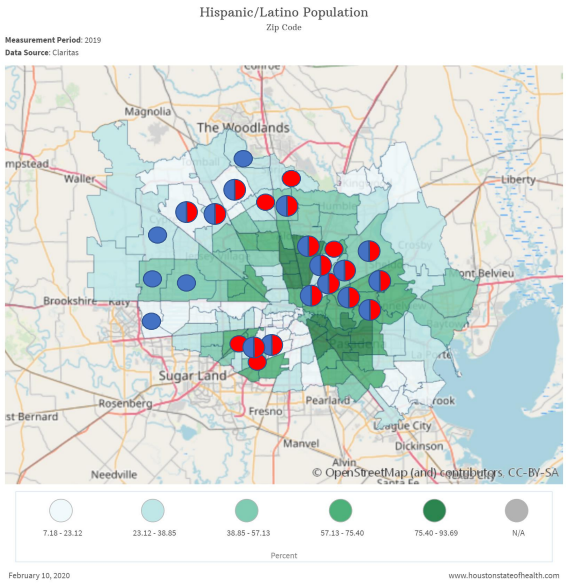


Figure 3: Case study.

6.6 Case Study

We visualize the top-20 detected hotspots before (blue) and after (red) adding our list-wise fairness penalties in fig. 3. The circle of both blue and red indicates that the location are captured in both ranking lists. The background [9] shows the population of Hispanic/Latino ranked by percentage at Harris County in the Greater Houston Area, Texas. A total of 5 locations has changed in the top-20 list and it is obvious that they switch to the locations with more Hispanic/Latino population in general. Even though the Houston dataset is sparse and it is hard to obtain a fair ranking list as we introduced in section 6.3, the results are still visible in the figure. It is worth mentioning that the east of Harris County, where several hotspots are detected in both neutral and fair ranking lists, is the worst-hit area suffering Hurricane Harvey. Since the model still

[9]Downloaded from the website: http://www.houstonstateofhealth.com/

keeps these top predicted locations, it demonstrates the effectiveness of the spatial-temporal point process in predicting the future events. Similar results are obtained on the white population map and presented in appendix D due to space limit.

7 CONCLUSION

In this paper, we present a novel list-wise fairness criterion to obtain a fair ranking list for predicting top-k locations via spatial-temporal point process. We propose a strict definition of the unfairness consistency property of a fairness metric and prove that our list-wise fairness criterion satisfies this property. Extensive experiments on the real-world datasets demonstrate the effectiveness of the list-wise fairness criterion. Future work includes extending our list-wise fairness to other fields such as scalable recommender system and developing efficient methods for fairness optimization.

8 ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation under Grant No.1927513, No. 1943486, and the Louisiana Board of Regent under Grant No. LEQSF (2017-20)-RD-A-29.

REFERENCES

[1] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

[2] Wim Bernasco, Shane D Johnson, and Stijn Ruiter. 2015. Learning where to offend: Effects of past on future burglary locations. *Applied Geography* 60 (2015), 120–129.

[3] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2212–2220.

[4] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *Proc. of the IEEE International Conference on Data Mining Workshops*. 13–18.

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proc. of the 3rd Innovations in Theoretical Computer Science Conference*. 214–226.

[6] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 259–268.

[7] Eric W Fox, Martin B Short, Frederic P Schoenberg, Kathryn D Coronges, and Andrea L Bertozzi. 2016. Modeling e-mail networks and inferring leadership using self-exciting point processes. *J. Amer. Statist. Assoc.* 111, 514 (2016), 564–584.

[8] Andrew M Freed. 2005. Earthquake triggering by static, dynamic, and postseismic stress transfer. *Annu. Rev. Earth Planet. Sci.* 33 (2005), 335–367.

[9] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*. 3315–3323.

[10] Zubin Jelveh and Michael Luca. 2014. Towards diagnosing accuracy loss in discrimination-aware classification: An application to predictive policing. *Fairness, Accountability and Transparency in Machine Learning* 26, 1 (2014), 137–141.

[11] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *Proc. of the IEEE 11th International Conference on Data Mining Workshops*. 643–650.

[12] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 417–432.

[13] Jeffrey C Lagarias, James A Reeds, Margaret H Wright, and Paul E Wright. 1998. Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on optimization* 9, 1 (1998), 112–147.

[14] Scott Linderman and Ryan Adams. 2014. Discovering latent network structure in point process data. In *Proc. of the International Conference on Machine Learning (ICML)*. 1413–1421.

[15] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830* (2015).

[16] George Mohler, Jeremy Carter, and Rajeev Raje. 2018. Improving social harm indices with a modulated Hawkes process. *International Journal of Forecasting* 34, 3 (2018), 431–439.

[17] George Mohler, Michael D Porter, Jeremy Carter, and Gary LaFree. 2018. Learning to rank spatio-temporal event hotspots. In *Proceedings of the 7th international workshop on urban computing*.

[18] George Mohler, Rajeev Raje, Jeremy Carter, Matthew Valasik, and Jeffrey Brantingham. 2018. 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)*. 2454–2459.

[19] Tao Qin, Tie-Yan Liu, and Hang Li. 2010. A general approximation framework for direct optimization of information retrieval measures. *Information retrieval* 13, 4 (2010), 375–397.

[20] Nikhil Rao, Hsiang-Fu Yu, Pradeep K Ravikumar, and Inderjit S Dhillon. 2015. Collaborative filtering with graph information: Consistency and scalable methods. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*. 2107–2115.

[21] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. 2017. When worlds collide: Integrating different counterfactual assumptions in fairness. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*. 6414–6423.

[22] Jin Shang and Mingxuan Sun. 2018. Local Low-Rank Hawkes Processes for Temporal User-Item Interactions. In *Proc. of the IEEE International Conference on Data Mining (ICDM)*. 427–436.

[23] Jin Shang and Mingxuan Sun. 2019. Geometric Hawkes Processes with Graph Convolutional Recurrent Neural Networks. In *Proc. of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4878–4885.

[24] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2219–2228.

[25] Daniel A Spielman and Shang-Hua Teng. 2004. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)* 51, 3 (2004), 385–463.

[26] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG type ranking measures. In *Proc. of the Conference on Learning Theory (COLT)*. 25–54.

[27] Maximilian J Werner, Agnès Helmstetter, David D Jackson, and Yan Y Kagan. 2011. High-resolution long-term and short-term earthquake forecasts for California. *Bulletin of the Seismological Society of America* 101, 4 (2011), 1630–1648.

[28] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 1–6.

[29] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*. 2921–2930.

[30] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proc. of the International World Wide Web Conference (WWW)*. 1171–1180.

[31] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proc. of the ACM International Conference on Information and Knowledge Management (CIKM)*. 1569–1578.

[32] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proc. of the International Conference on Machine Learning (ICML)* (Atlanta, GA, USA). 325–333.

[33] Lei Zou, NSN Lam, Shayan Shams, Heng Cai, Michelle A Meyer, Seungwon Yang, Kisung Lee, Seung-Jong Park, and Margaret A Reams. 2018. Social and geographical disparities in Twitter use during Hurricane Harvey. *International Journal of Digital Earth* (2018), 1–19.

A EXPERIMENT SETTING

We apply the Nelder-Mead simplex method in *MATLAB*® by using the function "fminsearch" [10], which can find local minimum of unconstrained multivariable functions using derivative-free method. Specifically, the code contains three parts: a script file that set all the variables and initials such that apply the "fminsearch" over objective function; a function file that is exactly the objective; and another function file calculates the log-likelihood, graph regularizer and the fairness penalties. We initialize the parameter as $\boldsymbol{\alpha} = \mathbf{0}$, $\theta = 0.8$, and $\omega = e^{-2}$ for all the datasets. We set geometric trade-off parameter $\rho = 1$ and vary the fairness trade-off parameters as $\gamma = 10^s$, $s = [0, 1, 2, ..., 8]$. We follow [18] to define the *fair* model learned with $\gamma = 10^8$ and the neutral model without fairness penalties($\gamma = 0$). The other experiments settings about datasets are introduced in section 6.1.

We show the convergence curve of the algorithm in fig. 4 on Portland Dataset. We can see that the method works well and the fairness loss, correlation and the value of objective finally stably converge to a local minimal.

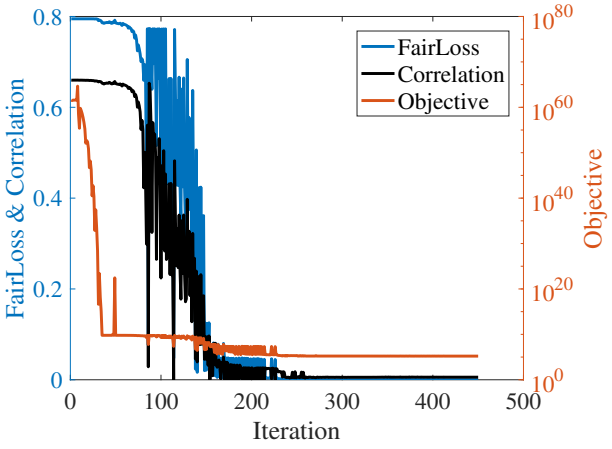


Figure 4: Convergence of the Algorithm on Portland Dataset.

B PROOFS OF THE LEMMAS

In this section, we present the proofs of lemma 7 and lemma 8. We first give two key claims that are useful to prove lemma 7, and then provide the proof of lemma 8. The proofs of claims are in appendix C.

B.1 Proof of Lemma 7

We first give two key claims.

Claim 9. *Suppose that* $F(x) = \int_1^x P(r)dr$ *and the probability* $p = \Pr[Y = 1] > 0$. *For any sufficiently large m, the following inequality*

$$|NDCG(\lambda_g, \mathcal{G}, \mathcal{Y}) - \frac{DCG(\lambda_g, \mathcal{G}, \mathcal{Y})}{F(mp)}| \leq O(m^{-1/3}), \tag{20}$$

holds with probability $(1 - 2e^{-2n^{1/3}})$.

[10] https://www.mathworks.com/help/matlab/ref/fminsearch.html

Claim 10. *Suppose that* $F(x) = \int_1^x P(r)dr$ *and* $y'(s) = \Pr_{G\sim P_G}[Y = 1 \mid \hat{\lambda}_G = s]$ *is Hölder continuous in* $s \in [0, 1]$ *with constants* $a, C > 0$. *Then,*

$$|\sum_{r=1}^{m} y'(1 - r/m)P(r) - \mathcal{N}(m)| \leq Cm^{-a/3}F(m) + 10. \tag{21}$$

Proof. Let $\mathcal{G}$ be the instance space and $g_1, ..., g_m (g_i \in \mathcal{G})$ be the m locations i.i.d. drawn from underlying distribution P_G. Let $x_{(r)} = \hat{\lambda}_{g_{(r)}}$ and we have $x_{(1)} \geq x_{(2)} \geq ... \geq x_{(m)}$ by definition. According to the Chernoff bound which is a special case of Bernstein inequalities, for each r we have $|x_{(r)} - (1 - r/m)| > m^{-1/3}$ with probability $Q = 2e^{-2m^{1/3}}$. Then, a union bound over r yields

$$\Pr[\forall r \in [m], |x_{(r)} - (1 - r/m)| \leq n^{-1/3}] \geq 1 - mQ. \tag{22}$$

Since $y'(s)$ is Hölder continuous in $s \in [0, 1]$ with constants $a, C > 0$, we have:

$$\Pr[|\sum_{r=1}^{m}(y'(x_{(r)})P(r) - y'(1-r/m)P(r))| \leq Cm^{-a/3}\sum_{r=1}^{m}P(r)] \geq 1-mQ. \tag{23}$$

Considering claim 10 and eq. (23) together, we obtain:

$$\Pr[|\sum_{r=1}^{m} y'(x_{(r)})P(r) - \mathcal{N}(m)| \leq 2Cm^{-a/3}F(m) + 10] \geq 1 - mQ. \tag{24}$$

Considering the fact that $y'(s) = \Pr_{G\sim P_G}[Y = 1 \mid \hat{\lambda}_G = s] = \mathbb{E}[Y|\hat{\lambda}_G = s]$, hence $\sum_{r=1}^{m} y'(x_{(r)})P(r)$ is the expectation of the $\mathrm{DCG}(\lambda_g, \mathcal{G}, \mathcal{Y}) = \sum_{r=1}^{m} y_{(r)}P(r)$ conditioned on $x_{(1)}, ..., x_{(m)}$. Note that conditioning on $x_{(1)}, ..., x_{(m)}$, $y_{(r)}(r = 1, ...m)$ are independent. Thus, since that $g_1, ...g_m$ are arbitrary and for $\forall r, (P(r))^2 \leq P(r)$, by applying Hoeffding's inequality which is another special case of Bernstein inequalities, we have for $\forall \epsilon > 0$,

$$\Pr[|\mathrm{DCG}(\lambda_g, \mathcal{G}, \mathcal{Y}) - \sum_{r=1}^{m} y'(x_{(r)})P(r)| \geq \epsilon] \leq 2\exp(-\frac{2\epsilon^2}{F(m)}). \tag{25}$$

Let $\epsilon = F(m)^{2/3}$ and combine eq. (24) and eq. (25), we obtain

$$\Pr[|\mathrm{DCG}(\lambda_g, \mathcal{G}, \mathcal{Y})-\mathcal{N}(m)|>2Cm^{-a/3}F(m)+2F(m)^{2/3}] \leq mQ+2e^{-2F(m)^{1/3}}. \tag{26}$$

Thus,

$$\Pr[|\frac{\mathrm{DCG}(\lambda_g, \mathcal{G}, \mathcal{Y})}{F(mp)}-\mathcal{N}(m)|\geq 4Cp^{-1}m^{-\min(a/3,1)}] \leq mQ+2e^{-2F(m)^{1/3}}, \tag{27}$$

and the lemma 7 is proved by combining claim 9 and eq. (27). □

B.2 Proof of Lemma 8

We first quote two propositions from [26].

Proposition 11. *(Claim 29 at [26]) Given a fixed integer* $k \in \mathbb{N}^* = \{0\} \cup \mathbb{N}$. *For any sufficiently large n,*

$$\int_{\frac{2}{n}}^{1} \frac{|\log^k x|dx}{(\log(nx))^{k+1}} \leq O(\frac{1}{\log^{k+1} n}), \tag{28}$$

and

Proposition 12. *(Claim 30 at [26]) Span($\{\log^k x\}_{k\geq 0}$), is dense in $L^2[0,1]$.*

Proof. Let $\Delta y'(s) = y'(s) - \tilde{y}'(s)$. Note that $F(mp) = Li(mp+1)$, where $Li(\cdot)$ is the offset logarithmic integral function and has the property $Li(n) \sim \frac{n}{\log n}$. Hence, given the normalized pseudo expectation $\mathcal{E}(m)$ in definition 6 and the observation that $|\Delta y'(s)| \leq 1$, we obtain:

$$\begin{aligned}\mathcal{E}(m) - \tilde{\mathcal{E}}(m) &= \frac{m}{Li(mp+1)} \int_{\frac{1}{m}}^{1} \frac{\Delta y'(1-s)ds}{\log(1+ms)} \\ &= \frac{m}{Li(mp+1)} \int_{\frac{2}{m}}^{1} \frac{\Delta y'(1-s)ds}{\log(1+ms)} + O(\frac{1}{Li(m)}). \end{aligned} \tag{29}$$

By expanding $\frac{1}{\log(1+ms)}$ at ms, we have:

$$|\int_{\frac{2}{m}}^{1} \frac{\Delta y'(1-s)ds}{\log(1+ms)} - \int_{\frac{2}{m}}^{1} \frac{\Delta y'(1-s)ds}{\log m + \log s}| \leq \int_{\frac{2}{m}}^{1} \frac{ds}{ms\log^2(ms)} \leq O(\frac{\log m}{m}), \tag{30}$$

and by expanding $\frac{1}{\log m + \log s}$ at $\log m$, we obtain the following

$$\begin{aligned}&|\int_{\frac{2}{m}}^{1} \frac{\Delta y'(1-s)ds}{\log m + \log s} - \sum_{z=1}^{u} \frac{(-1)^{z-1}}{\log^z m} \int_{\frac{2}{m}}^{1} \Delta y'(1-s)\log^{z-1} s ds| \\ =&|\int_{\frac{2}{m}}^{1} \frac{\Delta y'(1-s)\log^u s ds}{(\log m + \varepsilon_{m,s})^{u+1}}| \leq \int_{\frac{2}{m}}^{1} \frac{|\Delta y'(1-s)\log^u s|ds}{(\log m + \log s)^{u+1}} \leq O(\frac{1}{\log^{u+1} m}) \end{aligned} \tag{31}$$

holds for $\forall u \in \mathbb{N}^*$, where $\varepsilon_{m,s} \in (\log s, 0)$ and we obtain the last inequality by proposition 11. Also, by proposition 12 we known that unless $\Delta y'(s) = 0$ almost everywhere, there exist a constant $k \in \mathbb{N}^*$ and a non-zero constant B so that

$$(-1)^k \int_0^1 \Delta y'(1-s)\log^k s ds = 0. \tag{32}$$

Assume K is the smallest k satisfying eq. (32) and note that

$$\int_0^{\frac{2}{n}} \log^k x dx = k! \sum_{i=0}^{k} (-1)^{k-i} \frac{x\log^i x}{i!}\Big|_0^{\frac{2}{n}} = O(\frac{\log^k n}{n}), \tag{33}$$

we finally have the following inequality by combining all the equations above:

$$|\mathcal{E}(m) - \tilde{\mathcal{E}}(m) - \frac{B}{\log^K m}| \leq O(\frac{\log^K m}{m}) + O(\frac{1}{\log^{K+1} m}), \tag{34}$$

and that completes the proof of lemma 8. □

C PROOFS OF THE CLAIMS

C.1 Proof of Claim 9

Proof. Let $w = \sum_{(g_i,y_i)}^{m} I[y_i = 1]$ represent the number of $y_i = 1$ in the dataset. Considering i.i.d. sampling and the definition $\Pr[Y=1] = p$, by Chernoff bound we obtain:

$$\Pr[|w/m - p| > m^{-1/3}] \leq 2e^{-2m^{1/3}}. \tag{35}$$

Hence, with probability larger than $1 - 2e^{-2m^{1/3}}$, we have:

$$|\text{NDCG}(\lambda_g,\mathcal{G},\mathcal{Y}) - \frac{\text{DCG}(\lambda_g,\mathcal{G},\mathcal{Y})}{F(mp)}| \leq |\frac{\text{DCG}(\lambda_g,\mathcal{G},\mathcal{Y})}{w} - \frac{\text{DCG}(\lambda_g,\mathcal{G},\mathcal{Y})}{F(mp)}| \leq \text{DCG}(\lambda_g,\mathcal{G},\mathcal{Y}) \cdot \max(|\frac{1}{F(m(p-m^{-1/3}))} - \frac{1}{F(mp)}|, |\frac{1}{F(m(p+m^{-1/3}))} - \frac{1}{F(mp)}|). \tag{36}$$

Based on the observation that $\text{DCG}(\lambda_g,\mathcal{G},\mathcal{Y}) \leq F(m)$ and the Taylor expansion of $\frac{1}{F(m(p\pm m^{-1/3}))}$ at mp, claim 9 is proved. □

C.2 Proof of Claim 10

Proof. Based on the fact that $|P'(r)|$ and $P(r)$ are monotone decreasing functions and $P(1) + |P'(1)| < 10$, we have:

$$\begin{aligned}&|\sum_{r=1}^{m} y'(1-r/m)P(r) - \mathcal{N}(m)| = |\sum_{r=1}^{m} y'(1-r/m)P(r) - \int_1^m y'(1-s/m)P(s)ds| \\ =&|\sum_{r=1}^{m-1} \int_r^{r+1} \left(y'(1-r/m)P(r) - y'(1-s/m)P(s)\right) ds| + y'(0)P(m) \\ \leq&|\sum_{r=1}^{m-1} \int_r^{r+1} y'(1-s/m)(P(r) - P(s))ds| \\ &+ \sum_{r=1}^{m-1} \int_r^{r+1} |y'(1-r/m) - y'(1-s/m)|P(r)ds + y'(0)P(m) \\ \leq& \sum_{r=1}^{m-1} \int_r^{r+1} |P(r) - P(s)|ds + Cm^{-a/3} \sum_{r=1}^{m-1} P(r) + P(m) \\ \leq& \sum_{r=1}^{m-1} |P'(r)| + Cm^{-a/3}F(m) + P(m) \leq Cm^{-a/3}F(m) + |P'(1)| + \sum_{r=2}^{m} |P'(r)| + P(m) \\ \leq& Cm^{-a/3}F(m) + |P'(1)| + P(1) - P(m) + P(m) \leq Cm^{-a/3}F(m) + 10. \end{aligned} \tag{37}$$

□

D ADDITIONAL CASE STUDY FIGURE