Interpretable Hawkes Process Spatial Crime Forecasting with TV-Regularization

Hao Sha
Comp. and Info. Science
IUPUI
Indianapolis, USA
haosha@iupui.edu

Mohammad Al Hasan
Comp. and Info. Science
IUPUI
Indianapolis, USA
alhasan@iupui.edu

Jeremy Carter
Public and Env. Affairs
IUPUI
Indianapolis, USA
carterjg@iupui.edu

George Mohler
Comp. and Info. Science
IUPUI
Indianapolis, USA
gmohler@iupui.edu

***Abstract*—Interpretable models for criminal justice forecasting are desirable due to the high-stakes nature of the application. While interpretable models have been developed for individual level forecasts of recidivism, interpretable models are lacking for the application of space-time crime hotspot forecasting. Here we introduce an interpretable Hawkes process model of crime that allows forecasts to capture near-repeat effects and spatial heterogeneity while being consumable in the form of easy-to-read score cards. For this purpose we employ penalized likelihood estimation of the point process with a total-variation regularization that enforces the triggering kernel to be piecewise constant. We derive an efficient expectation-maximization algorithm coupled with forward backward splitting for the TV constraint to estimate the model. We apply our methodology to synthetic data and space-time crime data from Indianapolis. The TV-Hawkes process achieves similar accuracy to standard Hawkes process models of crime while increasing interpretability and transparency.**

I. Introduction

Because of the high-stakes nature of criminal justice forecasting, interpretable models are desirable so that users can understand why algorithmic decisions are being made. Model transparency facilitates critical assessment and may help highlight areas where the model should be further assessed in terms of fairness [6], [32]. For individual-level forecasts (e.g. recidivism, parole), mixed-integer programming has been used to create integer score cards where a low number of features are combined with integer weights to score risk [22], [29]. In Table I we present an example scorecard for recidivism from the seminal papers [22], [29]. For example, if an incarcerated individual has 3 prior arrests and their age is 30, then the score would be $1 + 1 = 2$, indicating a moderate to high level of risk of recidivism (on a scale of -1 to 4). The model utilizes three principles that we adopt in this paper:

1) Features should be easy to interpret, preferably binary indicator variables (something is or is not true).
2) Weights determining a score should be simple and preferably integers for easier human interpretation.
3) A minimal number of features determining should be used while still achieving an acceptable level of accuracy.

Spatial crime hotspotting and forecasting models, on the other hand, assign risk to places and times rather than individuals. These models can then be used by law enforcement or other stake-holders to allocate resources for crime prevention [16], [19], community policing [1], collective efficacy initiatives [21], or other interventions. Point processes are one type of modeling approach for space-time crime forecasting that allow for estimation of near-repeat effects [15], exogenous clustering [7], time of day patterns [28], incorporation of spatial covariates [12], [20], and point processes themselves can be used as features in machine learning base forecasts [13]. Two of the top performing models in the 2017 National Institute of Justice crime forecasting competition were point process models [13], [7].

Prior arrests ≥ 2	1 point	
Prior arrests ≥ 5	1 point	$+ \ldots$
Prior arrests for local ordiance	1 point	$+ \ldots$
Age at release $\in [18, 24]$	1 point	$+ \ldots$
Age at release ≥ 40	-1 point	$+ \ldots$
	Total	$= \ldots$

TABLE I: **Individual-level interpretable model** of recidivism presented in [22].

Did crimes occur in the past	1 days?	(# $\times$ 3 points)	$\ldots$
	2-3 days?	(# $\times$ 2 points)	$+ \ldots$
	4-30 days?	(# $\times$ 1 points)	$+ \ldots$
Did a crime ever occur?		(2 points)	$+ \ldots$
		Total	$= \ldots$

TABLE II: **Interpretable spatial crime forecast** score card. Each spatial grid cell in a city is scored using the card and the grid cells with the highest score are flagged as hotspots. Number of events in a time interval is denoted by # in the score card.

While interpretable forecasting at the individual level has been well-studied in the past several years [33], [30], [3], currently there are no interpretable versions of point process based spatial forecasting models in criminal justice. Here we propose to address this gap in research. We introduce an interpretable Hawkes point process model for crime forecasting that allows for the creation of easy-to-read score cards analogous to those in [29], [22]. For this purpose we use maximum penalized likelihood estimation and penalize the Hawkes process triggering kernel using the total-variation norm. Example output of the model is shown in Table VII.

The interpretable score card assigns an integer score to each spatial grid cell (or alternatively patrol beat or street segment) in a city on a given day. The score is broken down into a simple calculation that allows the end user to see how the score was created. For example, if a grid cell had 4 crimes in the past 2-3 days and 7 crimes in the past 4-30 days then the score would be $4 \cdot 2 + 7 \cdot 1 + 2 = 17$.

The outline of the paper is as follows. In Section II we provide the details of our TV-Hawkes process methodology. We derive an expectation-maximization algorithm for model estimation where within each EM iteration we solve a TV-penalized Poisson regression using forward backward splitting. In Section III we conduct several experiments on synthetic and real data. We show that the interpretable Hawkes process model achieves similar accuracy to standard non-interpretable Hawkes processes while providing a higher degree of model transparency. In Section IV we discuss directions for future work in the area of intepretable spatial crime forecasting.

II. Methodology

A. EM-FASTA Single-cell Model

We consider a Hawkes process model of crime with intensity:

$$\lambda(t) = \mu + \sum_{t>t_i} g(t - t_i). \tag{1}$$

Here $\lambda(t)$ is decomposed into a baseline intensity (rate) μ that models spontaneous events, along with a sum of intensities $g(t - t_i)$ that model repeat effects (e.g. the elevated risk following each crime at time t_i). The Hawkes process can be viewed as a branching process where spontaneous events occur according to a Poisson process with rate μ and then each event triggers a generation of offspring events with Poisson intensity $g(t - t_i)$.

The log-likelihood of the Hawkes process in Equation 1 is given by:

$$\begin{aligned} L &= \sum_{i=1}^{n} \log \lambda(t_i) - \int_0^T \lambda(t)dt - \mu_{TV}\|\nabla g\| \\ &= \sum_{i=1}^{n} \log(\mu(t_i) + \sum_{t_i>t_j} g(t_i - t_j)) - \\ &\int_0^T \mu(t)dt - \int_0^T \sum_{t>t_i} g(t - t_i)dt - \mu_{TV}\|\nabla g\| \\ &= f - \mu_{TV}\|\nabla g\|, \end{aligned} \tag{2}$$

where μ_{TV} is the regularization for the total variational constraint $\|\nabla g\| = \int_0^\infty |\nabla g(t)|dt$. Here we let f represent the non-penalized Hawkes process likelihood. We optimize Eq. 2 using forward-backward splitting, where the forward part is solved by the EM method described below and the backward part is solved using FASTA. For the triggering kernel $g(t)$ we discretize time and define $g(t_i - t_j) = g_m$ such that $m\delta t \leq (t_i - t_j) < (m + 1)\delta t$. Here the TV norm $\|\nabla g\|$ is then discretized as in [14], [8]:

$$\|\nabla g\| \approx \sum_m |g_m - g_{m-1}|. \tag{3}$$

Given the branching process representation of the Hawkes process, we can let $p_{i,i}$ be the probability that event i is spontaneous, and $p_{i,j}$ be the probability of event i being triggered by event j [15]. The expected complete un-penalized data log-likelihood is then given by:

$$\begin{aligned} E_p(L) = \quad & \sum_{i=1}^{n} p_{i,i} \log(\mu^k) - \mu^k T + \\ & \sum_{i=2}^{n} \sum_{j=1}^{i-1} p_{i,j} \log(g^k(t_i - t_j)) - \\ & \sum_{i=1}^{n} \int_{t_i}^{T} g^k(t - t_i)dt. \end{aligned} \tag{4}$$

With an initial guess for the parameter values of the model, expectation-maximization then proceeds by alternating at iteration k between the

E-step::

$$\begin{aligned} p_{i,i} &= \frac{\mu^k}{\mu^k + \sum_{j=1}^{i-1} g^k(t_i - t_j)} \\ p_{i,j} &= \frac{g^k(t_i - t_j)}{\mu^k + \sum_{j=1}^{i-1} g^k(t_i - t_j)}. \end{aligned} \tag{5}$$

and the **M-step:**

With g_m^k denoting the kernel at iteration k for the interval $[m\delta t, (m+1)\delta t)$, maximizing the complete data log-likelihood involves solving:

$$\begin{aligned} \frac{\partial E_p(L)}{\partial \mu^k} &= \frac{\sum_{i=1}^{n} p_{i,i}}{\mu^k} - T = 0 \\ \frac{\partial E_p(L)}{\partial g_m^k} &= \frac{\alpha_m}{g_m^k} - \beta_m \delta t = 0 \end{aligned} \tag{6}$$

where

$$\begin{aligned} \alpha_m &= \sum_{i=2}^{n} \sum_{j=1}^{i-1} p_{i,j} \mathbb{1}(m\delta t \leq t_i - t_j < (m+1)\delta t) \\ \beta_m &= \sum_{i=1}^{n} \mathbb{1}(T - t_i \geq m\delta t) \end{aligned} \tag{7}$$

and $\mathbb{1}$ is the indicator function. The M-step update rules are then:

$$\mu^{k+1} = \frac{\sum_{i=1}^{n} p_{i.i}}{T} \tag{8}$$

and

$$\begin{aligned} g_m^{k+1} &= \frac{\alpha_m}{\beta_m \delta t} \\ &= \frac{\sum_{i=2}^{n} \sum_{j=1}^{i-1} p_{i,j} \mathbb{1}(m\delta t \leq t_i - t_j < (m+1)\delta t)}{\sum_{i=1}^{n} \mathbb{1}(T - t_i \geq m\delta t)\delta t} \end{aligned} \tag{9}$$

We then satisfy the total-variation constraint as follows. Forward-backward splitting alternates between a forward gradient ascent step on the un-penalized log-likelihood

$$\hat{g}^{k+1} = g^k + \delta t \nabla f \tag{10}$$

and then a backward proximal gradient step:

$$g^{k+1} = \underset{g}{\text{prox}}(\hat{g}^{k+1}, \tau) = \underset{g}{\text{argmin}}\, \tau \nabla g + \frac{1}{2}||g - \hat{g}^{k+1}||^2 \tag{11}$$

where τ is the backward gradient stepsize. The forward step in Eq. 10 is equivalent to the EM step in Eq. 9. We can solve the backward step (Eq. 11) as is done in [8] using FASTA.

Putting the EM algorithm and the FBS splitting together, the overall algorithm is given by:

- **E-step** Estimate $p_{i,i}$ and $p_{i,j}$ using Eq 5.
- **M-step** Update μ using Eq 8. Estimate $\hat{g}^{k+1}$ using Eq 9. Update g^{k+1} by solving Eq. 11 using FASTA.

B. Multi-cell Model

In the above we considered the Hawkes process at a single spatial location. We can then discretize a spatial domain into grid cells of fixed size and estimate an intensity within each cell. We denote μ_l as the base intensity for cell l and we assume g_m is the same for every cell (though the kernel's contribution will vary because it depends on the event history in each cell). The EM algorithm for a multi-cell model is then given as:

E-step

$$\begin{aligned} p^l_{i,i} &= \frac{\mu^k_l}{\mu^k_l + \sum_{j=1;i,j\in S_l}^{i-1} g^k(t_i - t_j)} \\ p^l_{i,j} &= \frac{g^k(t_i - t_j)}{\mu^k_l + \sum_{j=1;i,j\in S_l}^{i-1} g^k(t_i - t_j)} \end{aligned} \tag{12}$$

where S_l is the set of indices for events in cell l and $g^k(t_i - t_j) = g^k_m$ such that $m\delta t \le (t_i - t_j) < (m+1)\delta t$.

M-step

$$\mu^{k+1}_l = \frac{\sum_{i=1}^n p^l_{i.i}}{T} \tag{13}$$

and

$$g^{k+1}_m = \frac{\sum_{l=1}^L \sum_{i,j\in S_l} p^l_{i,j} \mathbb{1}(t_i - t_j \in [m\delta t, (m+1)\delta t))}{\sum_{l=1}^L \sum_{i=1}^n \mathbb{1}(T - t_i \ge m\delta t)\mathbb{1}(i \in S_l)} \tag{14}$$

where L is the total number of cells.

III. Experiments

A. Synthetic Data

We first examine the TV-Hawkes process model on a synthetic dataset generated by a Hawkes process with exponential kernel $g(t) = \alpha w e^{-wt}$. In particular, we perform Ogata thinning simulation [17] with base intensity $\mu = 10.0$, $\alpha = 0.5$, and $w = 2.0$ in a time window $[0, T)$ where $T = 100.0$, generating a sequence of 1848 events.

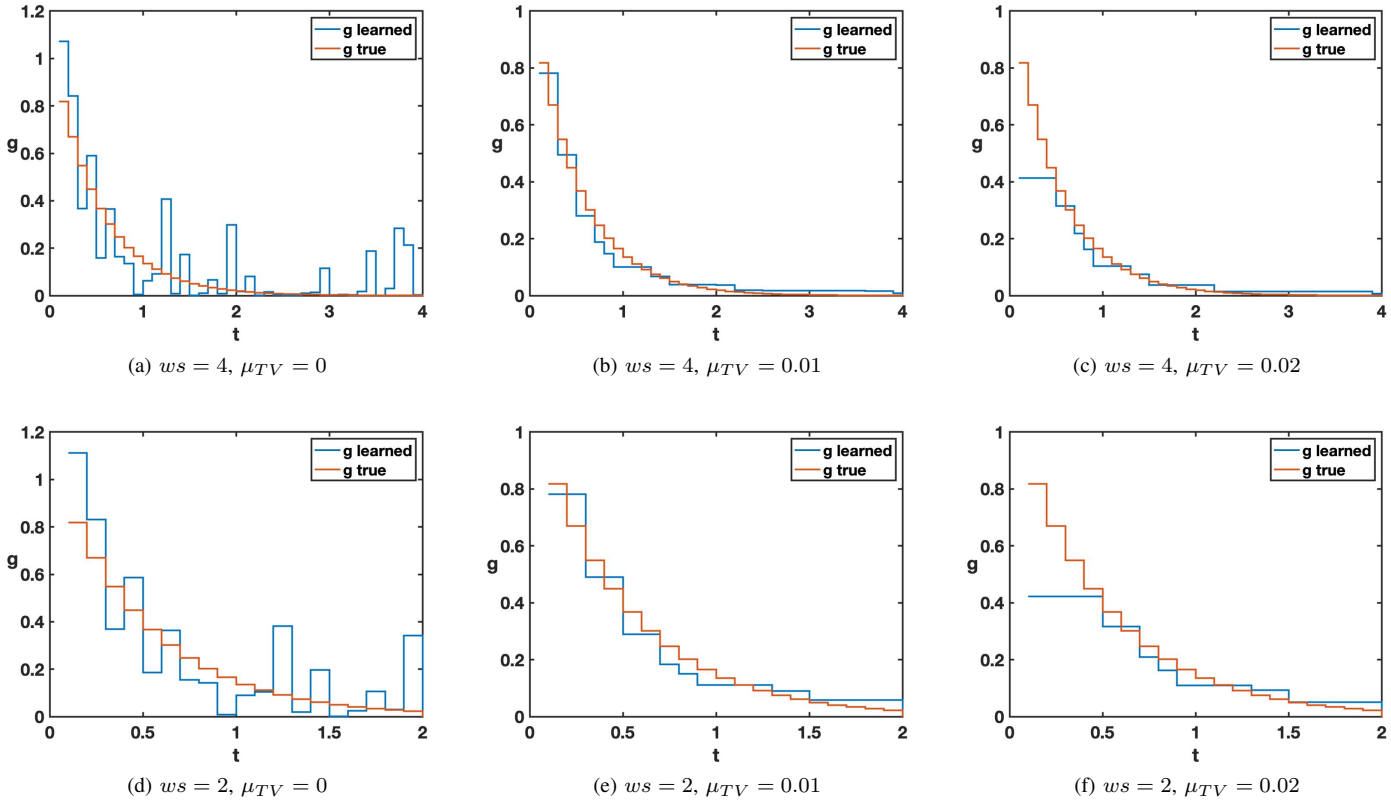
TABLE III: Fitting base intensity μ with different regularity strength μ_{TV}. ws is the cutoff length for the estimated kernel g_m. Time step $\delta t = 0.1$. μ_{TV} is the regularizer of the TV constraint. $\mu_{TV} = 0$ indicates no TV constraint. μ is the estimated base intensity, while the true μ is 10.

ws	4			2		
μ_{TV}	0	0.01	0.02	0	0.01	0.02
μ	7.81	**10.06**	11.53	9.19	10.39	11.83

Upon fitting the models, we test both truncating the kernels at window sizes of 4 and 2 (i.e. $ws = 4$ or 2 in Table III). We also vary the strength of the TV-constraint, with $\mu_{tv} = 0$, 0.01, and 0.02. We note that the TV-Hawkes estimate approaches the un-regularized Hawkes model when $\mu_{TV} = 0$. In the experiments, we adopt a fix time step $\delta t = 0.1$. The estimated μ values are shown in Table III. We can see that the models with a TV-constraint are in general better than the ones without the constraint, in terms of fitting the base intensity. In particular, when $ws = 4$ and $\mu_{TV} = 0.01$, the TV-Hawkes model gives $\mu = 10.06$, closely recovering the true $\mu = 10.0$. We also plot the estimated kernels alongside the true kernels in Fig. 1. In the left-most column, without a TV-constraint, the fitted curves exhibit some unrealistic peaks significantly higher than the true curves. In contrast, the kernels with a TV-constraint do not have such peaks and fit the true kernels very well (especially the middle column in Fig. 1). We can see that the TV-constraint effectively merges consecutive steps that are close in intensity. As a by-product, it also prevents a bin from being significantly different from its neighbors. In addition, we observe that when a stronger μ_{TV} is used, the fitted models tend to have a greater base intensity μ (Table III) but a lower first stair in the kernels (the right column in Fig. 1).

B. IMPD Crime Data

Next we apply the TV-Hawkes methodology to reported crime incidents in Indianapolis where each event consists of a date and geolocation (latitude and longitude). The dataset contains 53771 burglary, robbery and vehicle theft events and covers the time period of January 1, 2012 to December 31, 2015. We divide the Indianapolis metropolitan area into a grid, using boxes of 150×150 m as is done in [16]. The resulting grid contains 48614 cells. We then distribute the events into these cells based on the coordinates. As crime is highly concentrated in urban environments [31], we find that 72% of the cells are empty (without any reported crimes), while the non-empty cells have a mean and maximum sequence length of ~ 4 and 149, respectively. Next, we split the sequence into training (80%), validation (10%), and test (10%) sets based on the event dates. The model parameters (base intensity μ and kernel g) are estimated using the training set. Specifically we adopt a fix time step ($\delta t = 1$ day) for the kernel (g). We also apply a cutoff (window size, ws) to g. For instance, for $ws = 30$ days and $\delta t = 1$ day, the kernel g becomes the discrete g_m with 30 bins. In addition, we vary the strength of



(a) $ws = 4$, $\mu_{TV} = 0$ (b) $ws = 4$, $\mu_{TV} = 0.01$ (c) $ws = 4$, $\mu_{TV} = 0.02$

(d) $ws = 2$, $\mu_{TV} = 0$ (e) $ws = 2$, $\mu_{TV} = 0.01$ (f) $ws = 2$, $\mu_{TV} = 0.02$

Fig. 1: Synthetic experiments. Hawkes process with an exponential kernel. Orange curve: true kernel; Blue curve: estimated kernel. Top: kernel is truncated at 4; Bottom: kernel is truncated at 2. From left to right, the figures represent Hawkes model, Hawkes-tv with $\mu_{TV} = 0.01$, and 0.02.

the TV constraint, μ_{TV}, to control the number of steps in the step-wise kernels. These hyper-parameters are tuned on the validation set; while the final forecast results are evaluated on the test set.

1) Evaluation metric: To evaluate forecasting, we adopt the metric - predictive accuracy index (PAI) [5], which is an area-standardized measure of recall@k commonly used in crime hotspotting applications:

$$PAI = \frac{top\ k\ recall}{k\ /\ total\ \#\ boxes}. \tag{15}$$

It is the fraction of crime falling in the top k forecasted grid cells divided by the fraction of land area flagged as a hotspot for intervention. In particular, we select the top k boxes (ranked by intensity), and calculate the recall ($\frac{true\ positive}{true\ positive\ +\ false\ negative}$) (i.e. the top k recall). We then divide it by $\frac{k}{total\ \#\ boxes}$, the fraction of the boxes flagged as a hotspot.

2) Results: In Table IV, we show the experimental results on the validation set. The window size is taken as 14, 30, and 90 days. Different μ_{TV} values are tested, among which we show those leading to well-behaved kernel curves. The $\#stairs$ column indicates the number of flat components (stairs) in the estimated kernel. The ratio of $\#stairs$ and the window size is calculated (the fourth column), to indicate the compactness of the kernel. In general, a ratio closer to zero suggests better interpretability. To evaluate the forecast accuracy, we estimate the top-20 recall (the rate column) and the PAI score introduced in Eq. 15. From this table, we can see that a stronger TV constraint would lead to a simpler but less accurate model. However, the degradation in accuracy is not significant. For instance, for the 14-day window, our TV-Hawkes model reduces $\#stairs$ by two thirds, while only dropping the PAI score by 2%. Similarly, comparing to the unregularized Hawkes model, TV-Hawkes decreases $\#stairs$ by 83% for the 30-day window and 95% for the 90-day window, with a cost in PAI of only 4% and 8%, respectively.

To compare the models listed in Table IV, we quantify three properties - simplicity, interpretability, and accuracy. Specifically, we define $simplicity = 1/\#stairs$, $interpretability = 0.1/ratio$, and $accuracy = PAI/100$. Note the scaling factors in these definitions are introduced to make the quantities have the same scale. The simplicity is defined as the inverse of $\#stairs$, since a simpler model should have less parameters (stairs). The interpretability is inversely proportional to the ratio, as a model describing a larger window with less parameters (stairs) should be considered as more interpretable. In Fig. 2, we represent the models in radar charts with the three properties as axes. In these charts, the triangle area can be viewed as a comprehensive evaluation of

a model. A larger area can indicate a better balance among the three properties. Compared to the regular Hawkes models, our TV-Hawkes models are significantly higher along the simplicity and interpretability axes, while only slightly lower along the accuracy axis. Our TV-Hawkes models also show larger area than the Hawkes models. Among the models listed, TV-Hawkes-14, TV-Hawkes-30-3, and TV-Hawkes-90-4 have the largest triangle area in the corresponding windows. In Fig. 3, we plot the estimated kernels of these three models along with the ones given by the Hawkes process. We can see that multiple bins in the Hawkes kernels merge into one large step in the TV-Hawkes kernels. Meanwhile, the TV-Hawkes kernels show similar trends as the Hawkes kernels.

Next, we apply these models to the test set for forecasting. The results over the test set are shown in Table V. We also perform forecasting using the un-regularized Hawkes process and a hotspot mapping model where the flagged area is comprised of the hotspots that have the most crimes in the previous window. Our TV-Hawkes model performs better than the Hawkes model in top 20 accuracy when $ws = 14$ days, top 100 accuracy when $ws = 14$ or 30 days, and top 200 accuracy when $ws = 90$ days. In the other cases, our TV-Hawkes model has accuracy only slightly lower than the Hawkes model. Noticeably, our TV-Hawkes model outperforms the hotspot mapping model, by a margin that becomes significant in the top 100 and top 200 cases. The top 20 and 200 scores suggest that using a larger window might improve performance; however, such effect is not obvious in the top 100 case. We also observe that from top 20 to top 200, the PAI score decreases, although the rate increases. This suggests that it is more difficult for these models to forecast a crime in the locations with lower risk.

To further compare TV-Hawkes against Hawkes, we plot the intensity in grid cells for a random day picked from the test set, on a map of Indianapolis metropolitan area (roughly inside I-465) in Fig. 4. Note here we illustrate the models with a 30-day window, i.e. Hawkes-30 and TV-Hawkes-30-3 (Table IV). Color represents greater (red) or lower (green) intensity. The transparent regions have an intensity of zero, and are very unlikely to have a crime. The colored cells in both figures show very similar patterns. In both figures, we find that the high intensity cells (red or orange) are distributed in a predominantly yellow and green background. In particular, the high intensity cells indicate that the area near the Children's Museum, the area near the Motor Speedway, and the area in the east side between East 21st and 10th streets are among the riskiest. The blue triangles (hotspots) mark the top-20 locations most likely to have a crime. Both models show similar hotspots, except for one near the left boundary of the map.

3) Simplified score cards: Finally, we detail our method for constructing interpretable score cards. We construct the score card (SC-1) in Table VII based on the model TV-Hawkes-30-3, where $ws = 30$ days and $\mu_{TV} = 3e-4$. The model gives the following intensity:

$$\begin{aligned}\lambda(t) = {} & 0.0029 \times \#\text{ crimes in } (t, t-1] \\ & + 0.0025 \times \#\text{ crimes in } [t-2, t-3] \\ & + 0.0014 \times \#\text{ crimes in } [t-4, t-7] \\ & + 0.0013 \times \#\text{ crimes in } [t-8, t-15] \\ & + 0.0011 \times \#\text{ crimes in } [t-16, t-30] \\ & + \mu_0,\end{aligned} \tag{16}$$

where t is the current date and μ_0 is the mean base intensity. We have $\mu_0 = 4.9218e-4$ over all cells, and $\mu_0 = 0.0019$ over the cells with crimes. We can simplify SC-1 by scaling the coefficients to the range of $[0, 10]$, leading to the second score card (SC-2) in the introduction in Table II. We can further substitute the number of crimes in Eq. 16 by the mean crime counts in the corresponding intervals, and scale the resulted terms to between $[0, 10]$. This renders the third score card (SC-3) in Table VIII. We apply these three score cards to the IMPD crime data (test set), and show the forecasting results in Table VI. The first score card (SC-1) has the highest accuracy among the three, and the accuracy decreases as the score card becomes less complex. Noticeably, SC-1 outperforms the hotspot mapping model in rate@20 and PAI@20, by $\sim 2\%$. Compared to the TV-Hawkes and Hawkes models, the score cards are less accurate, however, they are much easier to use and interpret. We also see that from top 20 to 200, the rate increases while the PAI score decreases, a phenomenon similar to the more complicated models.

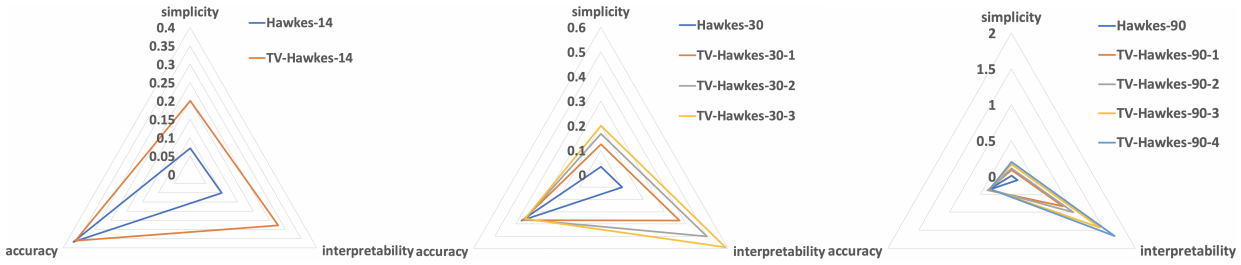
We also plot the intensity maps given by the score cards in Fig. 4. Notice that the maps on the bottom have more cells colored in orange or red than those on the top. This suggests that the score cards can in general assign higher intensity to the cells than the more complex Hawkes and TV-Hawkes models, which is likely the result of rounding the parameters in TV-Hawkes. The top 20 forecasted cells (blue triangles) show similar pattern across score cards. Although the forecasted locations may not exactly overlap with those given by the Hawkes and TV-Hawkes models, they appear to be in close proximity.

IV. Conclusion

Here we showed how to construct interpretable Hawkes process crime forecasting models using total-variation penalized likelihood estimation. The TV-Hawkes process allows one to balance accuracy, simplicity and interpretability. The methodology presented here has the advantage that feature engineering is not required and cut-offs for the step function comprising $g(t)$ are automatically generated during inference. This is in comparison to interpretable models of recidivism where hand-crafted feature engineering is required before the mixed-integer optimization. One disadvantage of our methodology is that we used post-processing to convert the model into an integer score. Future work in this area will focus on solving a non-linear TV-MPLE integer programming problem similar to what is done in [26] for logistic regression (where piece-wise linear approximate MIP is used).

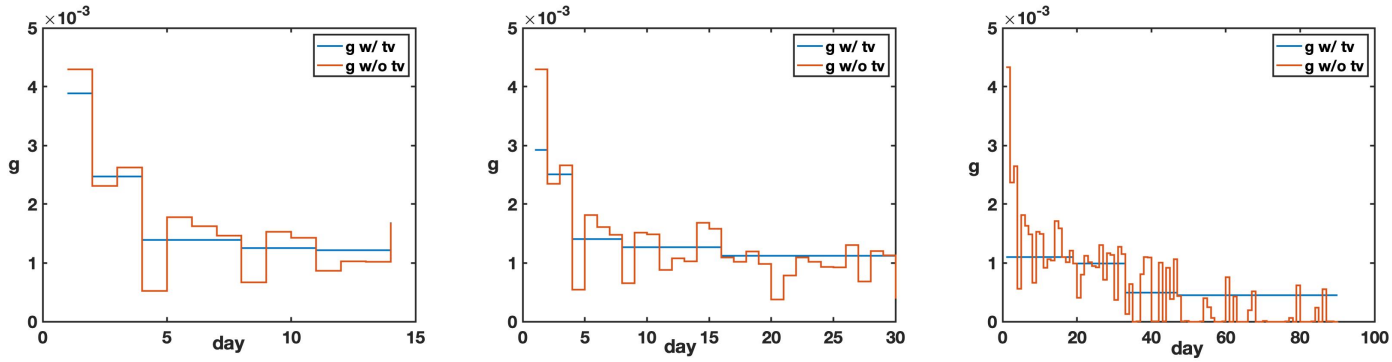
TABLE IV: IMPD experimental results on the validation set. ws is the window size for g_m (# of days). We use $\delta t = 1\ day$, and $150 \times 150\ m$ boxes. The ratio is $\#stairs/ws$. Top 20 boxes are used to calculate the recall.

model	ws	μ_{TV}	# stairs	ratio	rate	PAI
Hawkes-14	14	0	-	-	1.51%	36.7534
TV-Hawkes-14	14	1×1e-4	5	0.36	1.49%	36.2012
Hawkes-30	30	0	-	-	1.53%	37.2926
TV-Hawkes-30-1	30	1×1e-4	8	0.27	1.52%	36.8698
TV-Hawkes-30-2	30	2×1e-4	6	0.20	1.49%	36.1138
TV-Hawkes-30-3	30	3×1e-4	5	0.17	1.47%	35.6976
Hawkes-90	90	0	-	-	1.56%	37.9458
TV-Hawkes-90-1	90	1×1e-4	11	0.12	1.56%	37.9458
TV-Hawkes-90-2	90	5×1e-4	9	0.10	1.48%	35.8963
TV-Hawkes-90-3	90	1×1e-3	6	0.07	1.45%	35.2293
TV-Hawkes-90-4	90	1.2×1e-3	5	0.06	1.43%	34.8751



(a) 14 days (b) 30 days (c) 90 days

Fig. 2: Model comparison using IMPD dataset. The three axes in the radar chart are simplicity, interpretability, and accuracy. From left to right, the charts represent models with window sizes of 14, 30, and 90 days.



(a) 14 days (b) 30 days (c) 90 days

Fig. 3: Estimated kernels using the IMPD data. From left to right, the window sizes are 14, 30, and 90 days; $\mu_{TV} = 1 \times 1e-4$, $3 \times 1e-4$, and $1.2 \times 1e-3$. Blue curve: TV-Hawkes; Red curve: Hawkes.

Our results inform several demands and realities grounded in the application of identifying crime hotspots and crime forecasting. First, we believe these models will be useful in meeting the growing demand for transparency and perceived fairness in hotspot policing [24], [27]. Advocacy groups and scholars have voiced various concerns over the process through which crime hotspots and forecasting are calculated, thus our proposed methodology offers an alternative approach which begins to answer these calls. Second, despite advancements in information technology and availability of point-level data, the majority of police departments lack the appropriate personnel – either technical skills or time availability – to engage in meaningful crime analysis [2], [18]. Moreover, spatiotemporal analyses and identification of micro-time hotspots to direct police resources is even less likely to occur within the average police department [10], [23]. This reality is unfortunate given recent studies have shown the promise of near-repeat, micro-time hotspotting interventions to have crime reduction benefits [9], [11], [25]. Lastly, police agencies struggle to navigate effective hotspotting techniques in light of sparse event counts [7]. Inappropriate techniques are likely to lead to the identification of statistically random locations when the number of potential crime places exceeds the number of events [4]. For these reasons, we believe our proposed methodology better enables

TABLE V: IMPD experimental results on the test set. ws is the window size for g_m (# of days). We use $\delta t = 1$ day, and 150×150 m boxes. Rate and PAI are calculated using top 20, 100, and 200 recalls.

		TV-Hawkes		Hawkes		Hotspot	
top	ws	rate	PAI	rate	PAI	rate	PAI
20	14	**1.54%**	**37.5418**	1.53%	37.1646	1.37%	33.2356
	30	**1.61%**	39.0435	**1.61%**	**39.1838**	1.47%	35.6655
	90	1.68%	40.8551	**1.73%**	**41.9390**	1.68%	40.8238
100	14	**5.76%**	**27.9778**	5.73%	27.8763	3.50%	17.0237
	30	**5.83%**	**28.3535**	5.73%	27.8562	4.44%	21.5816
	90	5.70%	27.6959	**5.77%**	**28.0456**	4.90%	23.7972
200	14	9.72%	23.6277	**9.74%**	**23.6723**	5.01%	12.1832
	30	9.88%	24.0245	**9.89%**	**24.0362**	6.78%	16.479
	90	**10.05%**	**24.4383**	10.01%	24.3247	8.58%	20.8585

TABLE VI: Forecasting results with the score cards (SC). The results are evaluated on the IMPD crime data (test set), with a 30-day window. We set $\delta t = 1$ day, and use 150×150 m boxes. Rate and PAI are estimated on top 20, 100, and 200 forecasted hotspot grid cells. As comparison, results by other models in the 30-day window are listed.

	SC-1	SC-2	SC-3	TV-Hawkes	Hawkes	Hotspot
rate@20	1.50%	1.36%	1.25%	**1.61%**	**1.61%**	1.47%
PAI@20	36.3596	33.0772	30.4465	39.0435	**39.1838**	35.6655
rate@100	4.25%	4.06%	4.01%	**5.83%**	5.73%	4.44%
PAI@100	20.6653	19.7543	19.4889	**28.3535**	27.8562	21.5816
rate@200	6.47%	6.33%	5.88%	9.88%	**9.89%**	6.78%
PAI@200	15.7164	15.3641	14.2906	24.0245	**24.0362**	16.479

Did crimes occur in the past	1 days?	(# × 29 points)	...
	2-3 days?	(# × 25 points)	+ ...
	4-7 days?	(# × 14 points)	+ ...
	8-15 days?	(# × 13 points)	+ ...
	16-30 days?	(# × 11 points)	+ ...
Did a crime ever occur?		(19 points)	+ ...
		Total	= ...

TABLE VII: Interpretable crime forecast score card. Each spatial grid cell in a city is scored using the card and the grid cells with the highest score are flagged as hotspots.

Did crimes occur in the past	1 days?	(3 points)	...
	2-3 days?	(2 points)	+ ...
	4-7 days?	(1 points)	+ ...
	8-15 days?	(1 points)	+ ...
	16-30 days?	(2 points)	+ ...
Did a crime ever occur?		(1 points)	+ ...
		Total	= ...

TABLE VIII: (Further simplified) Interpretable crime forecast score card. Each spatial grid cell in a city is scored using the card and the grid cells with the highest score are flagged as hotspots.

police agencies to leverage their existing data to maximize police resources through empirically-driven interventions.

TV-Hawkes code available at https://github.com/daDiz/TV-Hawkes

ACKNOWLEDGMENTS

This research was supported by NSF grants SCC-1737585 and ATD-1737996. Author GM serves on the board of PredPol.

REFERENCES

[1] Barak Ariel, Cristobal Weinborn, and Lawrence W Sherman. "soft" policing at hot spots—do police community support officers work? a randomized controlled trial. *Journal of Experimental Criminology*, 12(3):277–317, 2016.

[2] Jyoti Belur and Shane Johnson. Is crime analysis at the heart of policing practice? a case study. *Policing and society*, 28(7):768–786, 2018.

[3] Richard A Berk and Justin Bleich. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Pub. Pol'y*, 12:513, 2013.

[4] Anthony A Braga, Andrew V Papachristos, and David M Hureau. The concentration and stability of gun violence at micro places in boston, 1980–2008. *Journal of Quantitative Criminology*, 26(1):33–53, 2010.

[5] Grant Drawve and Alese Wooditch. A research note on the methodological and theoretical considerations for assessing crime forecasting accuracy with the predictive accuracy index. *Journal of Criminal Justice*, 64(C):1–1, 2019.

[6] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[7] Seth Flaxman, Michael Chirico, Pau Pereira, Charles Loeffler, et al. Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: a winning solution to the nij "real-time crime forecasting challenge". *The Annals of Applied Statistics*, 13(4):2564–2585, 2019.

[8] Tom Goldstein, Christoph Studer, and Richard Baraniuk. A field guide to forward-backward splitting with a fasta implementation. *arXiv preprint arXiv:1411.3406*, 2014.

[9] Elizabeth Groff and Travis Taniguchi. Using citizen notification to interrupt near-repeat residential burglary patterns: the micro-level near-repeat experiment. *Journal of Experimental Criminology*, 15(2):115–149, 2019.

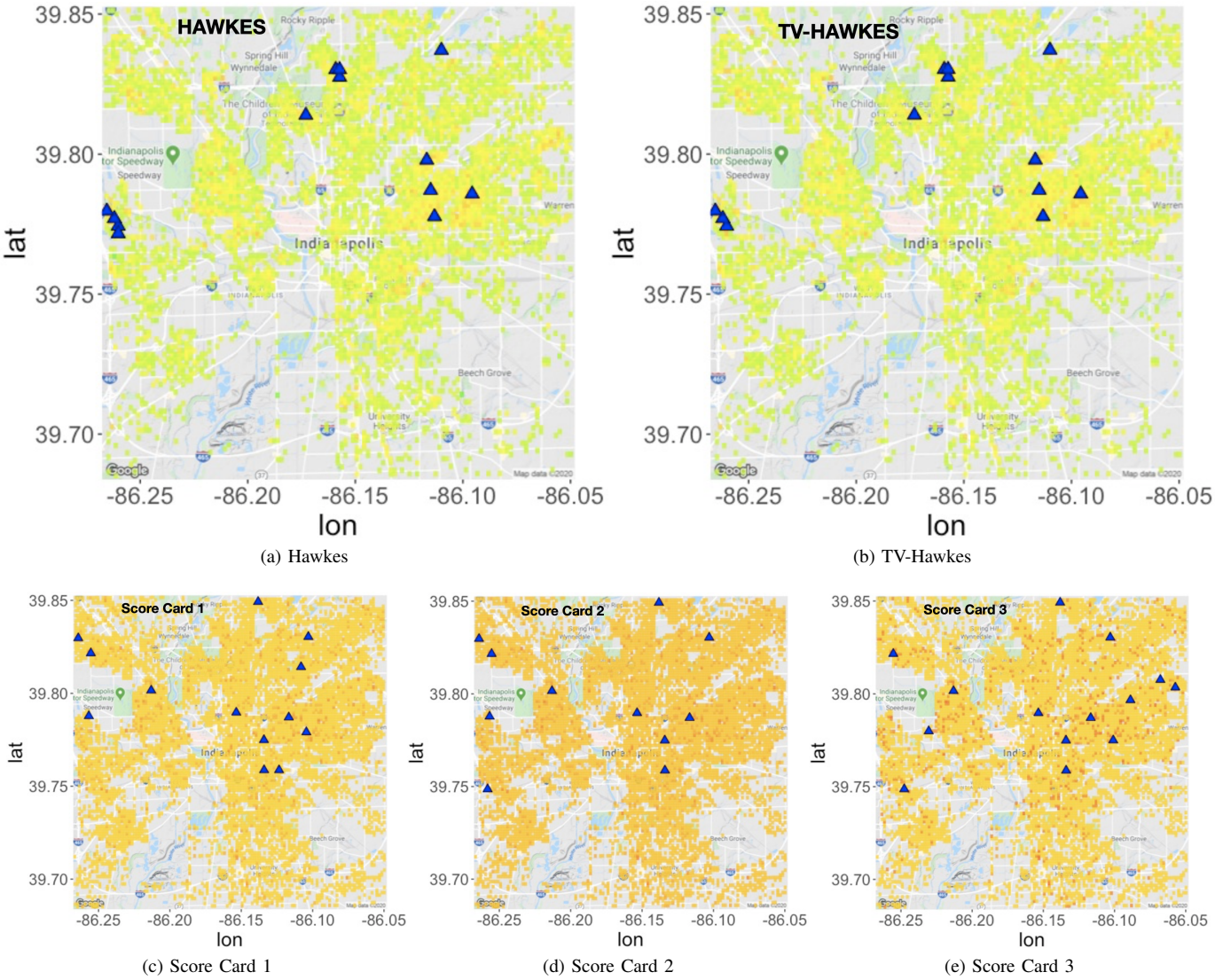
[10] Cory P Haberman and Jerry H Ratcliffe. The predictive policing challenges of near repeat armed street robberies. *Policing: A Journal of Policy and Practice*, 6(2):151–166, 2012.

[11] Shane D Johnson, Toby Davies, Alex Murray, Paul Ditta, Jyoti Belur, and Kate Bowers. Evaluation of operation swordfish: a near-repeat target-hardening strategy. *Journal of experimental criminology*, 13(4):505–525, 2017.

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

[13] George Mohler and Michael D Porter. Rotational grid, pai-maximizing crime forecasts. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 11(5):227–236, 2018.

[14] George O Mohler, Andrea L Bertozzi, Thomas A Goldstein, and Stanley J Osher. Fast tv regularization for 2d maximum penalized like-



(a) Hawkes (b) TV-Hawkes

(c) Score Card 1 (d) Score Card 2 (e) Score Card 3

Fig. 4: Intensity for a random day (from the test set). Top left: intensity estimated using regular Hawkes process; Top right: intensity estimated using TV-Hawkes process (30-day window). Bottom: from left to right, intensity estimated using Score Card 1, 2, and 3. The grid cells are colored based on the fourth root of intensity (strong: red, medium: yellow, weak: green). The blue triangles represent the top-20 forecasted cells. The cells with zero intensity are set as transparent.

lihood estimation. *Journal of Computational and Graphical Statistics*, 20(2):479–491, 2011.

[15] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

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

[17] Y. Ogata. On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

[18] Jerry Ratcliffe. Predictive policing. *Police Innovation: Contrasting Perspectives*, page 347, 2019.

[19] Jerry H Ratcliffe, Ralph B Taylor, Amber Perenzin Askey, Kevin Thomas, John Grasso, Kevin J Bethel, Ryan Fisher, and Josh Koehnlein. The philadelphia predictive policing experiment. *Journal of Experimental Criminology*, pages 1–27, 2020.

[20] Alex Reinhart and Joel Greenhouse. Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1305–1329, 2018.

[21] Tammy Rinehart Kochel and David Weisburd. The impact of hot spots policing on collective efficacy: Findings from a randomized field trial. *Justice Quarterly*, 36(5):900–928, 2019.

[22] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[23] Rachel B Santos and Roberto G Santos. Examination of police dosage in residential burglary and residential theft from vehicle micro-time hot spots. *Crime Science*, 4(1):27, 2015.

[24] Rachel Boba Santos. Predictive policing: Where's the evidence? *Police Innovation: Contrasting Perspectives*, page 366, 2019.

[25] Roberto G Santos and Rachel Boba Santos. An ex post facto evaluation of tactical police response in residential theft from vehicle micro-time hot spots. *Journal of Quantitative Criminology*, 31(4):679–698, 2015.

[26] Toshiki Sato, Yuichi Takano, Ryuhei Miyashiro, and Akiko Yoshise.

Feature subset selection for logistic regression via mixed integer optimization. *Computational Optimization and Applications*, 64(3):865–880, 2016.

[27] Aaron Shapiro. Reform predictive policing. *Nature News*, 541(7638):458, 2017.

[28] Shinichiro Shirota, Alan E Gelfand, et al. Space and circular time log gaussian cox processes with application to crime event data. *The Annals of Applied Statistics*, 11(2):481–503, 2017.

[29] Berk Ustun, Stefano Traca, and Cynthia Rudin. Supersparse linear integer models for interpretable classification. *arXiv preprint arXiv:1306.6677*, 2013.

[30] Glenn D Walters. Predicting criminal justice outcomes with the psychopathy checklist and lifestyle criminality screening form: a meta-analytic comparison. *Behavioral sciences & the law*, 21(1):89–102, 2003.

[31] David Weisburd. The law of crime concentration and the criminology of place. *Criminology*, 53(2):133–157, 2015.

[32] Andrew P Wheeler. Allocating police resources while limiting racial inequality. *Justice Quarterly*, pages 1–27, 2019.

[33] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):689–722, 2017.